

dr hab. inż. Dariusz Czerwiński, prof. uczelni
Politechnika Lubelska
ul. Nadbystrzycka 38,
20-618 Lublin
tel. +48 81 538 48 96

Lublin, dnia 03 lipca 2026 r.

RECENZJA ROZPRAWY DOKTORSKIEJ

mgr inż. Bartłomieja Klina pt. „Zwiększanie efektywności komunikacji człowiek-komputer (human-to-machine) w języku naturalnym w autonomicznych systemach robotycznych poprzez akwizycję i deskrypcję zagregowanej informacji wielokanałowej”

Przedmiotem recenzji jest rozprawa doktorska mgr inż. Bartłomieja Klina. Promotorem rozprawy jest dr hab. inż. Ryszard Beniak, prof. Politechniki Opolskiej, a promotorem pomocniczym dr hab. inż. Michał Podpora, prof. Uniwersytetu Opolskiego. Rozprawa jest ulokowana w dziedzinie nauk inżynieryjno-technicznych, w dyscyplinie automatyka, elektronika, elektrotechnika i technologie kosmiczne.

Opiniowana rozprawa doktorska została przygotowana w języku polskim. Według numeracji autora obejmuje 230 stron zasadniczego tekstu, a przekazany egzemplarz zawiera dodatkowo stronę tytułową, podziękowania, streszczenie w języku polskim i angielskim oraz spis treści. Dysertacja zawiera:

- wstęp, motywację oraz hipotezę badawczą,
- przegląd literatury dotyczący robotycznego przetwarzania mowy w scenariuszach multi-talker,
- rozdziały teoretyczne dotyczące akustyki, derewerberacji, tłumienia echa, synchronizacji wieloczujnikowej oraz lokalizacji źródeł dźwięku,
- opis stanowisk badawczych, komponentów i eksperymentów,
- model matematyczny komponentów szkieletu aplikacji Smart Beamforming,
- porównanie wariantów o nazwach *baseline* oraz *full*, opis wdrożenia oraz podsumowanie,
- bibliografię obejmującą 89 pozycji.

Cel, zakres i charakter rozprawy

Tematyka rozprawy mieści się w obszarze współczesnych systemów robotycznych, przetwarzania sygnałów akustycznych, komunikacji człowiek-robot oraz integracji heterogenicznych danych sensorycznych. Zasadniczym problemem badawczym jest zwiększenie skuteczności komunikacji werbalnej człowieka z autonomicznym systemem robotycznym w sytuacjach, w których w otoczeniu

systemu znajduje się więcej niż jeden rozmówca, występuje hałas, nakładanie się wypowiedzi, zmienna geometria sceny oraz okresowa degradacja toru wizyjnego.

Celem pracy było opracowanie, implementacja oraz walidacja szkieletu aplikacji *Smart Beamforming*. Szkielet ten ma umożliwiać akwizycję, synchronizację i fuzję informacji wielokanałowej, adaptacyjne sterowanie wiązkami mikrofonowymi, utrzymywanie kontekstów rozmówców oraz wykorzystywanie metryk jakości percepcji w pętli sterowania robota. W tym sensie praca ma charakter teoretyczno-eksperymentalny i wdrożeniowy. Autor nie ograniczył się do zaproponowania pojedynczego algorytmu, lecz przedstawił architekturę systemową obejmującą warstwę percepcji, diagnostyki, sterowania i awaryjnej degradacji działania systemu. Jest to zagadnienie silnie związane ze sterowaniem zachowania robota w tak zwanych warunkach niesprzyjających.

Zakres pracy doktorant opisał i zrealizował poprzez następujące zadania:

- analizę ograniczeń klasycznego modelu komunikacji REPL w robotach humanoidalnych,
- przegląd rozwiązań z obszary słyszenia robotów, formowania wiązki, separacji źródeł, fuzji audio-wideo oraz mechanizmów kolejności wypowiedzi,
- opracowanie modułów diagnostycznych, w szczególności *Recovery Time Calculator* oraz *WER Calculator*,
- opracowanie warstwy *Virtual Context Factory* służącej do utrzymywania kontekstów rozmówców,
- zaprojektowanie modułów sterowania nadzorczego *SuperManager*, *FallbackManager* oraz *FailsafeManager*,
- budowę kolejnych wersji stanowisk badawczych oraz przeprowadzenie eksperymentów laboratoryjnych i wdrożeniowych,
- porównanie wariantu bazowego *baseline* i wariantu pełnego *full* z wykorzystaniem wskaźników dedykowanych pod KPI, czyli: RT, WER, DER i Functional Delay.

