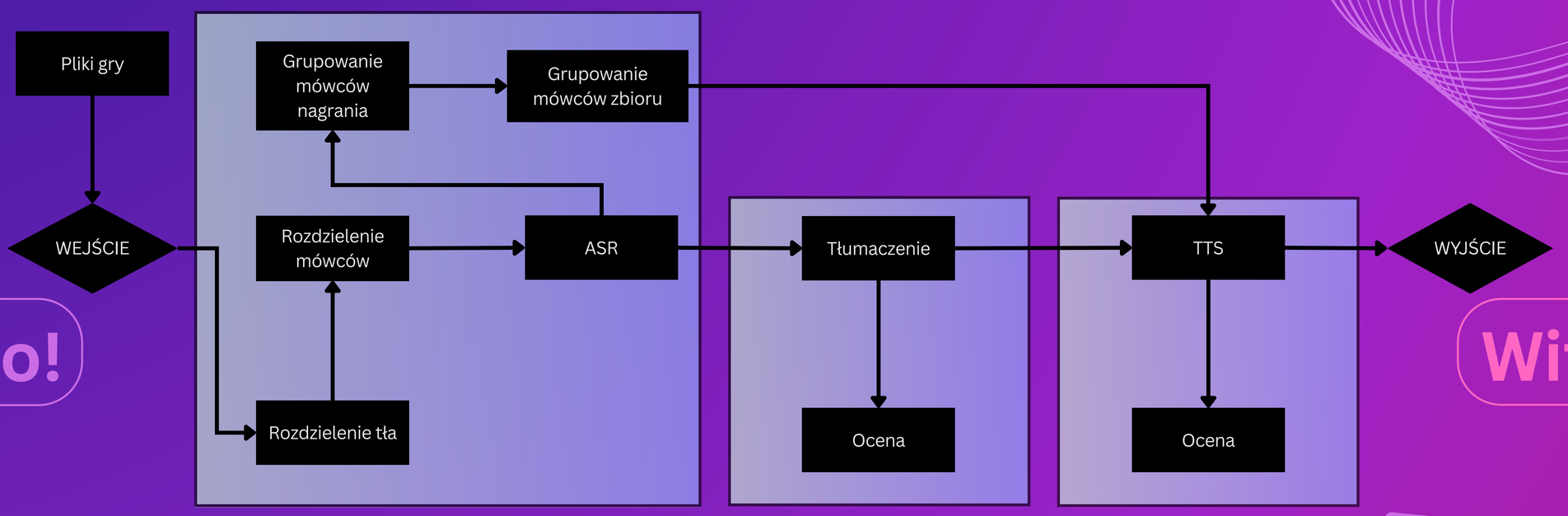
DubAI

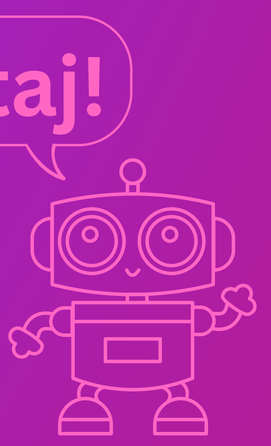
Cel projektu

Czytając recenzje gier komputerowych zauważyliśmy, że wiele użytkowników zwraca uwagę na **brak polskiego dubbingu**, co negatywnie wpływa na immersję. Jednocześnie tradycyjna lokalizacja głosowa wiąże się z wysokimi kosztami i długim czasem realizacji. Zmotywowani tymi wyzwaniami, postanowiliśmy stworzyć **system automatycznego generowania dubbingu** w języku polskim. W ramach projektu opracowaliśmy rozwiązanie, które w sposób szybki, prosty i przystępny kosztowo umożliwi lokalizację audio z zachowaniem naturalności i wysokiej jakości. Stworzony system jest narzędziem, które może wspierać zarówno **twórców indie,** jak i **większe studia.**

Dane

W ramach projektu wykorzystano próbki dialogów z gry **Wiedźmin 3: Dziki Gon**, stworzonej przez polskie studio CD Projekt RED. Tytuł ten słynie z rozbudowanej warstwy narracyjnej, setek unikalnych postaci i tysięcy nagranych kwestii dialogowych, co czyni go doskonałym materiałem do analizy systemów dubbingowych. Dzięki zastosowaniu danych z gry o tak wysokim standardzie produkcyjnym możliwa była wiarygodna symulacja warunków praktycznych, w których system generowania dubbingu mógłby być wykorzystywany.





