2024

Spawozdanie

PROJEKT ADM

KRZYSZTOF KOWALSKI

Cel projektu

Wykonanie dwóch rozmów pomiędzy dwoma osobami: każda osoba 3 x po około 30 sekund naprzemiennie mówią, jedna rozmowa rano, kolejna wieczorem (już nie zważając na czas, a analogiczność wypowiedzi). Następnie identyfikacja początków i końców naszych wypowiedzi oraz zliczenie liczby słów w każdym z przedziałów.

Drugim etapem projektu było wybranie obu tych samych fragmentów po jednym z każdej z rozmów i dokonanie analizy wypowiadanych słów, przerw między nimi oraz wyciągnięcie wniosków z otrzymanych wyników.

Parametry techniczne

- Rozmowy nagrywano wbudowanym w telefon Iphone dyktafonem na jednym kanale, którego częstotliwość próbkowania to 48 000 Hz.

- Rozmowa prowadzona była między kobietą oraz mężczyzną. Jako pierwsza mówiła kobieta.

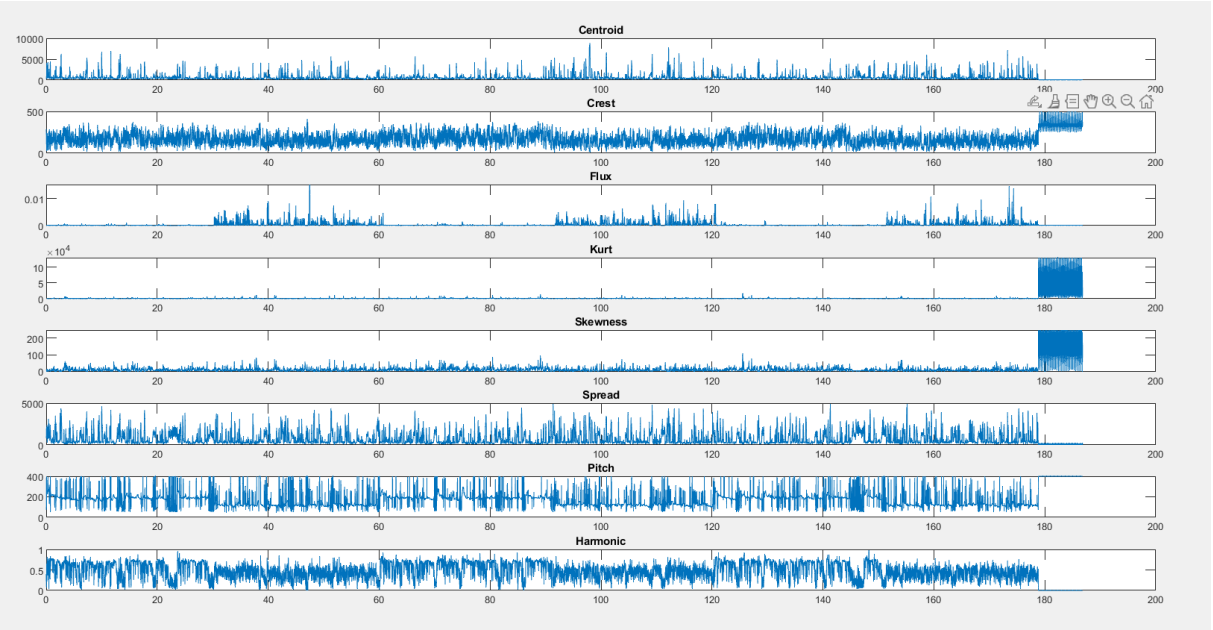
- Nagrania zostały zapisane do domyślnego formatu w którym zapisują Iphone – m4a.

- Damska kwestia pochodzi z książki architektonicznej o stylach, natomiast męska z piątego tomu kryminału Chyłka Remigiusza Mroza – „Inwigilacja".

