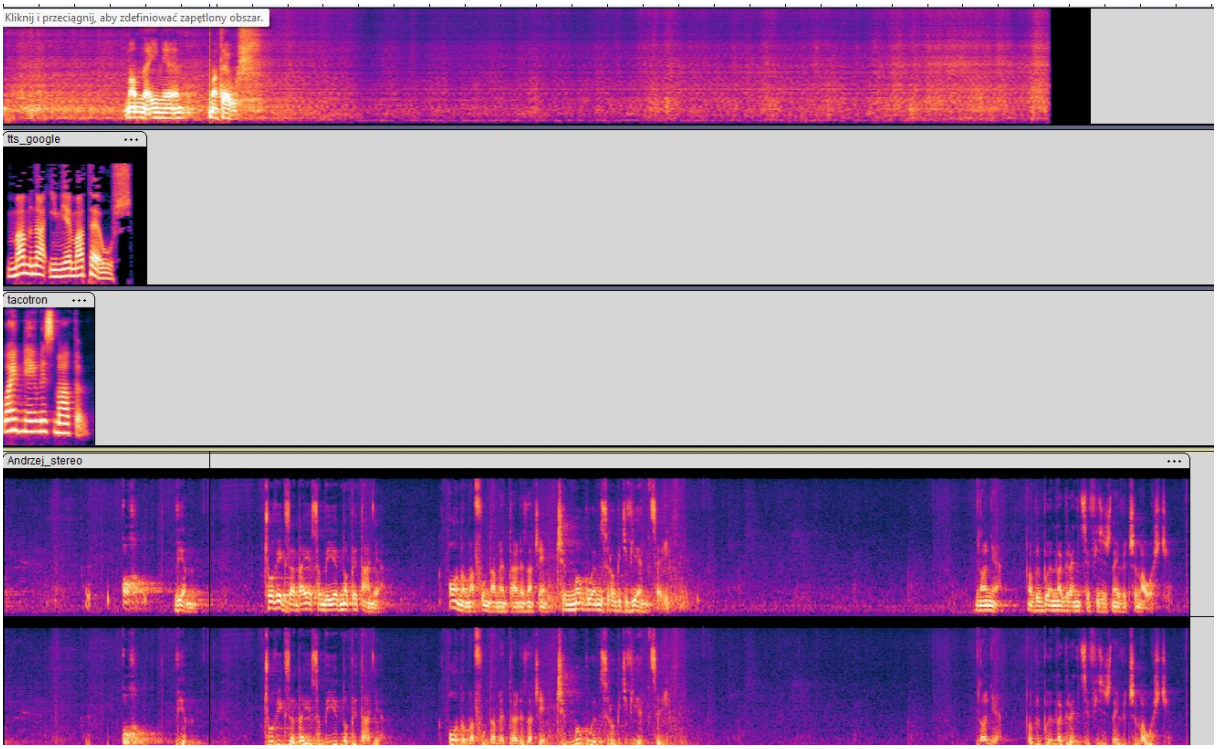
WMM laboratorium 3
Mateusz Ostaszewski
325203

Zad 1

Na podstawie własnych wrażeń słuchowych, subiektywnie oceniłem wyniki syntezy mowy trzech modeli. Najlepiej wypadł model Coqui.ai, gdyż jego wyjściowy dźwięk brzmi najbardziej ludzko. Jest to spowodowane poprawną wymową pojedynczych słów, dobrą dykcją oraz naturalnie brzmiącym głosem. W mojej opinii, drugim najlepszym modelem jest TTS Google. Jego główną wadą jest fakt, że głos wymawiający słowa brzmi zdecydowanie jak robot. Najgorzej wypadł model Tacotron2+Waveglow (NVIDIA). Największą jego wadą jest to, że uciął ostatnie słowo, co spowodowało, że wypowiedź stała się niezrozumiała. Dodatkowo, podobnie jak TTS Google, brzmi bardzo „robotycznie”.

Podsumowując, model Coqui.ai jest w stanie najbardziej naturalnie syntetyzować ludzką mowę spośród testowanych modeli.

