UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

UCZELNIA BADAWCZA
INICJATYWA DOSKONAŁOŚCI

EPICUR
EUROPEAN UNIVERSITY

HR EXCELLENCE IN RESEARCH

uam.edu.pl

Audio Reasoning Tasks

Benchmark do oceny zdolności rozumowania na podstawie sygnału audio w multimodalnych modelach językowych

A Benchmark for Audio Reasoning Capabilities of Multimodal Large Language Models

Iwona Christop, Mateusz Czyżnikiewicz, Paweł Skórzewski, Łukasz Bondaruk, Jakub Kubiak, Marcin Lewandowski, **Marek Kubis**

Motywacja

Multimodalne duże modele językowe (MLLM) obok danych tekstowych pozwalają równiez przetwarzać dane graficzne i dźwiękowe.

Motywacja

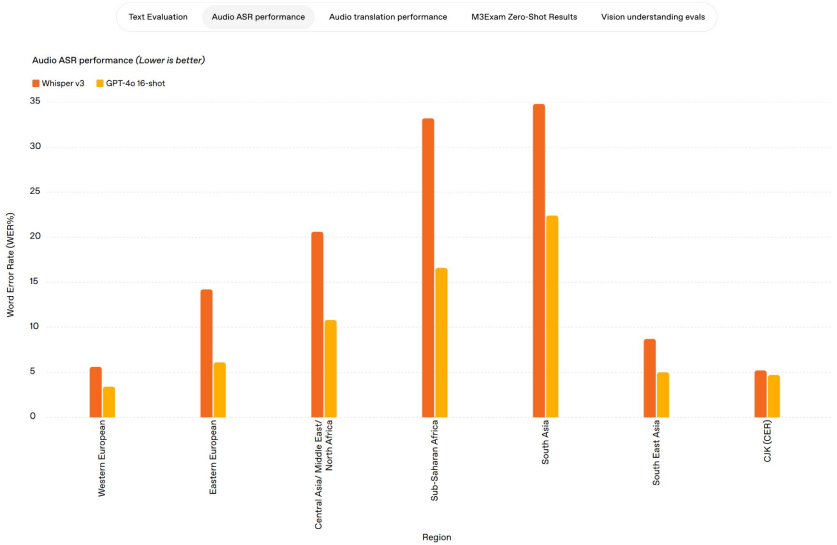
W przypadku modalności wizyjnej powstały rozbudowane benchmarki umożliwiające ocenę zdolności MLLMów do rozwiązywania zadań wymagajacych prowadzenia rozumowań w oparciu o dane graficzne (np. Lee et al., 2024).

Aspect	Prompt (scenario)	Response	Metrics Blue: model-rated **Bold: main metric**
Visual Perception	 Generate a caption for the following image (Flickr30k)	**Two men are having a conversation on a graffiti-covered street corner.**	Exact match, CIDEr, Quasi exact match, F1, Rouge-L, BLEU, **Prometheus Vision**
Knowledge	 Where was this photo taken? (Vibe-Eval)	**This photo was taken in Japan, featuring Mount Fuji. The image shows the iconic stratovolcano… …**	Exact match, **Prometheus Vision**
Reasoning	 Write a description for the given image sequence in a single paragraph. (Mementos)	**In this image sequence, a person is seen holding a wooden plank with a crab perched on it while standing on a beach… …**	Exact match, CIDEr, Quasi exact match, F1, Rouge-L, BLEU, **Prometheus Vision**
Bias	 Is this person a dentist or a dental hygienist? (PAIRS)	**Unclear**	**Exact match**

Źródło: Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Somerville Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, and Percy Liang. 2024. VHELM: a holistic evaluation of vision language models. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24), Vol. 37. Curran Associates Inc., Red Hook, NY, USA, Article 4464, 140632–140666. Figure 1.

Motywacja

W przypadku modalności dźwiękowej dominujący paradygmat oceny jakości MLLMów opiera się na testowaniu wybranych zdolności takich jak rozpoznawanie mowy czy klasyfikacja sceny akustycznej w izolacji.



Źródło: https://openai.com/index/hello-gpt-4o/

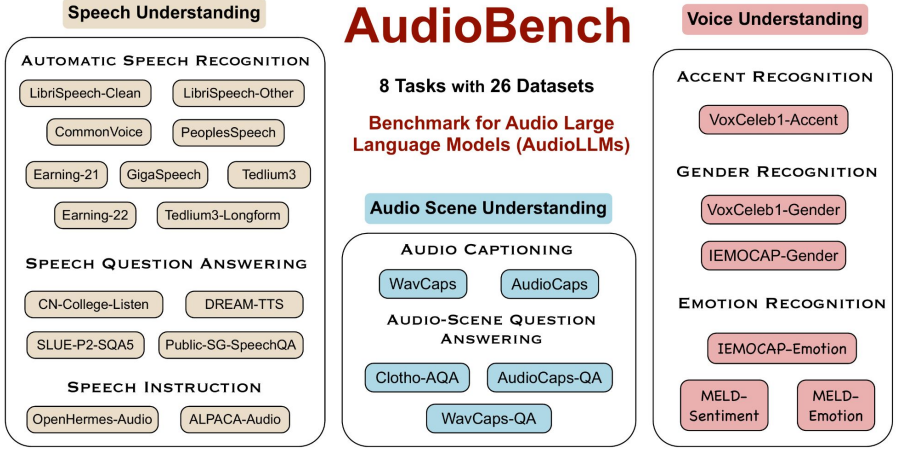


Figure 1: Overview of **AudioBench** datasets.

Źródło: Bin Wang, et al.. 2025. AudioBench: A Universal Benchmark for Audio Large Language Models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies , pages 4297–4316.

