Analiza Danych Multimedialnych
Projekt zaliczeniowy
Filip Hałys

Cel projektu

Celem projektu jest analiza danych multimedialnych w postaci dwóch (jednej porannej, drugiej wieczornej) identycznych konwersacji dwóch osób. Projekt podzielono na dwa etapy. Pierwszy z nich obejmował analizę pojedynczego nagrania, a więc automatyczną detekcję początku i końca fragmentu wypowiedzi danej osoby, zliczenie ilości słów, a także policzenie błędu detekcji danego fragmentu. Drugi etap polegał na analizie zmęczenia osób, czyli na porównaniu nagrania porannego z wieczornym na podstawie. Zmęczenie określono używając kilka cech, takich jak różnice między czasem wypowiadania tych samych słów, tendencję do wydłużania/skracania przerwy między słowami, czy istnienie zmian tonu podstawowego dla poszczególnych. Do wykonania projektu posłużono się programem MatLab.

Opis danych

Do analizy przygotowano dwa zestawy danych formacie .MP3: *morning.mp3* i *evening.mp3*. Są to dwa nagrania powstałe poprzez wyrecytowanie pierwszych osiemnastu zwrotek wiersza „Pani Twardowska" autorstwa Adama Mickiewicza. Każde nagranie składa się z naprzemiennego czytania wiersza przez 2 osoby zgodnie z zasadą: co 3 zwrotki (co około 30 sekund) zmiana. Ponadto każde z nich trwa około 3 minuty (poranne 2:52, wieczorne 2:46) Poniżej w Tab. 1 zaprezentowano fragmenty wiersza, które zostały przeczytane i nagrane z podziałem na osoby: os. 1 (głos męski – Filip) oraz os. 2 (głoś żeński – Urszula).

Osoba	Treść	Ilość słów
Os. 1	Jedzą, piją, lulki palą, Tańce, hulanka, swawola; Ledwie karczmy nie rozwalą, Ha, ha! Hi, hi! hejże! hola! Twardowski siadł w końcu stoła, Podparł się w boki jak basza: „Hulaj dusza! hulaj!" woła, Śmieszy, tumani, przestrasza. Żołnierzowi, co grał zucha, Wszystkich łaje i potrąca, Świsnął szablą koło ucha: Już z żołnierza masz zająca.	50
Os. 2	Na patrona z trybunału, Co milczkiem wypróżniał rondel, Zadzwonił kieską, pomału: Z patrona robi się kondel. Szewcu w nos wyciął trzy szczutki, Do łba przymknął trzy rureczki, Cmoknął: cmok! i gdańskiej wódki Wytoczył ze łba pół beczki. Wtem, gdy wódkę pił z kielicha, Kielich zaświstał, zazgrzytał; Patrzy na dno: — „Co u licha? Po coś tu, kumie, zawitał?"	57
Os. 1	Diablik to był w wódce na dnie: Istny Niemiec, sztuczka kusa; Skłonił się gościom układnie, Zdjął kapelusz i dał susa. Z kielicha aż na podłogę Pada, rośnie na dwa łokcie,	55

	Nos jak haczyk, kurzą nogę, I krogulcze ma paznokcie. „A, Twardowski… witam bracie!” To mówiąc, bieży obcesem: „Cóż to, czyliż mię nie znacie? Jestem Mefistofelesem.	
Os. 2	Wszak ze mnąś na Łysej Górze Robił o duszę zapisy: Cyrograf na byczej skórze Podpisałeś ty i bisy. Miały słuchać twego rymu; Ty, jak dwa lata przebiegą, Miałeś pojechać do Rzymu, By cię tam porwać jak swego. Już i siedem lat uciekło, Cyrograf nadal nie służy: Ty, czarami dręcząc piekło, Ani myślisz o podróży.	54
Os. 1	Ale zemsta, choć leniwa, Nagnała cię w nasze sieci: Ta karczma *Rzym* się nazywa… Kładę areszt na waszeci”. Twardowski ku drzwiom się kwapił Na takie *dictum acerbum*; Diabeł za kontusz ułapił: „A gdzie jest *nobile verbum*?” Co tu począć? kusa rada, Przyjdzie już nałożyć głową… Twardowski na koncept wpada I zadaje trudność nową.	53
Os. 2	„Patrz w kontrakt, Mefistofilu, Tam warunki takie stoją: Po latach tylu a tylu, Gdy przyjdziesz brać duszę moją, Będę miał prawo trzy razy Zaprząc ciebie do roboty, A ty najtwardsze rozkazy Musisz spełnić co do joty. Patrz, oto jest karczmy godło, Koń malowany na płótnie; Ja chcę mu wskoczyć na siodło, A koń niech z kopyta utnie.	57