Przebieg projektu

1. **Część pierwsza – znalezienie początków i końców każdej z wypowiedzi oraz zliczenie liczby słów w każdym ze znalezionych przedziałów.**

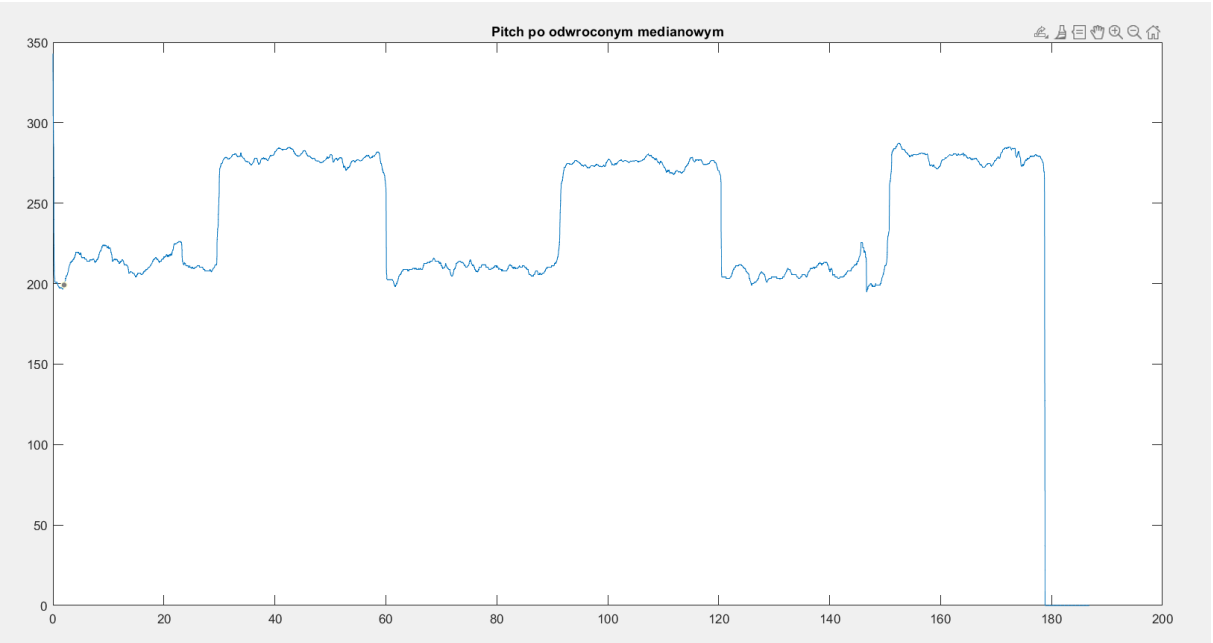
Tą część wykonywano na nagraniu wieczornym. Po wczytaniu sygnału oraz zapoznaniu się z jego przebiegiem zdecydowano się na dodanie do końca wypowiedzi lekkiego szumu sinusoidalnego, dzięki któremu lepiej uda się zlokalizować koniec ostatniej męskiej wypowiedzi – ponieważ jest ona na końcu nagranie wykonywanie filtru medianowego sprawi, że koniec piku zniekształci się. Dodanie szumu na końcu zapobiegnie temu zdarzeniu.



Rysunek 1Obliczone różne metody na sygnale

Po zapoznaniu się z powyższym wykresem zdecydowano na wykorzystanie metody **Pitch** do zlokalizowania początków i końców wypowiedzi, natomiast do zliczania słów skorzystano z metody **Flux.**

Na wyniku metody Pitch dokonano filtracji medianowej o rozmiarze okna 319 a następnie odwrócono wynik tej filtracji (góra-dół).