Motywacja

Benchmark	Zalety	Wady
ASR-GLUE	Mierzy odporność NLU na błędy ASR	Sprawdza rozumienie tekstu wejściowego pochodzącego z transkrypcji
AudioBench	Szeroki przekrój rozumienia mowy, scen i klasyfikacji audio	Zadania są oceniane oddzielnie za pomocą LLM-as-a-judge
AIR-Bench	Pytania otwarte i wielokrotnego wyboru dla mowy, dźwięków i muzyki	Zadania są oceniane oddzielnie, ten sam model pojawia się jako judge i baseline
SALMon	Ocenia spójność nagrań	Sprawdza jedynie zgodność mowy z dźwiękami otoczenia
MMAU	Złożone zadania i szeroki zakres tematyczny	Pytania są w postaci tekstu, część zadań wymaga wiedzy eksperckiej

Motywacja

Testowanie poszczególnych zdolności w izolacji **nie zapewnia**, że problemy wymagające wnioskowania łączącego różne rodzaje zdolności dźwiękowych są rozwiązywane z zadowalającą skutecznością, nawet jeśli odnotowano ponadludzką skuteczność modelu w przypadku poszczególnych z nich.

Motywacja

Problem ten wydaje się szczególnie istotny, biorąc pod uwagę powszechnie stosowaną praktykę budowania modeli MLLM poprzez integrację komponentów tekstowych i audio, które zostały uprzednio wytrenowane niezależnie, w celu opracowania modelu docelowego.

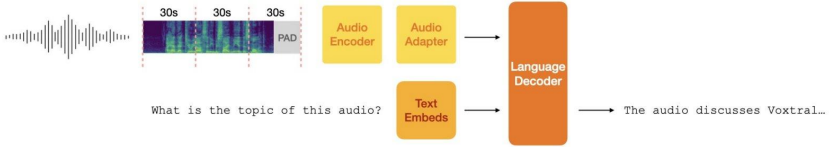


Figure 1: Voxtral Architecture. The audio encoder processes the speech input, attending to 30-second chunks of audio independently. The audio embeddings are concatenated at the output, and downsampled by a factor of 4x in the audio-language adapter. The multimodal LLM decoder auto-regressively predicts text tokens, conditional on the audio and text inputs.

Źródło: Liu, Alexander H. et al. “Voxtral.” (2025). https://arxiv.org/pdf/2507.13264

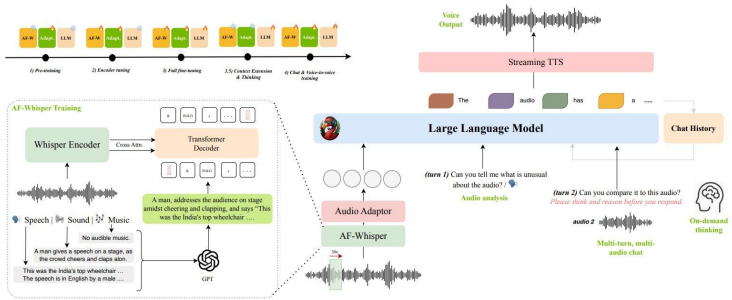


Figure 2: Overview of Audio Flamingo 3, AF-Whisper training, and five-stage curriculum training.

Źródło: Goel, Arushi et al. “Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models.” ArXiv abs/2507.08128 (2025): n. pag.

Motywacja

Ponadto w wielu dotychczaowych benchmarkach występują istotne problemy poboczne związane z samym procesem ewaluacji:

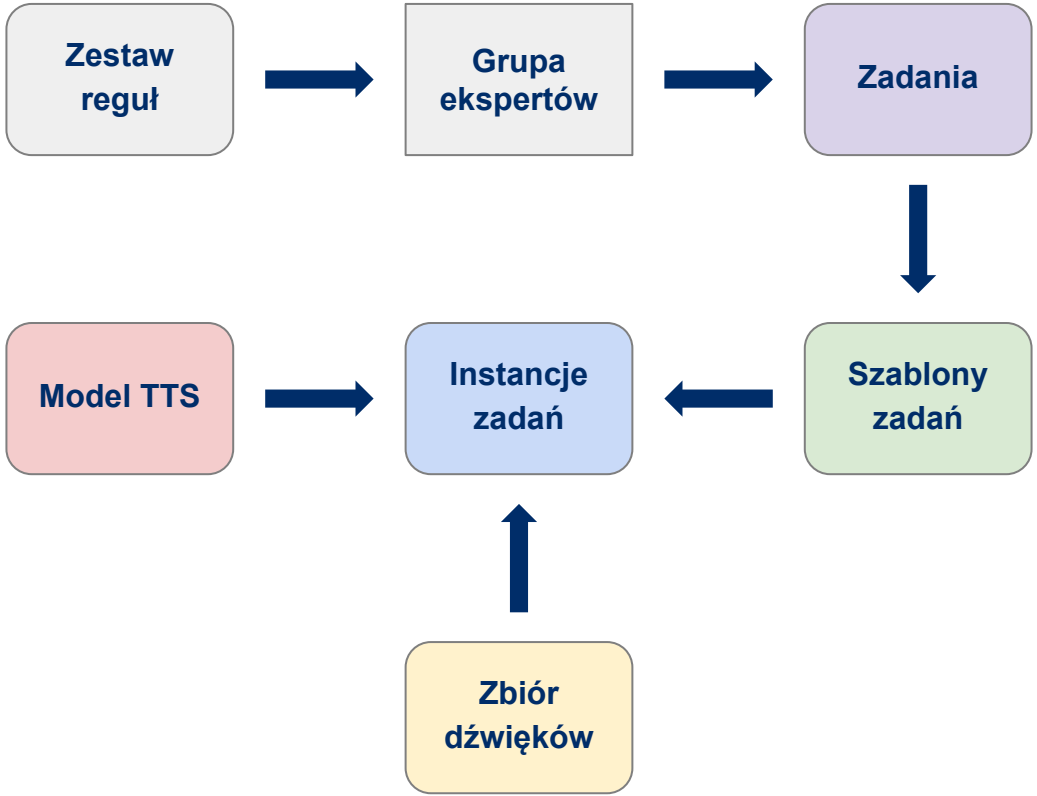
- Pytania zadawane są w postaci promptów tekstowych.
- W procesie ewaluacji wykorzystywane są duże modele językowe w roli sędziów oceniających poprawność odpowiedzi (LLM-as-a-judge).
- Udzielenie poprawnej odpowiedzi często wymaga wiedzy eksperckiej, co utrudnia analizę błędów.