Zad 2



Powyższe nagrania w kolejności:
- Coqui.ai
- TTS Google

- Tacotron2+Waveglow (NVIDIA)
- oryginalny

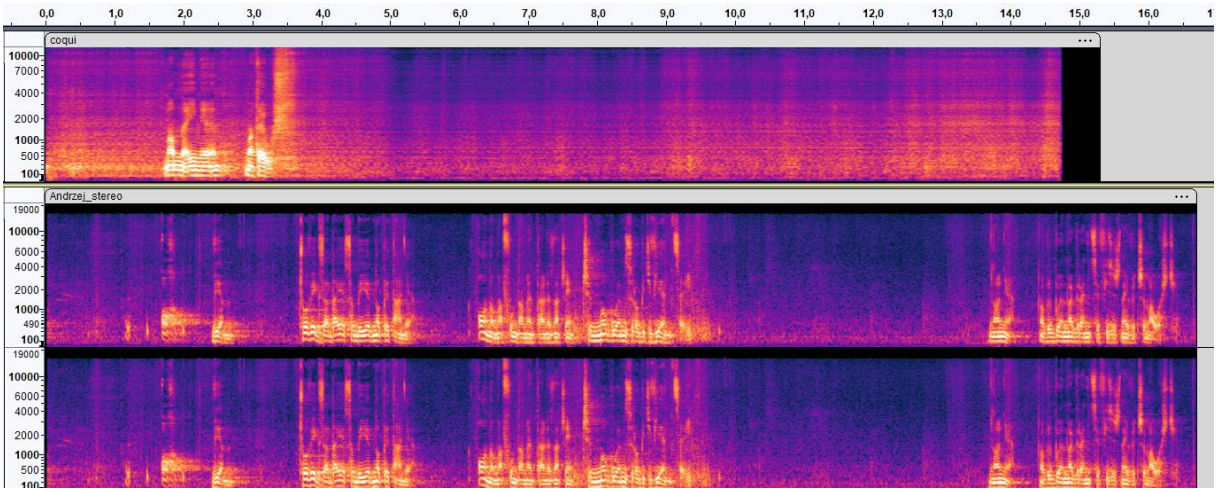
Oryginalne nagranie charakteryzuje się szerokim zakresem częstotliwości, obejmującym zarówno niskie, jak i wysokie pasma. Energia jest równomiernie rozłożona, co zapewnia pełne i bogate brzmienie.

Model Coqui.ai wykazuje podobne cechy do oryginalnego nagrania, oferując szeroki zakres częstotliwości i dobrze rozłożoną energię w pasmach. Syntezowany głos brzmi naturalnie.

Model TTS Google, choć dostarcza przyzwoitej jakości dźwięk, ma węższy zakres częstotliwości, z dominującymi wysokimi tonami. Energia w pasmach jest skoncentrowana głównie w wyższych częstotliwościach, co powoduje, że dźwięk brzmi bardziej „robotycznie” i mniej naturalnie.

Najgorsze wyniki uzyskał model Tacotron2+Waveglow (NVIDIA). Jego zakres częstotliwości jest węższy, z dominującymi średnimi częstotliwościami. Energia jest nierównomiernie rozłożona.

Zad 3



Zakres częstotliwości: Oba nagrania obejmują szeroki zakres częstotliwości, jednak syntezowany głos ma nieco inną charakterystykę rozłożenia energii. Energia w pasmach: Oryginalne nagranie Andrzeja ma bardziej równomiernie rozłożoną energię, co świadczy o pełnym i naturalnym brzmieniu mowy. Syntezowany głos ma miejscami mniejszą energię, co może wpływać na odbiór naturalności.

Warto zaznaczyć, że oryginalny tekst różni się od syntezowanego.