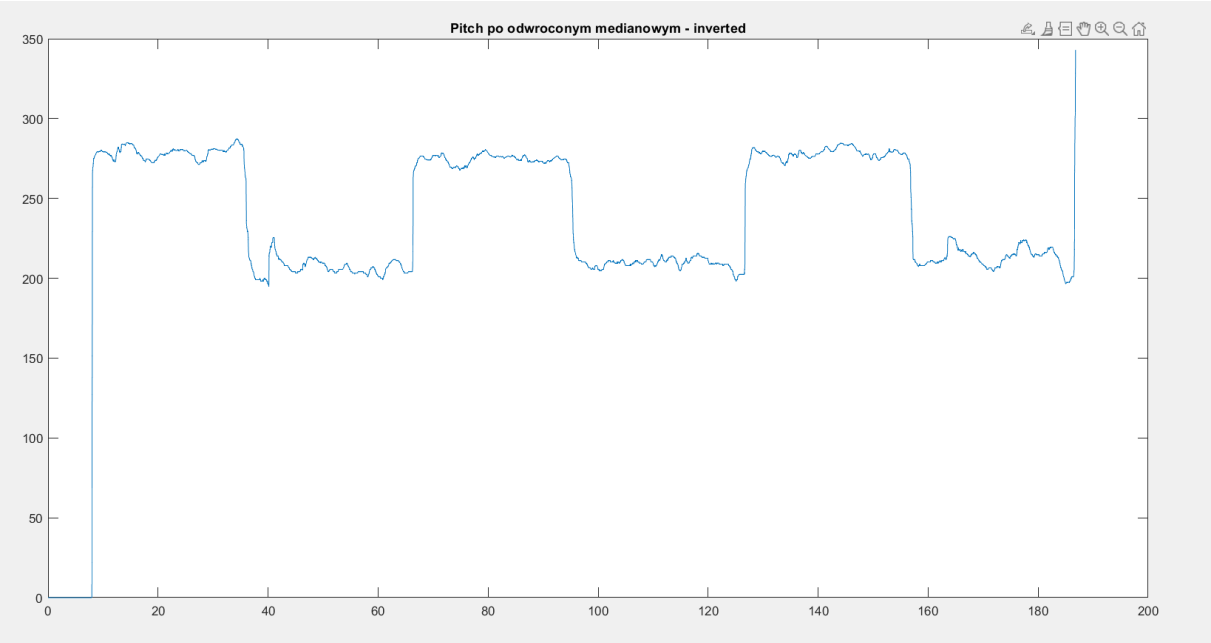


Rysunek 2 Otrzymany wynik po odwróceniu filtracji

Ponieważ chciano znaleźć wypowiedzi osoby mówiącej jako druga widoczna na powyższym rysunku wielkość pików jaka pozwoli zlokalizować początki szukanych wypowiedzi to około 270. Taką wartość zastosowano. Początek obszaru drugiego znaleziono poprzez przesunięcie czasu o

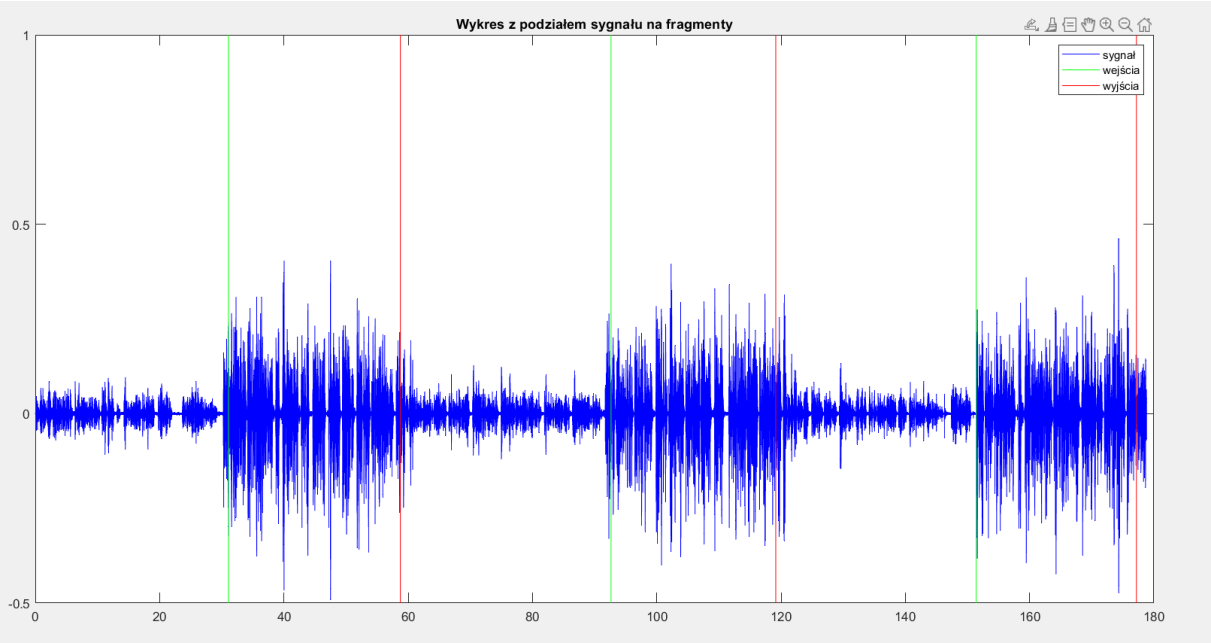
długość wypowiedzi (około 30 sekund), natomiast trzeciego analogicznie tylko stosując się do czasu drugiej wypowiedzi.