Obsługa krótkich wypowiedzi oraz dialogów wymaga oczyszczenia i segmentacji nagrania. Pierwszym krokiem jest usunięcie dźwięków tła za pomocą modelu **Demucs**. Aby rozwiązać problem nakładających się głosów w dialogach przeprowadzana jest ich detekcja modelem od **Pyannote** i separacja przy użyciu modelu **Mossformer2**. Na oczyszczonym nagraniu przeprowadzana jest transkrypcja oraz segmentacja z użyciem **Whispera**. Otrzymane wypowiedzi pogrupowano przy użyciu algorytmu **DBSCAN**, co ułatwiło identyfikację spójnych segmentów jednego mówcy.

Do tłumaczenia wykorzystano polski wielojęzyczny model językowy **Bielik**, który pozwala na dużą swobodę przy tłumaczeniu. Tłumaczenie zostało wzbogacone o kontekst sąsiednich wypowiedzi oraz rozpoznanych płci rozmówców. Jakość tłumaczeń oceniano przy użyciu metryk **COMET**, **BERTScore, chrF** i **BLEU**, zapewniających obiektywną ocenę pod względem zgodności semantycznej i płynności.

Syntezę mowy zrealizowano przy użyciu modelu **XTTS-v2**, obsługującego 17 języków i umożliwiającego generowanie naturalnego głosu na podstawie jedynie 6-sekundowego nagrania referencyjnego. Został on wybrany ze względu na dobrą jakość dźwięku, szybkość działania oraz lepszą kompatybilność z językiem polskim niż inne dostępne modele. Na końcu przeprowadzono subiektywną ocenę końcowego dźwięku w formie **testów odsłuchowych**.

Tłumaczenie

Metoda	Metryki			
	BLEU	chrF	COMET	BERTScore
Bez kontekstu	0,112	0,425	0,821	**0,827**
Z kontekstem	**0,118**	**0,431**	**0,822**	**0,827**

Przykład 1

Dialog:

Troll - Ick metal. Mouth stings.
Lambert - **You nuts?!**
Geralt - Shut up and follow my lead.

Oryginalne: Zwariowałeś?!
Bez kontekstu: Ty orzeszku?!
Z kontekstem: Czy ty zwariowałeś?!

Przykład 2

Dialog:

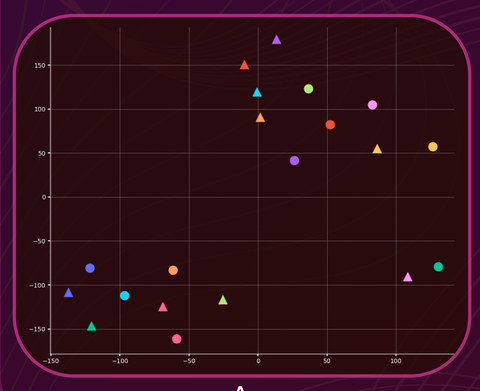
Yennefer - Philippa, your eyesight, only just recovered and magically simulated. Didn't you say you'd need some time to get accustomed?
Philippa Eilhart - **Did I?**
Yennefer - I'd forgotten how irritating she can be. Come, Ciri.

Oryginalne: Tak mówiłam?
Bez kontekstu: Czy ja?
Z kontekstem: Czy ja tak powiedziałam?