Tab. 1 – Fragmenty wiersza z podziałem na fragmenty, przeczytane przez daną osobę

Nagrania *morning.mp3* dokonano 13 maja około godziny 7:00, natomiast nagranie *evening.mp3* stworzono tego samego dnia około 20:00.

Oba nagrania zostały dołączone jako załączniki do tego sprawozdania.

Parametry sprzętowe

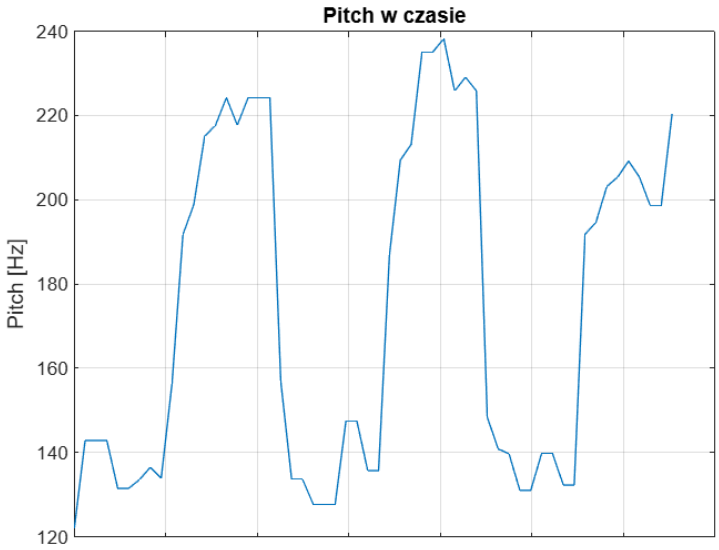
Wykorzystano dyktafon w telefonie Iphone 12 mini:

- Typ: Mikrofon wbudowany,
- Liczba mikrofonów wbudowanych: 3 – mikrofon główny (przy złączu), mikrofon do redukcji szumów (przy kamerce), mikrofon tylni (przy tylnym obiektywie),
- MEMS (Micro-Electro-Mechanical Systems),
- Częstotliwość 44,1 kHz,
- Kanał: mono.

Metodyka i przebieg projektu – analiza pojedynczej rozmowy

Do wykonania tej części projektu posłużono się nagraniem porannym: *morning.mp4*.

Na początku załadowano dane i zainicjalizowano *audioFeatureExtractor*, który wyznaczył cechy sygnału, takie jak tonacja (pitch), zmiany spektralne (spectra flux), harmonic ratio i spectra spread. Z otrzymanego ekstraktora wyodrębniono kolumnę z pitch. Nastepnie została ona (tonacja) przekształcona na średnie wartości w oknach o stałej długości, a także wygładzona filtrem medianowym. W kolejnym kroku zidentyfikowano fragmenty wypowiedzi, dla których pitch nie przekroczył 160 Hz (zwykle głos męski jest poniżej 160 Hz, a damski powyżej). Poniżej zaprezentowano wykres *Wyk. 1* pitch (oś y) na przestrzeni czasu (oś x). To właśnie na jego podstawie wydzielono 160 Hz jako wartość progową.



Wyk. 1 – Wykres pitch, na przestrzeni czasu

Na powyższym wykresie ewidentnie wybijają się 3 piki. Są to 3 fragmenty o wysokiej częstotliwości – głos os. 2 (żeński). Trzy przerwy między pikami (fragmenty o niskiej częstotliwości) to głos os. 1 (męski).

