**Sprawozdanie**

**Wstęp do multimediów (WMM) Laboratorium #3:   
Laboratorium – Automatyczne Rozpoznawanie oraz Synteza Mowy**

**Łukasz Szydlik 331446**

**Polecenia oraz kod:**

<https://colab.research.google.com/drive/11nujZpgUWw0dBVf5q5G4jiHo-mEeThze?usp=drive_link>

**Synteza Mowy**

**Subiektywna ocena po odsłuchu**

**Model TTS Google**

1. Zrozumiałość: bardzo wysoka – nie musiałem powtarzać odsłuchu.
2. Naturalność: lekko syntetyczny głos, ale dobre akcentowanie i płynna intonacja.
3. Błędy: brak przekłamań leksykalnych; interpunkcja odczytywana poprawnie.

**Model MMS TTS POL (Meta)**

1. Zrozumiałość: średnia – w szybkim tempie trudno odróżnić niektóre sylaby.
2. Naturalność: bardziej syntetyczny głos, z nierównym tempem i mniejszą ekspresją.
3. Błędy: zniekształcenia wyrazów, pomijane pauzy na przecinkach i kropkach.

**Analiza porównawcza spektrogramów**

A screenshot of a computer screen

AI-generated content may be incorrect.

**Nagranie oryginalne**

Pasmo niskie (0-300 Hz) - duża płynnie rozciągnięta energia, naturalne wahania głośności

Pasmo mowy ludzkiej (300-3 kHz) - płynne przejścia między głoskami

Pasmo wysokie (3-10 kHz) - energia spada powyżej 6 kHz

**Model TTS Google**

Pasmo niskie (0-300 Hz) – nie widać płynnych przejść, ale energia skupiona w tych pasmach

Pasmo mowy ludzkiej (300-3 kHz) - bardzo wyraźne, ostro zarysowane pasma; przerwy ciszy między frazami (widać czarne „dziury”)

Pasmo wysokie (3-10 kHz) - widoczne aż do 9–10 kHz – lepsza wyrazistość *s, sz, cz*

**Model MMS TTS POL (Meta)**

Pasmo niskie (0-300 Hz) – większość energii nie jest skupiona już w niższym paśmie

Pasmo mowy ludzkiej (300-3 kHz) - lepiej rozmyte niż w TTS Google, ale mniej płynności niż w oryginale

Pasmo wysokie (3-10 kHz) - płynnie rozmyte, ale dość dużo energii podczas całego nagrania

**Wnioski**

1. **Szerokość pasma**

- w Google TTS obecność energii aż do 10 kHz. Nadaje to głosowi klarowność, zwłaszcza spółgłoskom szczelinowym, ale może powodować wrażenie lekkiego „syczenia”.

- Meta TTS ma wysokie częstotliwości podobnie do oryginału.

1. **Rozkład energii**

**-** W nagraniu ludzkim amplituda wygląda naturalnie (samogłoski głośniejsze, spółgłoski cichsze).

**-** W obu syntezach energia jest równiej rozłożona, dlatego spektrogramy wyglądają jaśniej i „równiej”, lecz brzmienie staje się mniej naturalne.

1. **Struktura harmoniczna**– W oryginale harmoniczne są lekko rozmazane, bo w realnym głosie występują modulacje wysokości.  
   – *Google TTS* pokazuje niemal idealnie równoległe pasma, co zwiększa czytelność, ale odbiera naturalność.  
   – *Meta TTS* ma szersze, mniej kontrastowe harmoniczne, co daje efekt „miękkiego”, lecz mało klarownego brzmienia.
2. **Płynność i pauzy**- W oryginale między zdaniami występują oddechy i krótkie pauzy – spektrum zmierza do ciszy, ale nie „urwanej”.  
   *-* Google TTS wstawia niemal idealną ciszę (czarne pasy), co sprawia, że zdania brzmią wyraźnie rozdzielone, choć nieco sztucznie.  
   *-* Meta TTS rzadko „oddycha”, płynnie przechodzi między zdaniami – rytm jest gęstszy, co pogarsza rozumienie przy szybkim tempie.