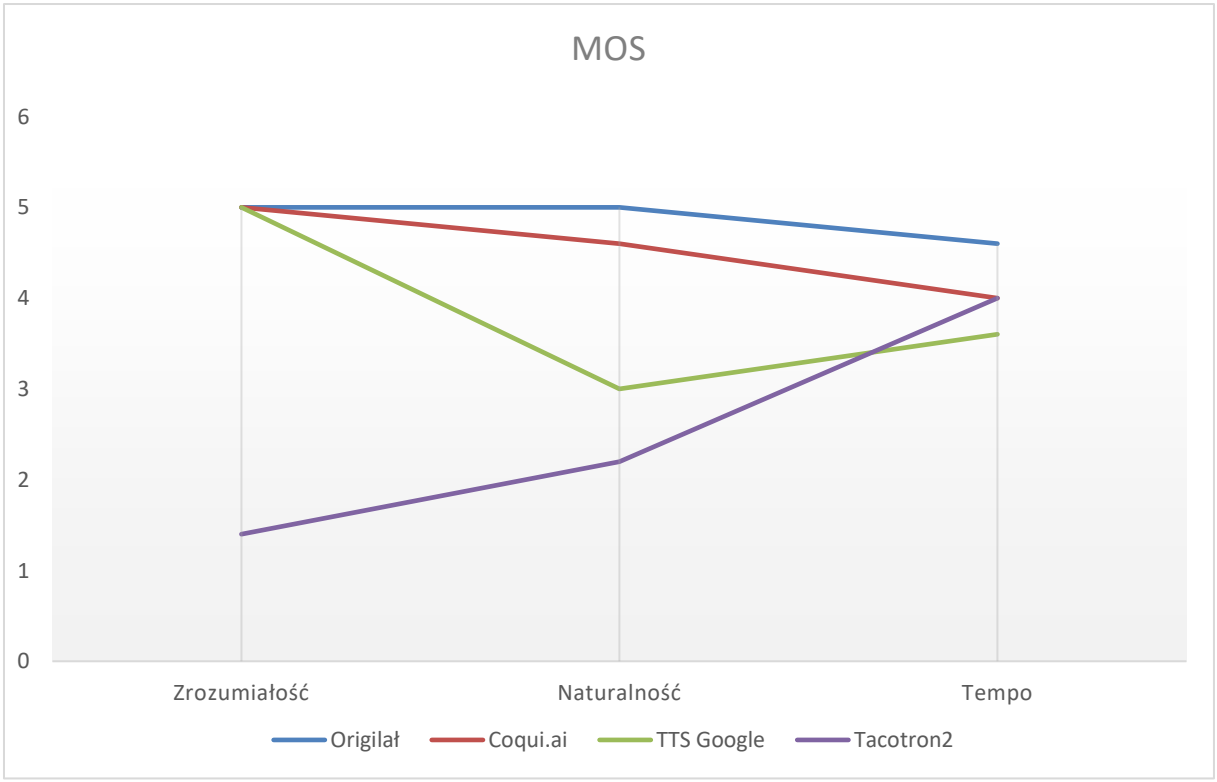
Zad 4

Oryginał			
Osoba	Zrozumiałość	Naturalność	Tempo
1	5	5	5
2	5	5	4
3	5	5	5
4	5	5	4
5	5	5	5

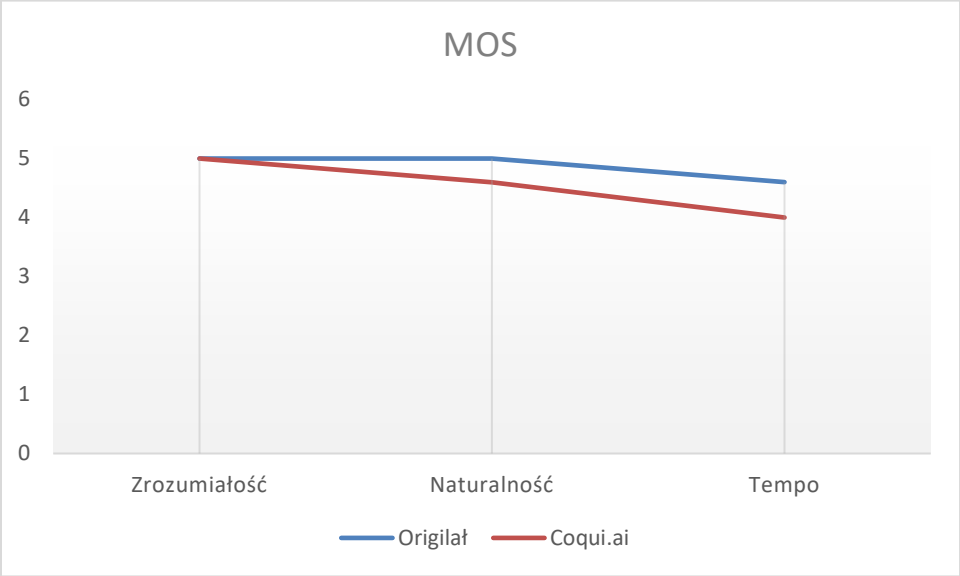
Coqui.ai			
Osoba	Zrozumiałość	Naturalność	Tempo
1	5	5	5
2	5	5	3
3	5	4	4
4	5	5	3
5	5	4	5