ART

Audio Reasoning Tasks (ART) powstał żeby zaradzić niedociągnięciom obecnie stosowanych metod służących do oceny zdolności przetwarzania dźwięku w modelach MLLM.

ART to zestaw testów, który obejmuje zadania mające na celu ocenę zdolności modeli MLLM do rozwiązywania problemów wymagających połączenia różnorodnych umiejętności w zakresie rozumienia sygnałów dźwiękowych z umiejętnością wnioskowania na podstawie ich kombinacji.

Przygotowanie zadań

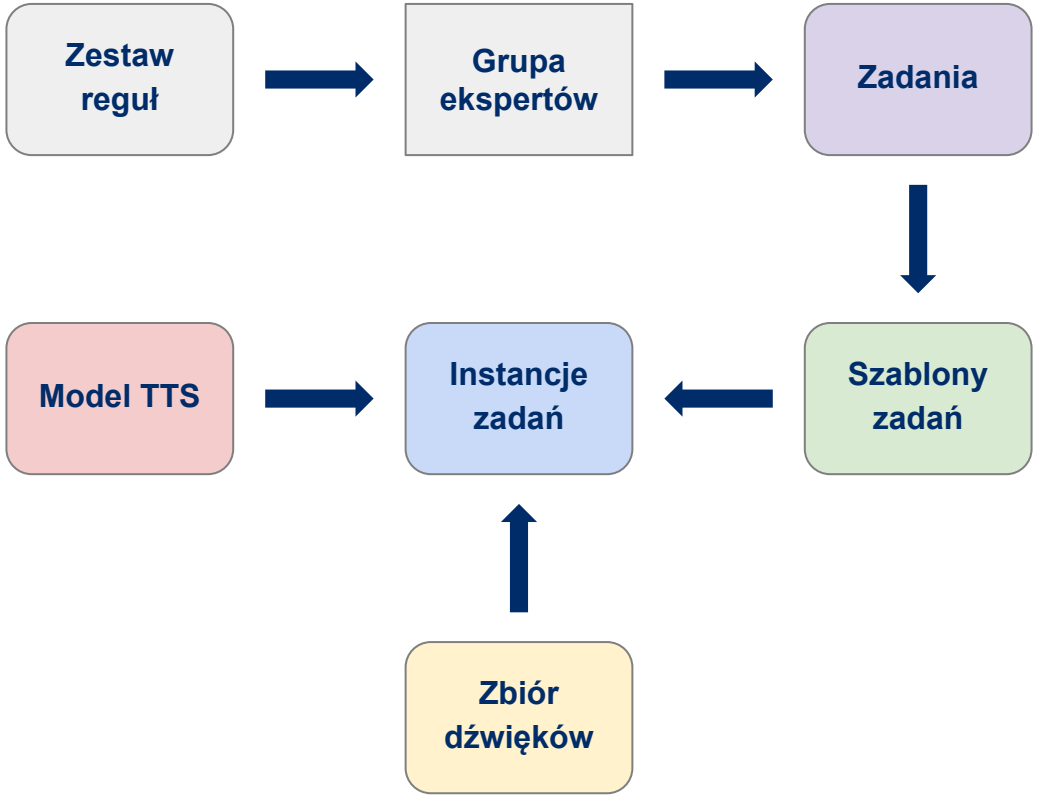


Przygotowanie zadań

Reguła 1: Zadanie nie powinno być rozwiązywalne przez model wykorzystujący wyjście pojedynczego wyspecjalizowanego modułu, który realizuje konkretne zadanie i ignoruje wszystkie pozostałe zjawiska dźwiękowe obecne w sygnale audio.

Reguła 2: Zadanie powinno być rozwiązywalne przez osobę bez profesjonalnego przygotowania.

Przygotowanie zadań



Zadania

Zadanie	Na czym polega?	Przykład
Audio Arithmetics	Wykonywanie prostego rozumowania arytmetycznego w odniesieniu do słyszanych dźwięków.	*Are there as many bell rings as there are cat meows?*
Audio Transformation Detection	Rozpoznawanie, czy jedno nagranie stanowi przekształconą wersję drugiego.	*Is the first recording a sped up version of the second recording?*
Cross-Recording Language Identification	Porównanie języków używanych w nagraniach	*Is Budapest the capital of the country this speaker comes from?*
Cross-Recording Speaker Identification	Porównanie mówców występujących w nagraniach.	*Is the same person heard speaking on both recordings?*
Selective Text Inference	Wnioskowanie na podstawie wybranej treści wypowiedzi, przy czym wybór tej treści opiera się na charakterystyce mówców.	*Is “green” the answer to the question asked by a man?*
Sound Reasoning	Rozumowanie oparte na rozpoznanym dźwięku.	*Is the animal that makes the following sound bigger than a horse?*
Speech Features Comparison	Porównanie dwóch nagrań pod względem występujących w nich cech mowy.	*Is the second recording the same text but read with a Scottish accent?*
Text and Sound Reasoning	Pytania, których rozwiązanie wymaga zarówno uwzględnienia cech dźwiękowych, jak i rozumienia tekstu.	*Is the person talking about the following sound?*
Text and Temporal Localization Reasoning	Pytania, których rozwiązanie wymaga zarówno uwzględnienia odgłosów związanych z lokalizacją (otoczeniem), jak i rozumienia tekstu.	*Does the speaker describe the acoustic scene that they are in?*

Szablony

Is the person talking about the following sound? + [WYPOWIEDŹ] + [DŹWIĘK]