W rozdziale 1 doktorant wprowadził problem komunikacji człowiek-maszyna oraz wskazał ograniczenia sekwencyjnego schematu REPL. W rozdziale tym sformułowano hipotezę badawczą, zgodnie z którą „możliwe jest wykorzystanie autonomicznego systemu robotycznego do jednoczesnej obsługi konwersacji z wieloma rozmówcami w warunkach rzeczywistych”. Rozdział 2 zawiera przegląd literatury w kontekście rozwiązania *Smart Beamforming* z nurtami słyszenia robotów oraz systemami takimi jak HARK, ManyEars i ODAS. Rozdziały 3-8 mają charakter teoretyczny i porządkują podstawy akustyki, propagacji dźwięku, derewerberacji,

tłumienia echa, synchronizacji danych wieloczuJNIKOWYCH i lokalizacji źródeł dźwięku. Rozdział 9 przedstawia stanowiska badawcze i komponenty opracowanego rozwiązania. Rozdział 10 opisuje prototyp ver. 4 oraz mechanizmy sterowania interakcją. Rozdział 11 zawiera finalne modele matematyczne, scenariusze symulacyjne oraz wyniki. Rozdział 12 przedstawia porównanie wariantów *baseline* i *full*. Rozdział 13 opisuje wdrożenie wyników pracy, a rozdział 14 stanowi syntetyczne podsumowanie.

Opinia na temat rozprawy

Dysertacja bazuje na doświadczeniach autora wynikających z prac wdrożeniowych z robotami humanoidalnymi oraz na publikacjach naukowych, w tym wskazanych w pracy artykułach dotyczących rozwiązania Smart Beamforming i interakcji werbalnej z udziałem wielu rozmówców. Jest to istotna zaleta rozprawy, ponieważ problem badawczy nie ma charakteru wyłącznie teoretycznego, lecz wynika z ograniczeń realnych systemów robotycznych stosowanych w środowiskach usługowych.

Hipoteza badawcza została sformułowana następująco: „*Możliwe jest wykorzystanie autonomicznego systemu robotycznego do jednoczesnej obsługi konwersacji z wieloma rozmówcami w warunkach rzeczywistych*”. W celu jej weryfikacji doktorant przyjął trzy kamienie milowe: K1 dotyczący skrócenia czasu rekonfiguracji wiązki (ang. Recovery Time), K2 dotyczący ograniczenia błędu rozpoznawania mowy (ang. Word Error Rate) oraz K3 dotyczący zarządzania kontekstem i płynnością interakcji, mierzonego między innymi przez parametry DER oraz Functional Delay.

W opinii recenzenta przyjęta hipoteza została co do zasady potwierdzona. Wyniki przedstawione w rozdziale 12 wskazują, że wariant *full* osiągnął $RT_{p95} = 0,580$ s i $WER_{p90} = 0,195$, podczas gdy wariant *baseline* osiągnął odpowiednio $RT_{p95} = 5,895$ s i $WER_{p90} = 0,873$ s. Doktorant wykazał również obniżenie parametru DER z 0,149 do 0,104 oraz Functional Delay z 1,056 s do 0,41 s. Wyniki te potwierdzają, że integracja metryk jakości percepcji z mechanizmami sterowania i redundancji może istotnie poprawić funkcjonalną skuteczność interakcji w scenariuszach z wieloma rozmówcami.

Na uwagę zasługuje systemowe ujęcie problemu. W wielu pracach z obszaru przetwarzania mowy centralne miejsce zajmuje pojedynczy algorytm separacji, lokalizacji lub odszumiania. Doktorant przyjął szerszą perspektywę: potraktował percepcję mowy jako proces, którego jakość musi być monitorowana i wykorzystywana w sterowaniu zachowaniem robota. W mojej ocenie jest to najważniejszy element wkładu naukowego i aplikacyjnego pracy.