Do znalezienia wyjść skorzystano z tego samego wyniku filtracji tym razem obracając ją dodatkowo prawo-lewo. Dzięki czemu można było zastosować analogiczne znajdowanie początków wypowiedzi, które dla wejściowego sygnału będą końcami. Należało ostatecznie od długości sygnału odjąć znalezioną próbkę.



Rysunek 3 Otrzymane wyniki po dodatkowym obrocie wyniku filtracji

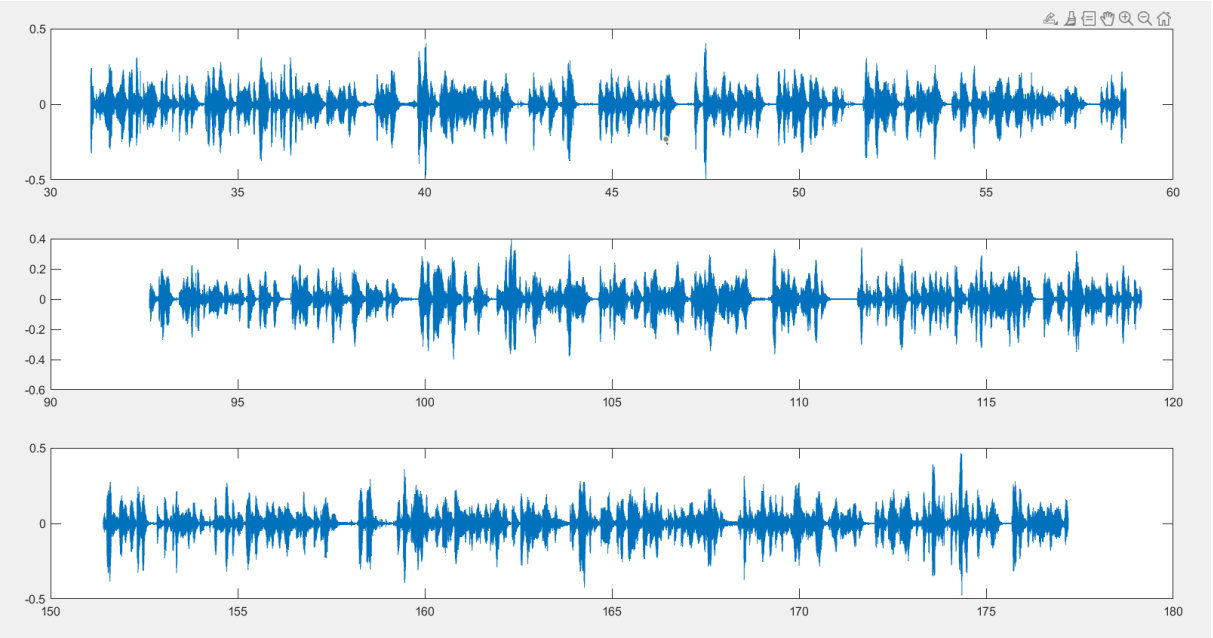
Po znalezieniu początków i końców można było zbadać poprawność znajdowania wypowiedzi nakładając wartości na wejściowy sygnał.