**Ocena jakości syntezy mowy**

W celu oceny jakości syntezy mowy utworzono ankietę poprzez Formularze Google.

Odbiorcom podano skalę oceniania oraz udostępniono nagranie oryginalne oraz wygenerowane przez modele. Syntezę modelu MMS TTS POL (Meta)nazwano - Nagranie 1, a TTS Google – Nagranie 2

Link do ankiety: <https://forms.gle/1rPhGmTvJXPqGLqY8>

**Wnioski**

1. Zrozumiałość – Jak możemy zauważyć lepszą zrozumiałość osiąga model Google TSS
2. Płynność i Naturalność – parametry te mocno zależą od subiektywnej oceny doznań użytkownika
3. Zdecydowanie lepiej byłoby przeprowadzić ankietę na większej ilości odbiorców. Wtedy można by zapewnić lepszą ocenę płynności i naturalności podanych modeli.

**Automatyczne Rozpoznawanie Mowy**

**Rzeczywisty tekst:**

Szukajcie prawdy jasnego płomienia. Szukajcie nowych, nieodkrytych dróg. Za każdym krokiem w tajniki stworzenia

|  |  |
| --- | --- |
| Rozpoznawanie mowy | |
| Model | Rozpoznany tekst |
| Whisper base (OpenAI) | Szukajcie prawdy jasnego płomienia. Szukajcie nowych nieodkrytych dróg. Za każdym krokiem w tajniki stworzenia |
| MMS (Meta) | szukajcie prawdy jasnego płomienia szukajcie nowych nieodkrytych dróg za każdym krokiem w tajniki stworzenia |
| QuartzNet (NVIDIA) | szukajcie praw de jasnego płomienia szukajcie nowych nieotkrytych drók za każdym krokiem w tajnki stworzenia |
| FastConformer (NVIDIA) | Szukajcie prawdy jasnego płomienia. Szukajcie nowych, nie odkrytych dróg. Za każdym krokiem w tajniki stworzenia, |

**Wyniki WER i CER:**

|  |  |  |
| --- | --- | --- |
| Rozpoznawanie mowy | | |
| Model | WER | CER |
| Whisper base (OpenAI) | 7,14 | 0,9 |
| MMS (Meta) | 42,86 | 5,41 |
| QuartzNet (NVIDIA) | 71,43 | 9,91 |
| FastConformer (NVIDIA) | 21,43 | 1,8 |

**WER - Word Error Rate**

Odsetek błędnych słów w transkrypcji, 0 % = transkrypcja idealna.

**CER - Character Error Rate**

Pokazuje, jak bardzo transkrypcja odbiega literowo.

**Wnioski**

1. Whisper base zdecydowanie wygrywa – 7 % WER to wynik porównywalny z pisownią. Nie uwzględnił jedynie przecinka. Oznacza to, że model ma dobrą znajomość polskich słów.
2. MMS dobrze rozpoznaje same wyrazy, ale ma kłopoty z regułami pisowni: brak wielkich liter na początku zdań i niedokładna interpunkcja podwyższają zarówno WER, jak i CER.
3. QuartzNet – starsza, konwolucyjna architektura CTC prawdopodobnie trenowana na ograniczonych polskich zasobach; stąd najwyższy WER i CER. Ma problem z poprawnym wykryciem wyrazów oraz ich pisownią. Wyraźnie odstaje od nowszych modeli.
4. FastConformer wypada nieźle na poziomie liter, ale traci na poziomie słów. Dobrze radzi sobie z interpunkcją. Kwestia przecinka na końcu nagrania mogłaby być kwestią sporną. Z pozostałych błędów to tylko jedynie błąd pisowni; napisanie wyrazu ‘nieodkrytych’ rozłącznie.
5. Różnica między WER a CER u wszystkich modeli pokazuje, że część błędów to pojedyncze litery (CER ≪ WER), ale większy odsetek dotyczy całych słów, co wpływa mocno na WER.
6. Próba była krótka (trzy zdania) – przy dłuższych nagraniach wyniki mogą się lekko zmienić, jednak różnice między modelami powinny się utrzymać.