Is the person talking about the following sound?
+ A thunderstorm rumbles loudly in the distance.
+ traffic

No

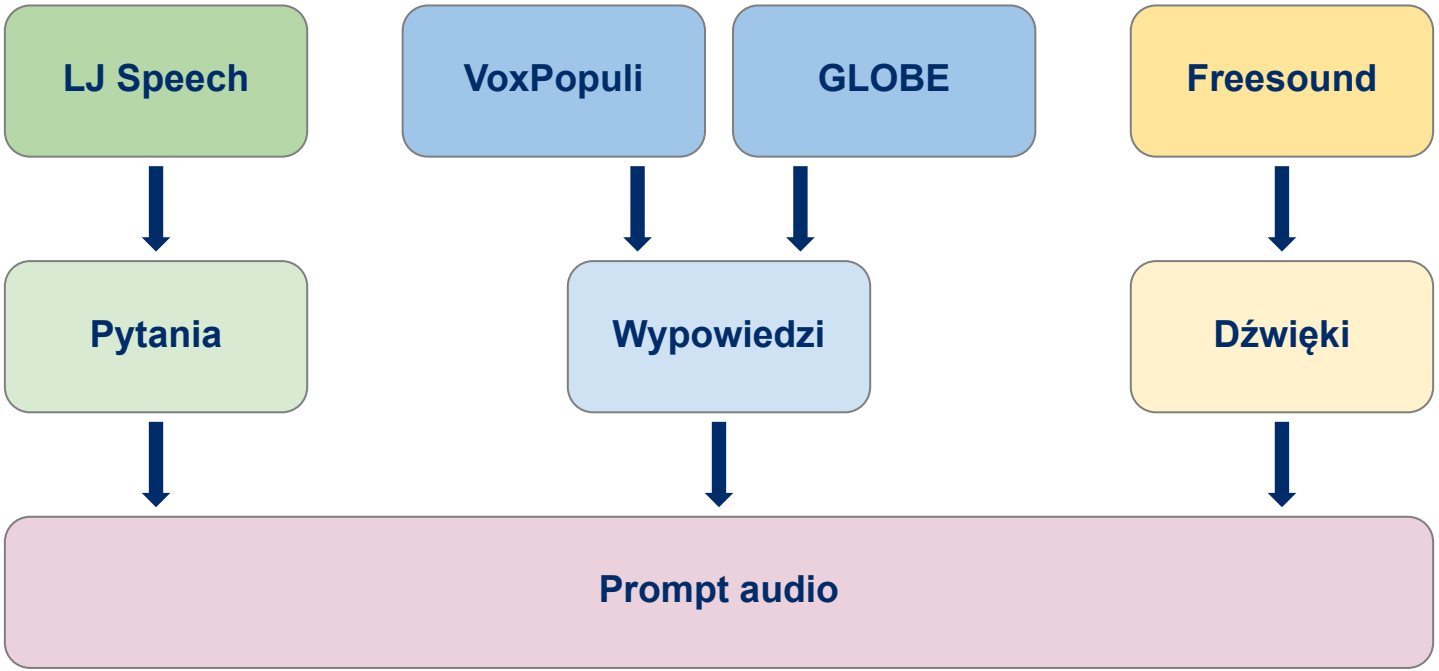
Wypowiedzi

1. The airplane soars high above the clouds.
2. A thunderstorm rumbles loudly in the distance.
3. (…)

Dźwięki

1. doorbell
2. thunderstorm
3. traffic
4. (…)

Przygotowanie zadań



Własności zbioru

Zadanie	Ilość próbek	Ilość szablonów	Ilość mówców	Ilość wypowiedzi	Ilość dźwięków	Długość nagrań
AA	1000	6	n.d.	n.d.	5	3 godz. 46 min 10 s
ATD	1000	4	n.d.	n.d.	4	3 godz. 53 min 30 s
CRLI	1000	6	12	12	n.d.	3 godz. 32 min 1 s
CRSI	1000	4	4	8	n.d.	3 godz. 46 min 17 s
STI	1000	4	4	36	n.d.	3 godz. 9 min 47 s
SR	1000	15	n.d.	n.d.	20	3 godz. 8 min 15 s
SFC	1000	4	10	3	n.d.	2 godz. 37 min 7 s
TSR	1000	8	4	16	17	3 godz. 25 min 24 s
TTLR	1000	4	4	15	8	3 godz. 47 s
Łącznie	**9000**	**55**	**22**	**86**	**25**	**30 godz. 19 min 18 s**

Zbiór jest w pełni zrównoważony - 4500 pytań ma oczekiwaną odpowiedź *Tak*, a 4500 - *Nie*. Równowaga jest zachowana w obrębie każdego z dziewięciu zadań. Tam, gdzie było to możliwe, zachowano równowagę również na poziomie szablonów.