Rysunek 4 Sygnał wejściowy z nałożonymi przedziałami

Wyniki świadczą o dosyć precyzyjnym znalezieniu szukanych wypowiedzi.

W celu obliczenia liczby słów należało wyciąć wypowiedzi.



Rysunek 5 Wycięte wypowiedzi

A następnie przejść pętlą po każdym z sygnałów i korzystając z wyniku metody Flux ustawić odpowiednio parametry pików oraz dystansów między słowami. Zastosowano podejście „uczenia maszynowego" aby zoptymalizować i nie przeoczyć potencjalnie dobrego wyniku detekcji słów stworzono dwie pętle przechodzące po dystansach i wysokościach z małym krokiem których zadaniem było minimalizowanie błędu. Następnie aby zważyć błąd dla wszystkich trzech wypowiedzi wyciągnięto średnią dla każdego z wyników (średnią z trzech najlepszych dystansów i trzech najlepszych wysokości). Tym sposobem minimalny odległość między pikami wynosiła 21, natomiast wysokość pików 0.00049.

NUMER FRAGMENTU	POCZĄTEK [S]	KONIEC [S]	ILOŚĆ SŁÓW WYPOWIEDZIANYCH	ILOŚĆ SŁÓW OBLICZONYCH	BŁĄD [%]
1	31.0720	58.7413	67	70	4.4776
2	92.6400	119.1573	72	69	4.1667
3	151.4027	177.1947	71	66	7.0423
SUMA	-	-	210	205	5.2381

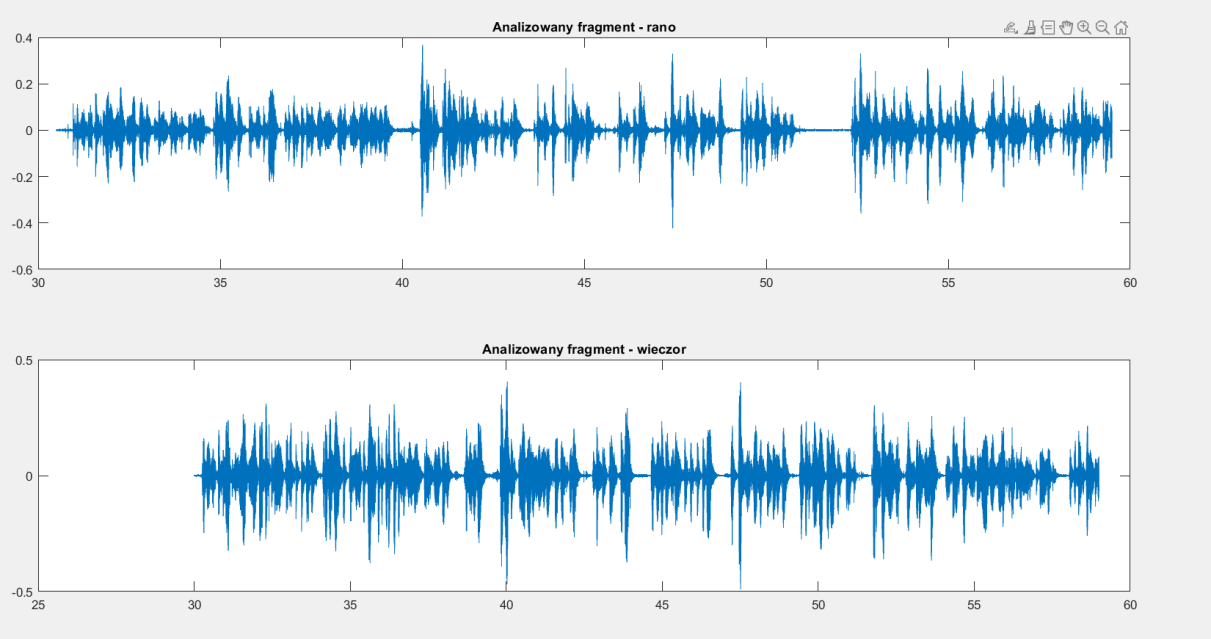
Wykorzystane nagranie/pliki to:

- Normalne.m4a

- zliczanie_slow.m

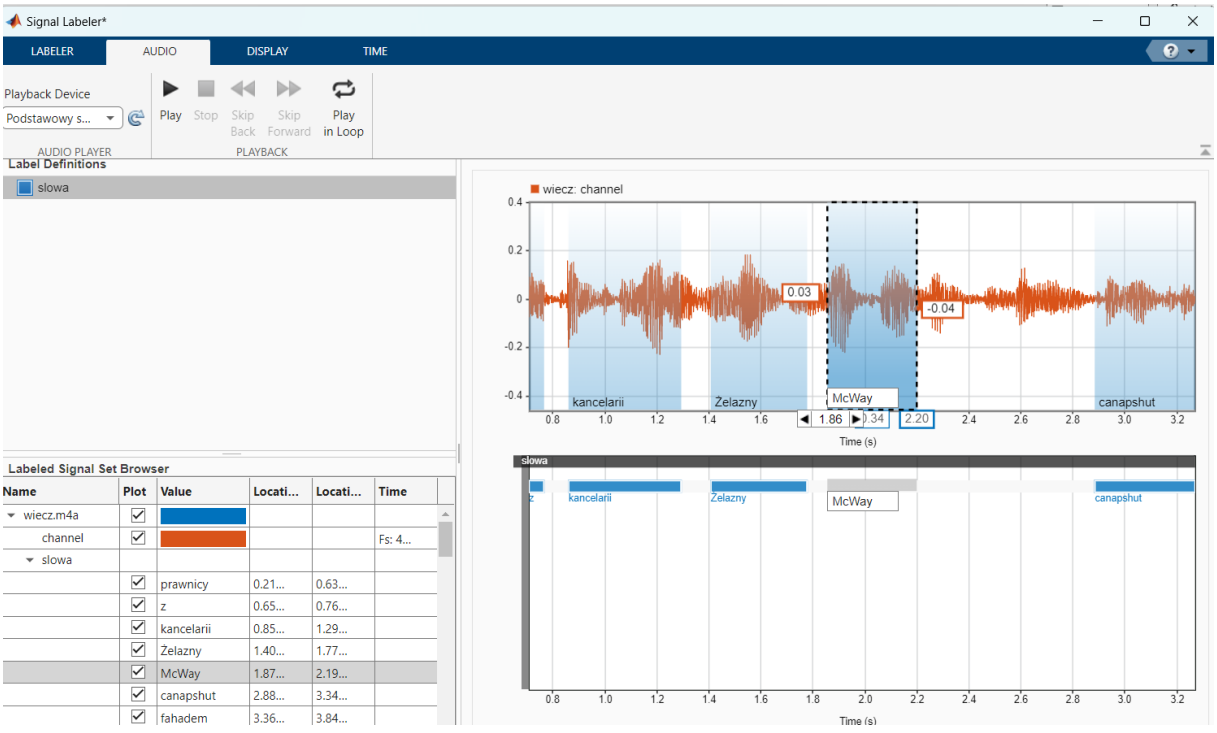
2. Część druga – analiza różnic rano, a wieczór

Do tej części wykorzystano pierwszą część męską z nagrania wieczornego oraz porannego.