W następnym etapie zidentyfikowane fragmenty mowy pogrupowano w segmenty. Pominięto krótkie segmenty trwające poniżej 10 sekund. Dla każdego utworzonego segmentu utworzono wektor identyfikatorów oraz maskę, która pozwoliła wydzielić fragmenty mowy przypisane do konkretnej osoby.

Następnie wykorzystano cechę spectral flux do wykrywania słów (ilości słów). Dla każdego z trzech segmentów os. 1 wycięto odpowiedni fragment flux, następnie go wygładzono, wyznaczono lokalny próg (wykorzystując wzór: *0.1*średnia+0.34*odch.stand.*), wykryto piki z minimalną wysokością i minimalnym odstępem (0.24s), a także zapisano czasy pików oraz zliczono liczbę wykrytych słów. Liczbę tę przedstawiono w sekcji wyniki, wraz z wykresami przedstawiającymi wyznaczone segmenty.

Metodyka i przebieg projektu – analiza i wpływ zmęczenia

Do przeanalizowania czasów wypowiadania każdego ze słów oraz przerw między nimi poprawiono manualnie, wcześniej wyznaczone granice między początkiem, a końcem wypowiadania słowa. Poprawki te naniesiono na poranne nagranie. Wieczorne granice wypowiadania słów zostały wyznaczone całkowicie

w sposób manualny. Dla obu nagrań posłużono się pierwszym fragmentem (os. 1, 52 słowa, ok. 30 sekund).

Po wyznaczeniu czasowych granic słów obliczono długość trwania procesu wypowiadania danego słowa (granica górna – granica dolna). Następnie wyznaczono różnicę między czasami analogicznych słów wieczornych i porannych (czas wieczorny – czas poranny), po czym sklasyfikowano te różnice na trzy grupy:

- dodatnie (dłużej wypowiadano słowo wieczorem) – 27 słów,
- ujemne (dłużej wypowiadano słowo o poranku) – 23 słowa,
- równe (taki sam czas wypowiadania słowa o poranku i wieczorem) – 2 słowa.

Ponadto wyświetlono również podstawowe statystki czasów, takie jak:

- minimalna różnica: -0.400s,
- maksymalna różnica: 0.290s,
- średnia różnica: 0.007s.

W kolejnym kroku przeanalizowano przerwy między słowami. Okazało się, iż przerwy NIE wydłużyły się wieczorem, a nawet nieznacznie się skróciły:

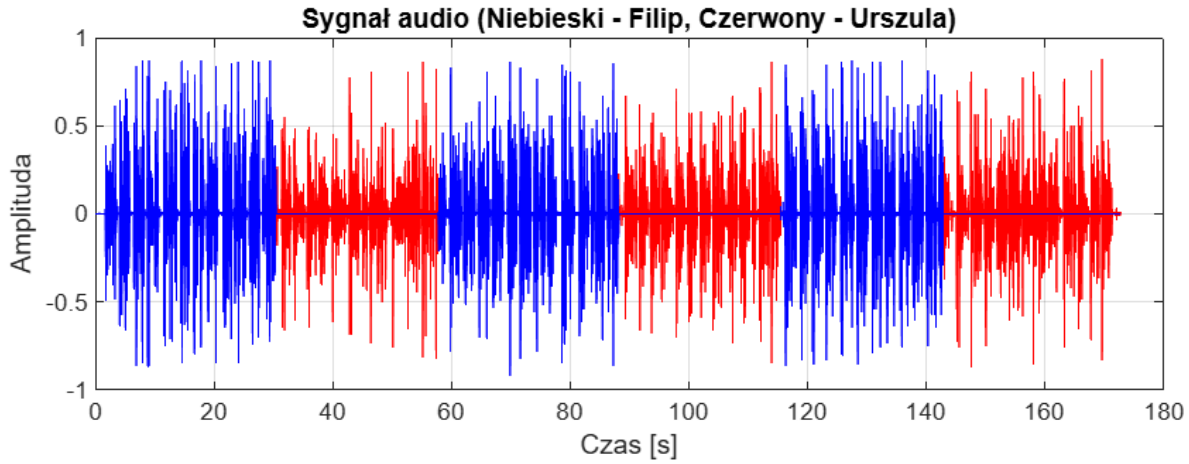
- średnia różnica przerw (średnia z różnic między czasami przerw: przerwy wieczorne minus przerwy poranne): -0.030s,
- minimalna różnica między przerwami: -0.3s,
- maksymalna różnica między przerwami: 0.2s.