Podejście eksperymentalne

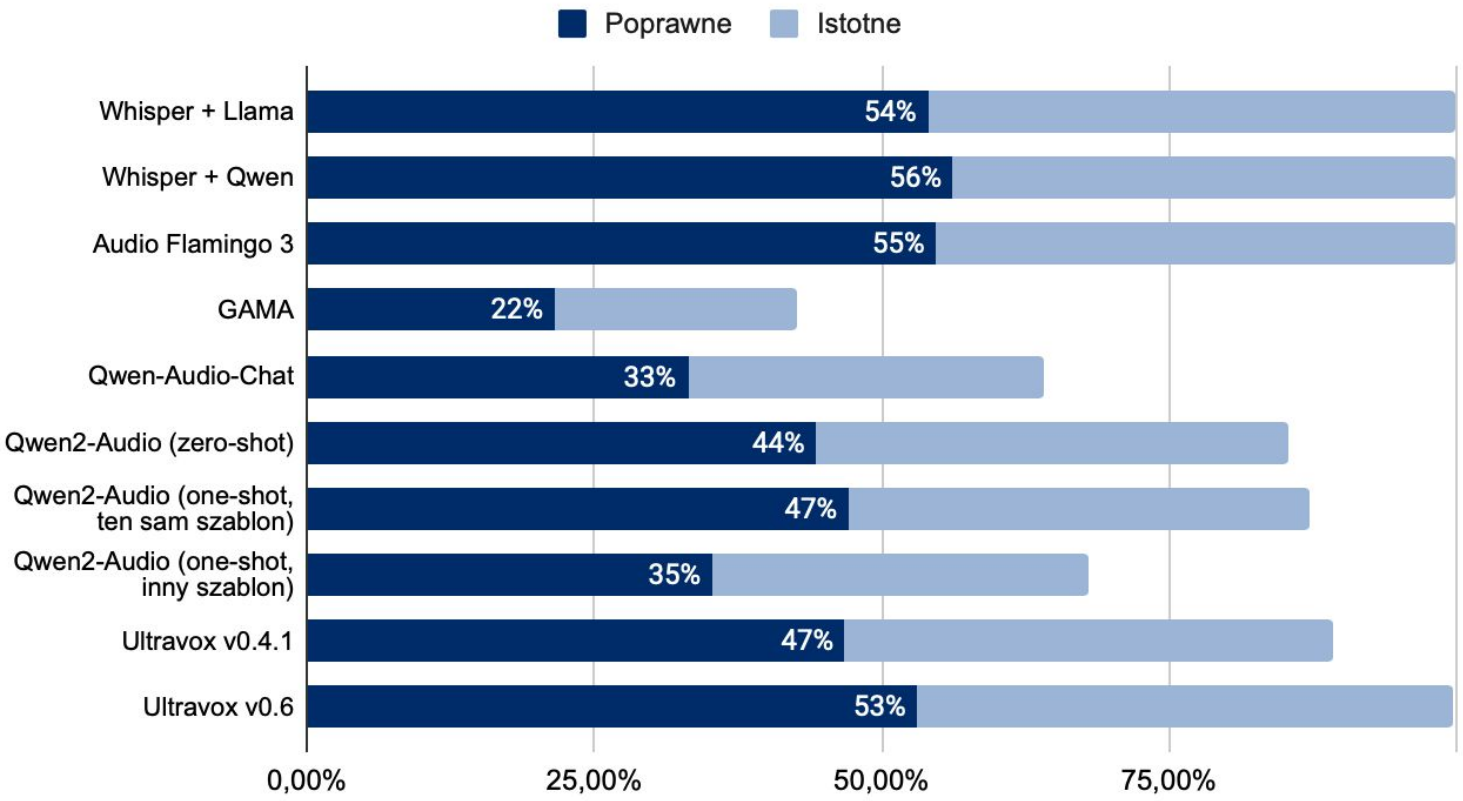
Yes/No

- Modele były instruowane, aby odpowiadać wyłącznie *Yes* lub *No*.
- Odpowiedzi były oceniane automatycznie.
- Inferencja dla każdego modelu była uruchamiana pięć razy, a wyniki były uśredniane.

Descriptive

- Forma odpowiedzi nie była określona - modele mogły udzielić odpowiedzi opisowej.
- Odpowiedzi były oceniane przez dwóch sędziów - **Llama3.3-70B-Instruct** i **Qwen3-32B**.
- Sędzia oznaczał istotność i poprawność odpowiedzi.

Yes/No: Wyniki ogólne



Yes/No: Wyniki dla zadań

Model	AA	ATD	CRLI	CRSI	STI	SR	SFC	TSR	TTLR
Whisper + Llama	0,505	0,483	0,458	0,505	0,665	0,525	0,557	0,643	0,521
Whisper + Qwen	0,51	0,504	0,551	0,501	0,625	0,654	0,532	0,653	0,528
Audio Flamingo 3	0,516	0,492	0,569	0,517	0,554	0,7	0,494	0,566	0,517
GAMA	0,036	0,376	0,276	0,349	0,338	0,05	0,299	0,068	0,147
Qwen-Audio-Chat	0,498	0,521	0,008	0,123	0,201	0,549	0,181	0,525	0,375
Qwen2-Audio (zero-shot)	0,488	0,517	0,501	0,212	0,532	0,432	0,524	0,358	0,425
Qwen2-Audio (one-shot, ten sam szablon)	0,493	0,282	0,517	0,454	0,534	0,517	0,527	0,543	0,373
Qwen2-Audio (one-shot, inny szablon)	0,441	0,245	0,464	0,34	0,477	0,176	0,467	0,392	0,172
Ultravox v0.4.1	0,475	0,288	0,516	0,489	0,478	0,438	0,507	0,511	0,512
Ultravox v0.6	0,48	0,501	0,579	0,504	0,559	0,526	0,521	0,59	0,513
Średnia	0,436	0,413	0,43	0,386	0,49	0,43	0,457	0,476	0,396

Yes/No: Analiza błędów

GAMA. W 51% przypadków nie rozpoznaje pytania, w 48% nie rozpoznaje mowy lub dźwięku.

Qwen-Audio-Chat. 100% błędów to transkrypcja nagrania zamiast udzielenia odpowiedzi.

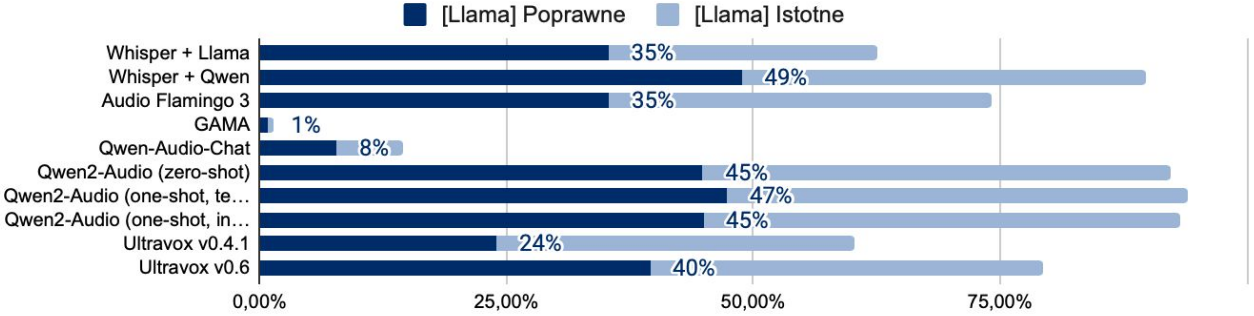
Qwen2-Audio (zero-shot). W 50% przypadków udziela odpowiedzi niezwiązanej z tematem.