Rysunek 6 Analizowane fragmenty

Następnie w celu znalezienia początków i końców każdego ze słów skorzystano z narzędzia **Signal Labeler,** gdzie korzystając z sieci Wav2vec zlokalizowano słowa. Po automatycznym wykrywaniu należało dokonać korekty – sieć nie posiada wysokiej precyzji dla wypowiedzi w języku polskim.



Rysunek 7 Etap manualnego poprawiania znalezionych przez sieć słów

W ten sposób otrzymano dwa pliki matlabowe: rano_ls.mat i wieczor_ls.mat, które w swojej strukturze zawierały słowa wraz z czasem ich początku i końca.

Następnie można było przejść do analizy czasu wypowiedzi każdego ze słów. Otrzymano w ten sposób

Slowo	SredniCzasRano	SredniCzasWieczorem	Roznica
{'Al-Jassam'}	0.46277	0.42756	-0.035208
{'Chyłka'}	0.3384	0.31048	-0.027917
{'Fahadem'}	0.45404	0.43267	-0.021366
{'Kordian'}	0.33321	0.32667	-0.0065369
{'McWay'}	0.3101	0.34379	0.033687
{'a'}	0.084274	0.08006	-0.0042143
{'aplikant'}	0.4481	0.355	-0.093104
{'był'}	0.12879	0.091938	-0.036854
{'dla'}	0.097667	0.10588	0.0082083
{'do'}	0.06539	0.071344	0.0059537
{'dopiero'}	0.27723	0.2645	-0.012729
{'ducha'}	0.30448	0.30829	0.0038125
{'i'}	0.079542	0.042153	-0.037389
{'jakby'}	0.28596	0.26296	-0.023
{'kancelarii'}	0.44663	0.4332	-0.013429
{'klientowi'}	0.46713	0.38071	-0.086417
{'kurtkę'}	0.35562	0.2136	-0.14202
{'mieszkania'}	0.39604	0.31354	-0.0825
{'mieszkanie'}	0.37344	0.35863	-0.014808
{'milczał'}	0.32173	0.31142	-0.010313
{'może'}	0.18335	0.16317	-0.020188
{'na'}	0.1165	0.080968	-0.035532
{'od'}	0.10744	0.10008	-0.0073625

Rysunek 8 Początek słów

Informuje, że w przypadkach gdy:

- **Różnica > 0**: Średni czas wypowiadania słowa wieczorem jest dłuższy niż rano. Słowo jest wypowiadane wolniej wieczorem.

- **Różnica < 0**: Średni czas wypowiadania słowa wieczorem jest krótszy niż rano. Słowo jest wypowiadane szybciej wieczorem.

- **Różnica = 0:** Średni czas wypowiadania słowa jest taki sam rano i wieczorem. Słowo jest wypowiadane w tym samym tempie.

Dokonano również analizy średniego czasu przerw między słowami.

- dla nagrania porannego: 0.13s

- dla nagrania wieczornego: 0.14s

Wnioski

- Stosując odpowiednie metody można z dosyć dobrą precyzją rozróżnić w rozmowie fragment w którym mówi kobieta od tego w którym mówi mężczyzna (połączenie metody Pitch z filtrem medianowym i odpowiednim jego obrotem).

- Metoda Flux pozwala na dokonanie detekcji słów w wypowiedzi z dosyć dobrą precyzją.

- Siec Wav2vec znakomicie radzi sobie z rozpoznawaniem słów w języku angielskim, niestety dla języka polskiego nie jest aż tak precyzyjna.

- Dokonując podziału na słowa dobrze widoczne na wykresy były slaby.

- Więcej różnic jest ujemnych, a zatem rano słowa wypowiadamy szybciej niż wieczorem, co jest sensowne po całym dniu i zmęczeniu.

- Przerwy między słowami są większe wieczorem, co również jest sensownym wnioskiem i może być spowodowane zmęczeniem.

- Okres dnia (wpływ zmęczenia) ma wpływ na długość wypowiadanych słów oraz na pauzy między słowami.

- Powyższe wnioski oraz przebieg projektu świadczą o pomyślności wykonanego ćwiczenia.