Faza 4: rozpoznawanie mowy

> Typ: procedura · Stan: zaliczona 2 października 2026 · Wymaga: zaliczona faza 3

Dokładamy ucho: mikrofon, nagranie i zamianę mowy na tekst. Używamy faster-whisper z modelem large-v3-turbo, uruchomionym jako mała lokalna usługa HTTP, z której korzysta czat głosowy i później okno awatara.

Skróty użyte na tej stronie (20)

Skrót	Po angielsku	Po polsku i co znaczy
AI	Artificial Intelligence	sztuczna inteligencja: programy, które naśladują część ludzkiego rozumowania i rozmawiają w języku naturalnym
API	Application Programming Interface	interfejs programistyczny aplikacji: sposób, w jaki programy rozmawiają ze sobą
CPU	Central Processing Unit	procesor centralny: główny procesor komputera
CUDA	Compute Unified Device Architecture	platforma firmy NVIDIA do obliczeń na jej kartach graficznych; od niej zależą sterownik i szybkość modeli
DLL	Dynamic-Link Library	biblioteka łączona dynamicznie: plik z kodem ładowany przez programy Windows
GB	Gigabyte	gigabajt: miliard bajtów, jednostka rozmiaru plików i pamięci
GET	HTTP GET method	metoda HTTP: prosi serwer o odczyt zasobu bez zmieniania czegokolwiek
GPU	Graphics Processing Unit	procesor graficzny, czyli karta graficzna: liczy modele językowe wielokrotnie szybciej niż zwykły procesor
HTTP	Hypertext Transfer Protocol	protokół przesyłania hipertekstu: język, którym przeglądarki i programy rozmawiają z serwerami WWW
IT	Information Technology	technologia informacyjna: informatyka, dział albo praca związana z komputerami i infrastrukturą
LLM	Large Language Model	duży model językowy: program przewidujący kolejne fragmenty tekstu, na którym opiera się czat AI

Skrót	Po angielsku	Po polsku i co znaczy
POST	HTTP POST method	metoda HTTP: wysyła dane do serwera, na przykład żeby utworzyć nowy zasób
Q4_K_M	4-bit K-quant, medium	kwantyzacja czterobitowa typu K w wariancie średnim: kompromis między rozmiarem pliku a jakością
Q5	5-bit quantization	kwantyzacja pięciobitowa: zapis parametrów modelu na około 5 bitach, większy plik i zwykle lepsza jakość niż Q4
STT	Speech-to-Text	zamiana mowy na tekst: rozpoznawanie mowy
tok/s	tokens per second	tokenów na sekundę: miara szybkości, z jaką model generuje odpowiedź
VAD	Voice Activity Detection	wykrywanie aktywności głosowej: filtr, który odróżnia mowę od ciszy
VR	Virtual Reality	wirtualna rzeczywistość
VRAM	Video Random Access Memory	pamięć karty graficznej: w niej musi zmieścić się model, żeby liczył się na karcie
WAV	Waveform Audio File Format	format pliku dźwiękowego bez kompresji

□□ Leksykon: nazwy i pojęcia z tej strony (11)

Nazwa	Co to jest
awatar VRM	Postać 3D, którą widzimy podczas rozmowy; VRM to format pliku takiej postaci, z gotowymi wyrazami twarzy i ruchem ust.
Bielik	Polski model językowy rozwijany przez SpeakLeash i ACK Cyfronet. U nas Bielik-Minitron-7B w wersji v3 jest modelem rozmowy.
czat	Rozmowa z modelem, w której on tylko odpowiada tekstem. Sam niczego nie uruchamia ani nie zmienia, dopóki nie damy mu narzędzi.
kwantyzacja	Zapisanie wag modelu na mniejszej liczbie bitów, żeby zajmował mniej pamięci i działał szybciej, kosztem niewielkiej utraty jakości. Oznaczenia Q4 i Q5 mówią z grubsza, ile bitów przypada na jedną wagę.
model językowy	Program wyuczony na dużych zbiorach tekstu, który dopisuje kolejne fragmenty tekstu pasujące do tego, co już jest. Rozmowa z nim to ciągłe dopisywanie, a nie szukanie faktów w bazie.
offline	Praca bez połączenia z internetem: model, głos i rozpoznawanie mowy działają na naszym komputerze. Test offline sprawdzamy, zanim cokolwiek podłączymy do sieci.
pip i PyPI	pip to instalator pakietów Pythona, a PyPI to publiczne składowisko tych pakietów.
Python	Język programowania, w którym piszemy klienty, serwery MCP i skrypty.