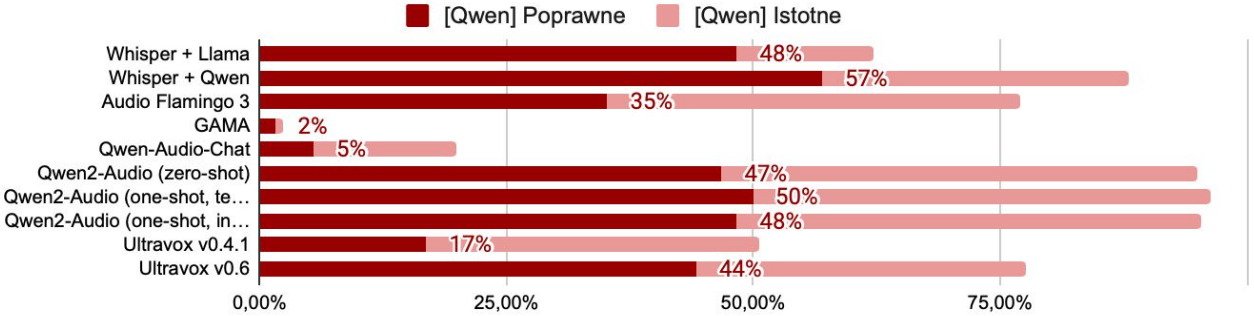
Qwen2-Audio (one-shot). Większość błędów polega na rozpoznaniu mówcy zamiast udzielenia odpowiedzi.

Ultravox v0.4.1. W 67% przypadków nie rozpoznaje mowy lub dźwięku.

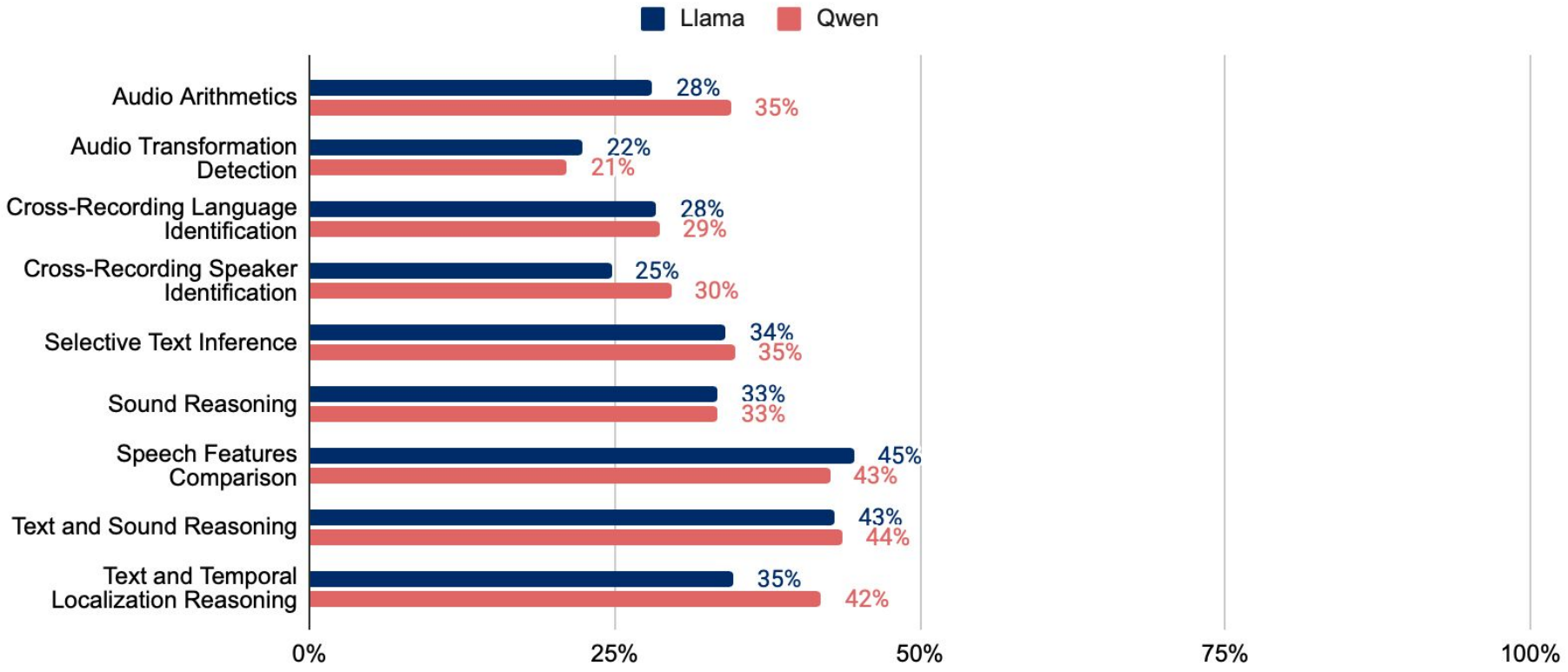
- W wielu przypadkach modele mają problem z poprawną interpretacją wejścia.
- Część modeli dokonuje transkrypcji lub rozpoznania mówcy, nadpisując właściwe zadanie.
- Odpowiedzi w innym języku lub całkowicie niezwiązane z zadanym pytaniem pokazują słabą kontrolę zachowania w przypadku niepewności.