TTS Google			
Osoba	Zrozumiałość	Naturalność	Tempo
1	5	3	3
2	5	2	4
3	5	3	4
4	5	4	4
5	5	3	3

Tacotron2			
Osoba	Zrozumiałość	Naturalność	Tempo
1	1	2	4
2	2	1	5
3	1	3	4
4	1	2	4
5	2	3	3



Zad 5



Model Coqui.ai ma zbliżone wyniki do oryginalnego nagrania. Warto jednak zaznaczyć, że zdaje się mieć on lekki problem z dobraniem tempa mówienia.

Zad 6

Model	Rezultat
Oryginał	O, wiem. Człowiek zadaje sobie jedno pytanie. Ono ma fundamentalne znaczenie, czy mogłem zrobić więcej? No nie, no by byłem na granicy fizycznej wytrzymałości. Więcej by się nie dało zrobić.
Whisper (base)	O, wiem. Człowiek zadaje sobie jedno pytanie. Ono ma fundamentalne znaczenie, czy mogłem zrobić więcej? No nie, no by byłem na granicy fizycznej wytrzymałości. Między by się nie dało zrobić.
Wav2vec	oh wiem człowiek zadaje sobie jedno pytanie ono ma fundamentalne znaczenie czy mogłem zrobić więcejno nie o wybyłem nagranice fizycznej wytrzymałości więce by się nie dało zrobić
QuartzNet	oo wiem człowiek zadaje sobie jedno pytania kona mafu nament alne naczeń czy mogłym zrobić więcejno nie abybyłem ma granicy fizycznejo trzymałości więce by się nie dało zrobić
FastConformer	O. wiem, człowiek zadaje sobie jedno pytanie. Ono ma fundamentalne znaczenie, czy mogłem zrobić więcej. No, nie wybór byłem na granicy fizycznej wytrzymałości, więcej by się nie dało zrobić.

Zad 7

Model	WER	CER
Whisper (base)	3,33	2,11
Wav2vec	56,67	10,53
QuartzNet	70	18,42
FastConformer	33,33	6,84

Zad 8

Whisper (base): Model ten uzyskał najniższe wskaźniki WER (3,33) i CER (2,11), co wskazuje na jego wysoką precyzję i dokładność w transkrypcji mowy. Z analizy wyników tekstowych widać, że Whisper (base) niemal idealnie odwzorował oryginalne nagranie, z jedynie minimalnymi błędami.

Wav2vec: Ten model wykazał znacznie wyższe wskaźniki błędów, z WER wynoszącym 56,67 i CER 10,53. Choć model ten potrafił uchwycić pewne elementy wypowiedzi, błędy w rozpoznawaniu słów i składni znacznie obniżyły jego dokładność.

QuartzNet: Wyniki QuartzNet są najgorsze spośród testowanych modeli, z WER na poziomie 70 i CER wynoszącym 18,42. Model ten miał największe problemy z poprawnym rozpoznawaniem mowy, co skutkowało licznymi błędami i niezrozumiałymi fragmentami tekstu.

FastConformer: Model ten osiągnął umiarkowane wyniki, z WER wynoszącym 33,33 i CER 6,84. Chociaż nie jest tak dokładny jak Whisper (base), jest znacznie lepszy od Wav2vec i QuartzNet. Analiza tekstu pokazuje, że model ten dobrze radzi sobie z rozpoznawaniem struktury zdania, ale nadal ma problemy z dokładnością poszczególnych słów.