Wyniki

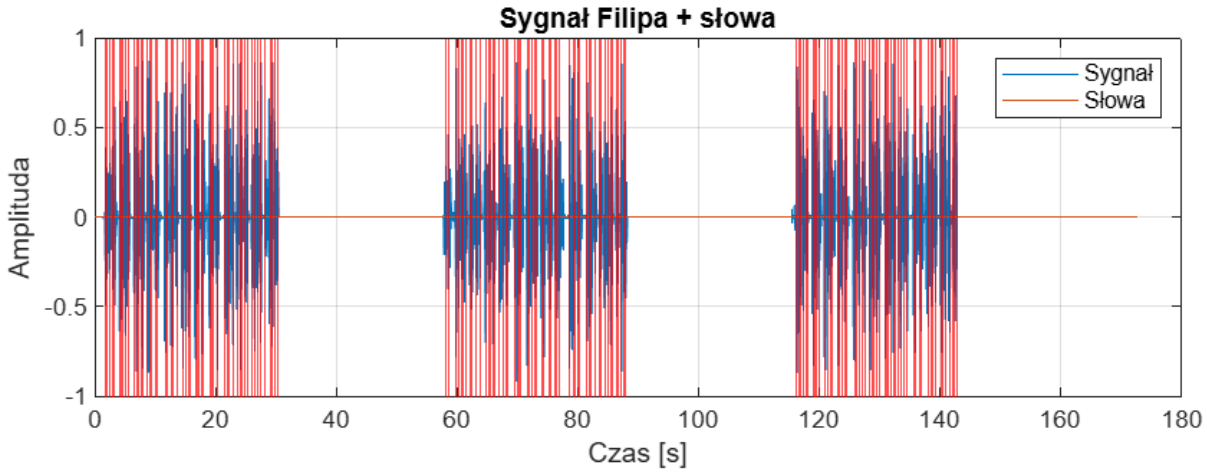
Na wykresie Wyk. 2 przedstawiono sygnał audio dla całego czasu okresu czasu trwania, wraz z podziałem na segmenty wypowiedziane przez os. 1 i os. 2. W tabeli Tab. 1 przedstawiono wyniki procesu detekcji ilości słów w każdym z trzech segmentów os. 1. Na Wyk. 3 zaprezentowano sygnały w segmentach os. 1 wraz z pionowymi kreskami określającymi granice czasów wypowiedzenia słów, które w sposób automatyczny zostały wyznaczone.

Numer fragmentu	Początek [s]	Koniec [s]	Ilość słów wypowiedzianych	Ilość słów obliczonych	Błąd [%]
1	0.00	30.40	52	56	7.69
2	57.76	88.16	55	56	1.82
3	115.52	142.88	53	54	1.89
Suma	-	-	160	166	3.75