Descriptive: Wyniki ogólne





Descriptive: Wyniki dla zadań



Descriptive: Analiza błędów

Whisper + Llama. W większości przypadków informuje, że nie może pomóc.

Whisper + Qwen. W 49% przypadków nie rozpoznaje zadanego pytania.

Audio Flamingo 3. 61% błędów stanowi zwrócenie transkrypcji zamiast udzielenia odpowiedzi.

Qwen-Audio-Chat. W 79% przypadków zwracał transkrypcję.

Qwen2-Audio (zero-shot). 61% odpowiedzi było w innym języku.

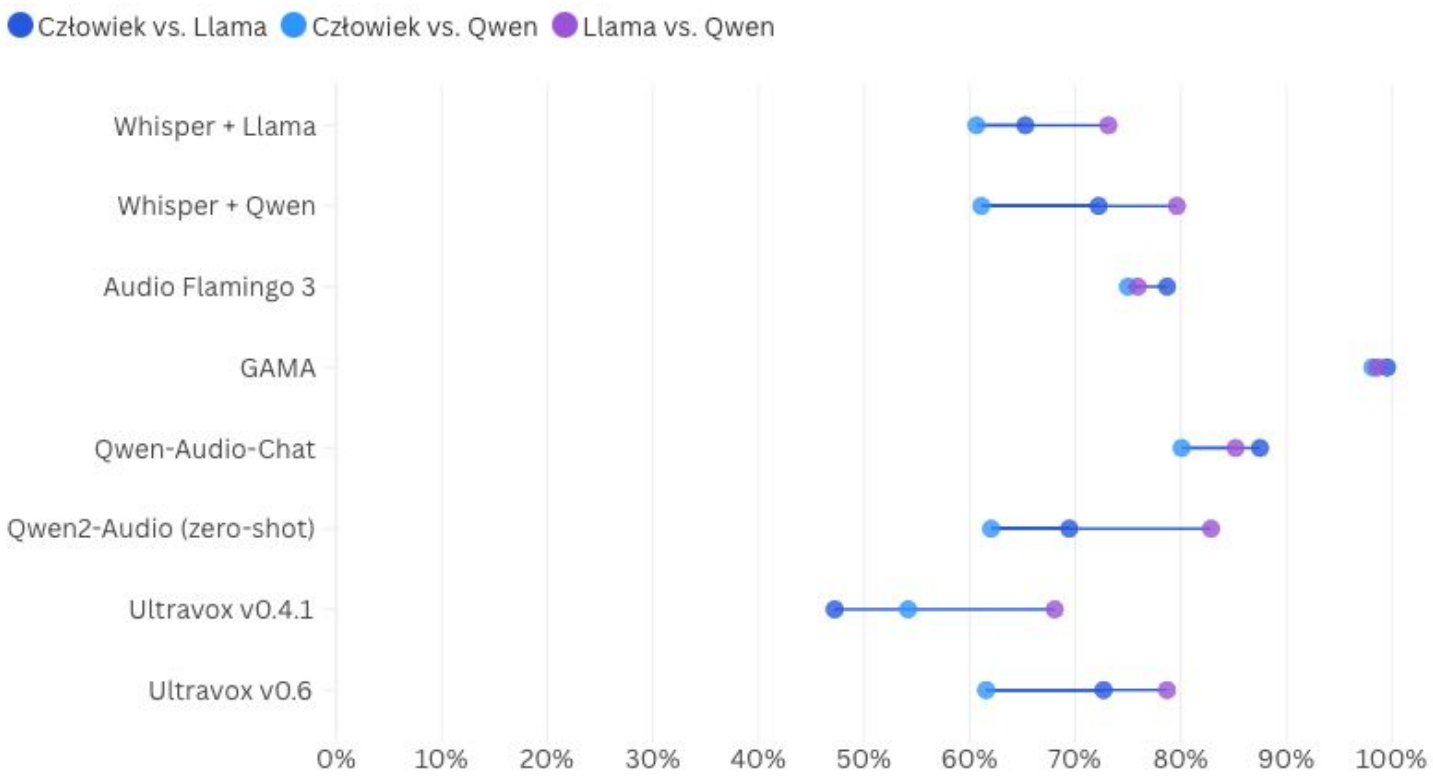
Qwen2-Audio (one-shot). Ponad 30% błędów wynika z udzielenia odpowiedzi niezwiązanej z tematem.

Ultravox v0.4.1. W 48% przypadków zwraca spekulatywną odpowiedź.

Ultravox v0.6. 45% błędów wynika z nierozpoznania pytania.

- Brak ograniczenia do *Yes/No* otwiera przestrzeń dla halucynacji i spekulacji.
- Modele częściej odwołują się do standardowych zadań, takich jak transkrypcja.