Nazwa	Co to jest
STT i TTS	Dwie połowy rozmowy głosowej: STT zamienia mowę na tekst, TTS zamienia tekst na mowę.
test akceptacyjny	Sprawdzenie opisane z góry, które faza musi przejść, żeby uznać ją za skończoną.
Whisper	Model rozpoznawania mowy firmy OpenAI. U nas działa jako faster-whisper z modelem large-v3-turbo, w małej lokalnej usłudze.

Zanim zaczniemy

Plan zakładał Whisper na procesorze, żeby nie ruszać VRAM. W praktyce i7-6700K ma tylko 4 rdzenie, a duży model na CPU liczy wypowiedź zbyt długo jak na rozmowę. Na GPU large-v3-turbo w trybie int8_float16 zajmuje około 0,35 GB VRAM i przepisuje 8 sekund mowy w około 0,5 sekundy.

Ta decyzja kosztowała coś w innym miejscu. Bielik Q5 razem z Whisperem dociskał kartę do granicy i prędkość spadła z około 70 do 40 tok/s. Wróciliśmy do Bielika Q4_K_M, który obok Whispera generuje około 89 tok/s.

Element	Wersja	Gdzie
faster-whisper	1.2.1	`C:\AI\stt\venv` (osobne środowisko, Python 3.12)
model	large-v3-turbo, `cuda`, `int8_float16`	`C:\AI\models\whisper`
biblioteki CUDA	nvidia-cublas-cu12 12.9.2.10, nvidia-cudnn-cu12 9.27.0.42	pakiety pip, katalogi DLL dodawane w kodzie
usługa	`C:\AI\stt\stt_server.py`	tylko `127.0.0.1:8770`

Krok 1: usługa STT

Usługa ładuje model raz i trzyma go w pamięci, więc kolejne wypowiedzi nie czekają na ładowanie. Słucha wyłącznie na adresie lokalnym i nie ma żadnych innych funkcji poza przepisaniem dźwięku.

- `GET /health` zwraca stan i nazwę modelu.
- `POST /stt` przyjmuje WAV 16 kHz, mono, 16 bit i zwraca `{text, audio_s, czas_s}`.
- Język jest ustawiony na polski, włączony jest filtr ciszy (VAD), a `initial_prompt` podpowiada słownictwo IT.

```
& C:\AI\stt\venv\Scripts\python.exe C:\AI\stt\stt_server.py
Invoke-RestMethod http://127.0.0.1:8770/health
```

Krok 2: mikrofon w czacie głosowym

Czat z fazy 3 dostał flagę `--mikrofon`. Enter zaczyna i kończy nagranie, a nagrania krótsze niż 0,4 sekundy są odrzucane. Flaga `--urzadzenie` wybiera konkretny mikrofon, bo domyślnym urządzeniem okazało się ciche wirtualne wejście gogli VR.

```
& C:\AI\tts\venv\Scripts\python.exe C:\AI\tts\chat_tts.py --mikrofon
```

Problemy i rozwiązania

Objaw	Przyczyna	Rozwiązanie
`TypeError: ... metadata_errors` przy transkrypcji	PyAV 17 zmienił API, z którego faster-whisper czyta pliki	podajemy dźwięk jako tablicę numpy, bez dekodowania przez PyAV
brak tekstu, cisza w nagraniu	domyślny mikrofon był urządzeniem wirtualnym	wybór urządzenia flagą, odrzucanie cichych nagrań
spadek LLM z 70 do 40 tok/s	Bielik Q5 i Whisper razem na 8 GB	Bielik Q4_K_M, około 89 tok/s
błąd ładowania cuBLAS lub cuDNN	DLL z pakietów pip nie są w PATH	funkcja `dll_cuda()` dopisuje ich katalogi przed importem

Test akceptacyjny

Odłączamy sieć, mówimy pytanie do mikrofonu i słyszymy odpowiedź. Tekst rozpoznany przez Whispera pokazuje się przed odpowiedzią, więc od razu widać, czy błąd jest w uchu, czy w modelu. Test przeszedł: rozpoznanie poprawne, model i usługa STT razem mieszczą się w VRAM (około 6,75 GB z 8 GB).

Następny krok: Fazy 5 i 6: awatar i pełna rozmowa offline.

Revision #4
Created 2026-10-01 23:24:15 UTC by Tomasz Konieczny
Updated 2026-10-20 09:05:55 UTC by Tomasz Konieczny