Tab. 1 – Tabela z podsumowaniem detekcji słów w ramach segmentów

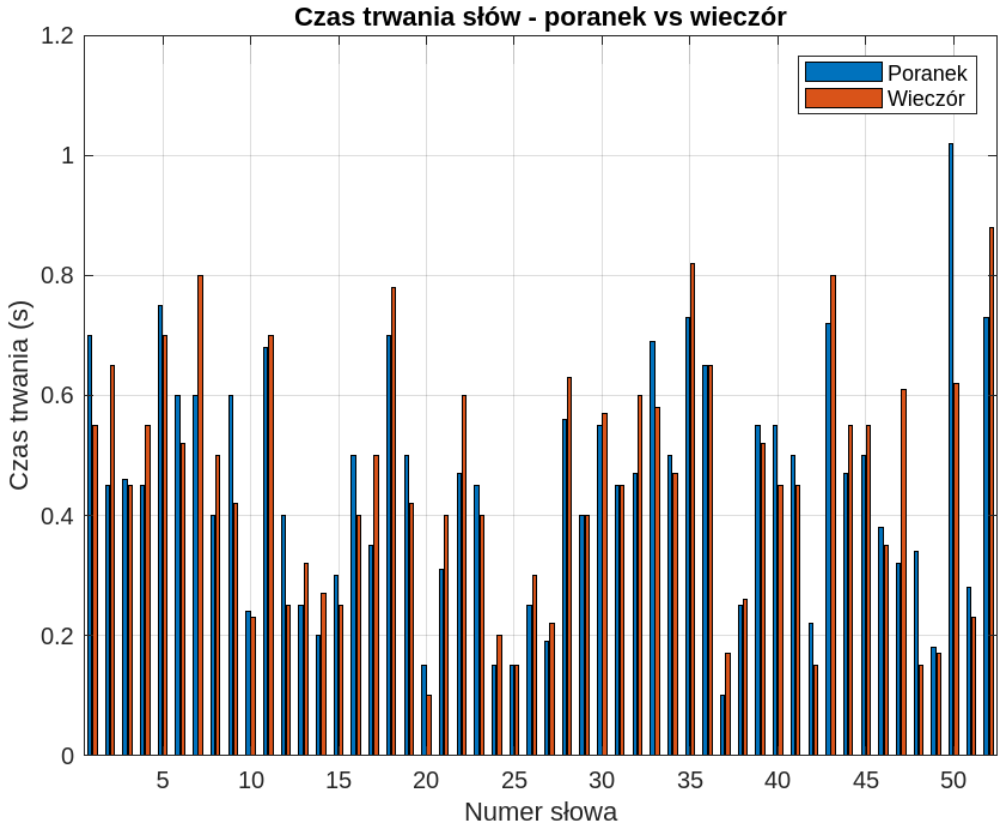


Wyk. 2 – Sygnał audio z podziałem na segmenty

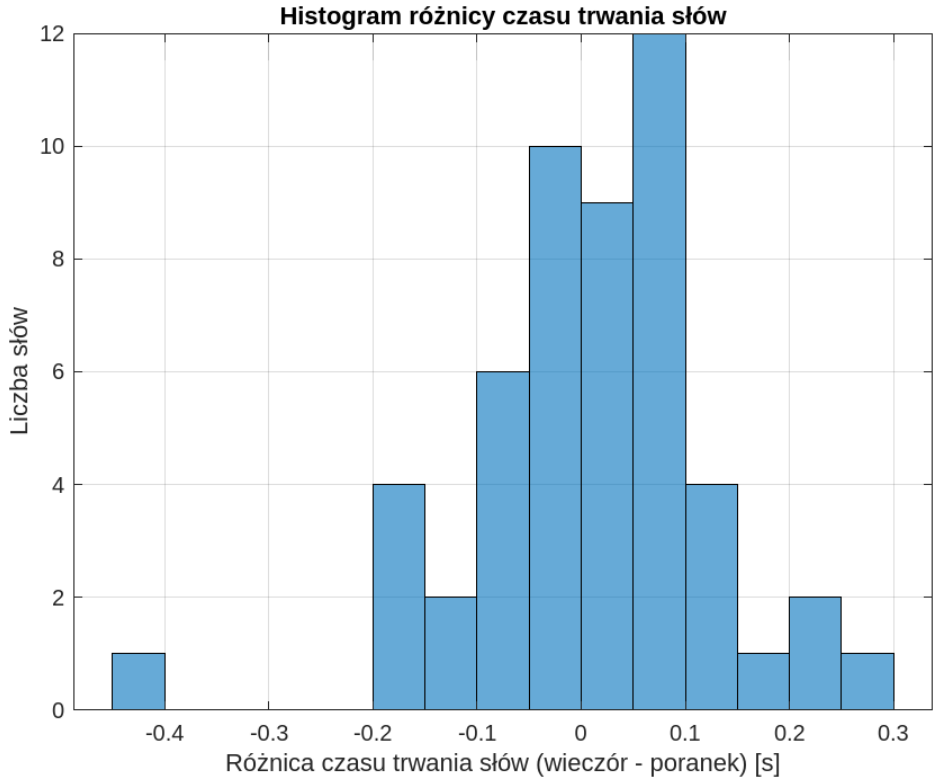


Wyk. 3 – Sygnał audio os. 1 z wyznaczonymi automatycznie słowami

Na kolejnych wykresach przedstawiono porównanie czasów wypowiadania słów między nagraniem porannym, a wieczornym. Na Wyk. 3 zaprezentowano wykres słupkowy z podziałem na czasy poranne i wieczorne. Natomiast na Wyk. 4 przedstawiono