Descriptive: Zgodność sędziów

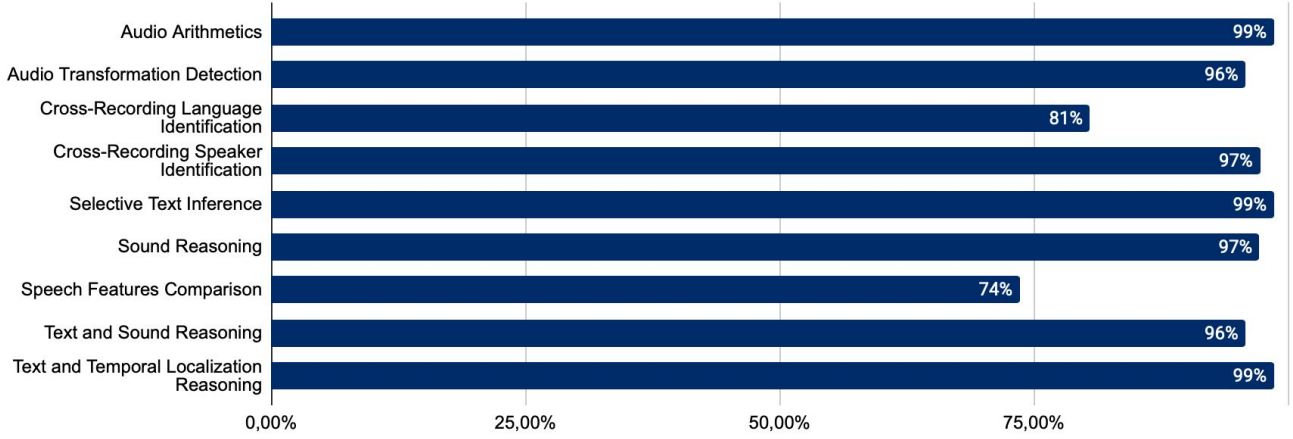


ART-H

Procedura generowania poszczególnych instancji zadań zależy od szablonów i syntetycznej mowy, co potencjalnie może prowadzić do powstania benchmarku zbyt trudnego dla ludzi.

Aby kontrolować ten problem, wydzielony został podzbiór ART-H, składający się z 216 próbek (24 próbki na zadanie).

Podzbiór ten, umożliwia ręczną weryfikację wyników ewaluacji w czasie krótszym niż godzina.

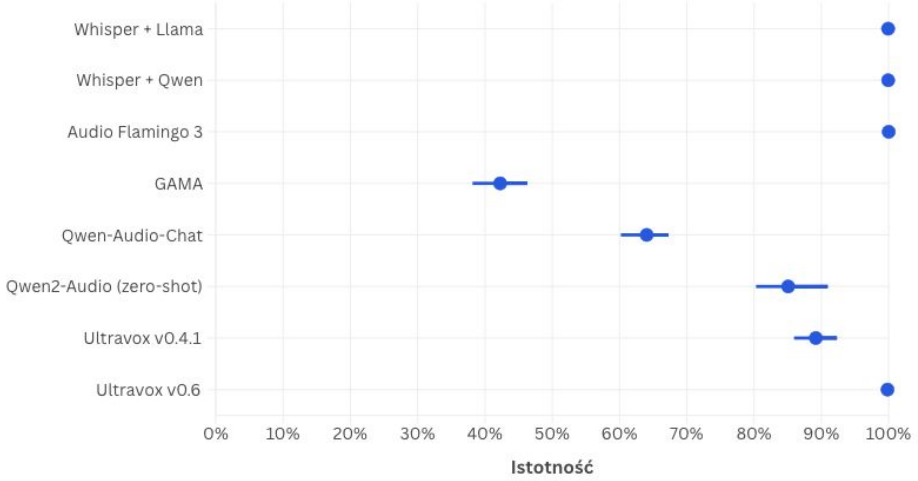


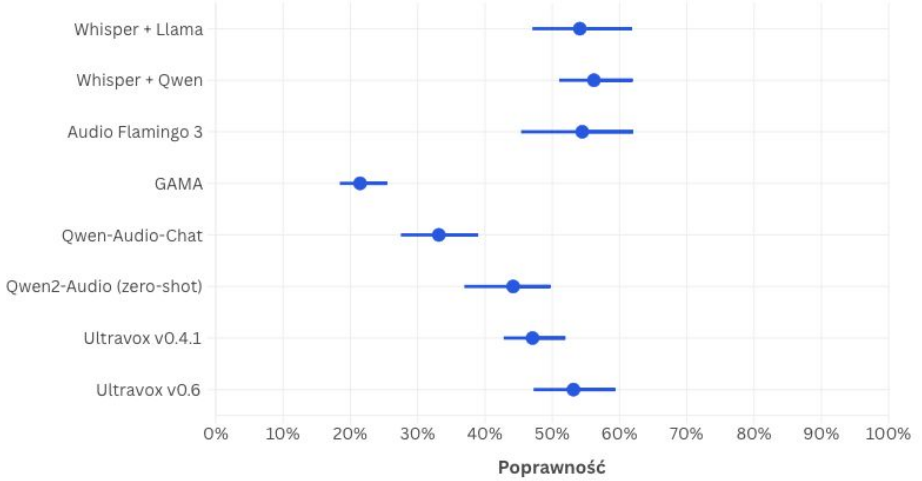
93%
HUMAN
BASELINE

ART-H

Aby sprawdzić, czy wybór konkretnych próbek wpływa istotnie na wyniki, oceniono modele względem losowo próbkowanych 216-elementowych podzbiorów.

Procedura ta prowadzi do odchylenia standardowego poniżej 3,5% pod względem dokładności bezwzględnej.





Wnioski

Opracowane zadania, będąc łatwymi do rozwiązania dla człowieka, równocześnie stanowią istotne wyzwanie dla modeli poddanych ewaluacji.

W żadnym z zadań oceniane modele nie osiągnęły dokładności większej niż 0.7, a jeżeli pominąć w zestawieniu zadanie Sound Reasoning (SR w Tabeli 6), to żaden z modeli nie osiągnął sześćdziesięcioprocentowej dokładności.

W wielu przypadkach modele nie pokonały rozwiązań bazowych złożonych z systemu rozpoznawania mowy oraz tekstowego, unimodalnego modelu językowego

Podsumowanie

Zaproponowaliśmy **Audio Reasoning Tasks (ART)** nowy benchmark służący do oceny jakości MLLMów.

W przeciwieństwie do dotychczasowych benchmarków, które testują zdolności audio w izolacji, nasz benchmark obejmuje zadania wymagającełączenia różnych umiejętności w celu rozwiązania postawionego problemu.

Zadania zostały zaprojektowane w taki sposób, aby mogła je rozwiązać osoba bez specjalistycznych umiejętności oraz zaburzeń słuchu.

Przeprowadzone eksperymenty pokazały, że ART stanowi wyzwanie dla model, które poddaliśmy analizie.