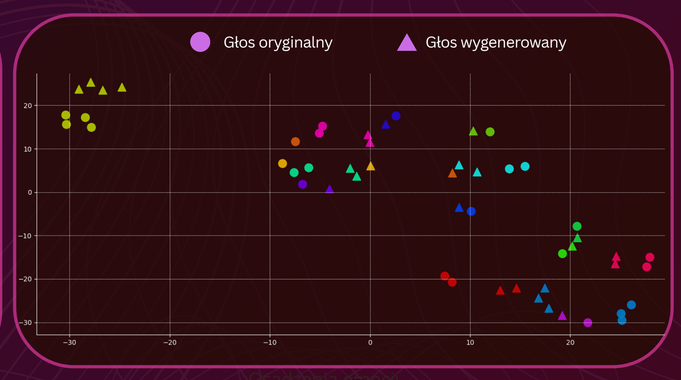
TTS

	Wskaźnik błędów słownych	Podobieństwo cosinusowe	Zgodność emocji
EN → PL	**14,81%**	**0,60**	**38%**
PL → EN	**6,13%**	**0,87**	**32%**

Wykresy przedstawiają **wizualizacje wektorów osadzeń** dla emocji (wykres A) oraz cech mówców (wykres B).



A



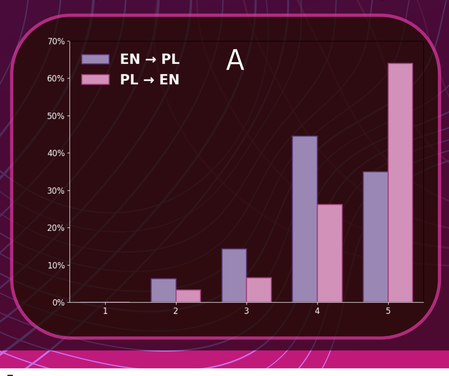
B

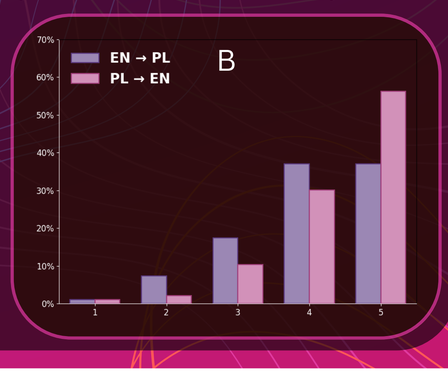
MOS (Mean Opinion Score)

- 63 respondentów
- sMOS - transfer głosu
- nMOS - naturalność

	EN → PL	PL → EN
sMOS	**4,08**	4,51
nMOS	**4,02**	**4,38**

Histogram ocen transferu cech głosu (wykres A) oraz naturalności (wykres B)





Podsumowanie

W ramach projektu zbudowano **system do automatycznego generowania dubbingu** na potrzeby gier komputerowych. System osiąga satysfakcjonujące rezultaty w języku angielskim, zarówno pod względem jakości dźwięku, jak i naturalności wypowiedzi. Nieco słabiej wypadają wyniki dla języka polskiego, co wynika głównie z ograniczonej dostępności wysokiej jakości danych treningowych oraz większej złożoności fonetycznej i fleksyjnej języka. Projekt potwierdził potencjał technologii syntezy mowy w kontekście przemysłu gier i otwiera dalsze możliwości rozwoju, w tym rozbudowę systemu o kolejne języki i lepsze dopasowanie do stylu oraz kontekstu narracyjnego.

Autorzy:
Klaudia Janicka - 262268@student.pwr.edu.pl
Natalia Iwańska - 262270@student.pwr.edu.pl
Kajetan Kołodziejczyk - 259171@student.pwr.edu.pl
Michał Wiktorowski - 262330@student.pwr.edu.pl
Wiktor Jeżowski - 260426@student.pwr.edu.pl

Opiekunowie projektu:
dr hab. inż. Maciej Piasecki
dr inż. Piotr Syga

Katedra Sztucznej Inteligencji
Politechnika Wrocławska