różnice między czasami wieczornymi, a porannymi w jednostce czasu (sekundy)

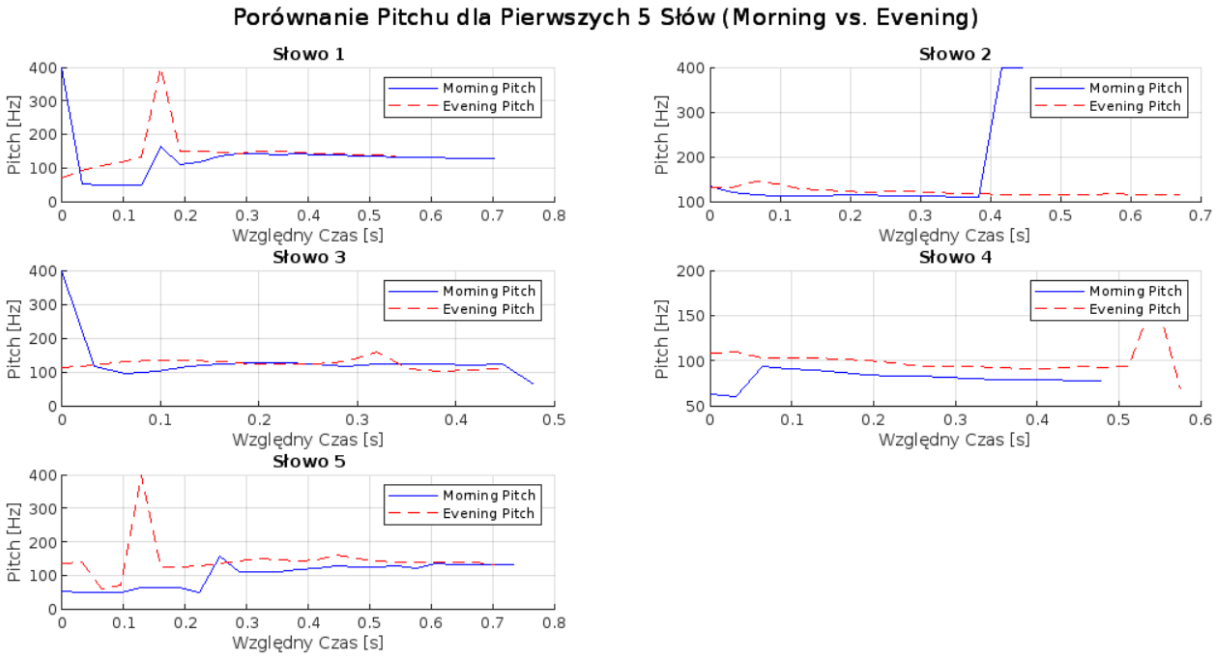


Wyk. 4 – Czas trwania słów – porównanie miedzy porankiem, a wieczorem



Wyk. 5 – Histogram z różnicą czasów trwania słów – porównanie między porankiem, a wieczorem

Na wykresie poniżej Wyk. 6 zaprezentowano wykresy liniowe tonacji dla 5 pierwszych słów z podziałem na poranek/wieczór.



Wyk. 6 – Porównanie tonacji dla nagrań porannego i wieczornego (pierwsze 5 słów)