Pozytywnie oceniam również konsekwentną ewolucję stanowisk badawczych od prototypu uruchomionego w recepcji firmy Weegree, przez stanowiska laboratoryjne,

aż po prototyp ver. 4. Taki sposób prowadzenia badań dobrze odpowiada formule doktoratu wdrożeniowego, ponieważ pozwalał łączyć obserwacje praktyczne z kontrolowanymi eksperymentami oraz iteracyjnym doskonaleniem architektury.

Analiza literatury obejmuje zarówno klasyczne podstawy akustyki i przetwarzania sygnałów, jak i nowsze prace z zakresu formowania wiązki, audiowizualnej separacji mowy, synchronizacji wieloczułnikowej oraz słyszenia robotów. Bibliografia jest obszerna i aktualna, chociaż w kilku miejscach wymaga większej konsekwencji redakcyjnej oraz wyraźniejszego rozdzielenia pozycji naukowych od źródeł dokumentacyjnych i opisów sprzętu. Doktorant wykazał się znajomością zagadnień z zakresu przetwarzania sygnałów, akustyki pomieszczeń, synchronizacji sensorów, robotyki, rozpoznawania mowy i projektowania architektur systemowych. Szczególnie wartościowe jest powiązanie modelu matematycznego komponentów z praktycznymi modułami systemu oraz z mierzalnymi wskaźnikami jakości.

Oceniając wkład autora rozprawy do rozwoju nauki można wskazać przede wszystkim: zaproponowanie architektury *Smart Beamforming* jako sterowanego metrykami szkieletu aplikacji do obsługi percepcji mowy w robotyce, wprowadzenie parametru Recovery Time jako operacyjnej miary czasu powrotu jakości po rekonfiguracji wiązek, integrację warstwy VCF z mechanizmami świadomości celów i kolejnością zabierania głosów oraz praktyczne połączenie mechanizmów awaryjności/niezawodności z jakością percepcji. Za istotny należy uznać również aspekt wdrożeniowy, tj. przygotowanie finalnej implementacji prototypu i jej uruchomienie w środowisku rzeczywistym.

Uwagi szczegółowe

Na podstawie analizy pracy mogę stwierdzić, że układ rozprawy jest zasadniczo właściwy, może zbyt mocno rozbudowana część teoretyczna dysertacji. Przedstawiony materiał dokumentuje znaczący zakres wykonanych badań teoretycznych, eksperymentalnych i wdrożeniowych. W trakcie lektury nasuwają się jednak uwagi redakcyjne i merytoryczne wyszczególnione poniżej. Pragnę podkreślić, że mają one charakter dyskusyjny i nie umniejszają pozytywnej oceny recenzowanej pracy. Uwagi redakcyjne należy potraktować jako wskazówki przy redagowaniu przyszłych opracowań naukowych oraz tekstów przygotowywanych przez autora.

Uwagi redakcyjne

- 1) W pracy występują liczne drobne błędy językowe, literówki i niezręczności składniowe. Przykładowo: „po krótcie” powinno być zapisane jako „pokrótce”, pojawia się zapis „Peppe r”, „heterogonicznych”, „dźwięku”, „rysynku”, „beamforingu”, a w rozdziale 2 występuje literówka „FallbackManager/FailsafeManagepr”, na str. 71 „pełni w frameworku”, lub

str. 81 „Architektura software’owa bazowego pipeline’u” – typowy slang informatyczny.

- 2) W kilku miejscach zdania mają charakter urwany lub potoczny. Przykładem jest sformułowanie ze streszczenia „Co jest w przeciwieństwie do klasycznego podejścia...”. Zdanie to warto połączyć z poprzednim lub przeredagować, np. „Jest to podejście odmienne od klasycznego...”.
- 3) Należy ujednolicić zapis terminów angielskich i polskich. W pracy występują równoległe formy „video” i „wideo”, „human - to -machine” oraz „human-to-machine”, a także różne zapisy nazw modułów i metryk. Podobnie jest ze stosowaniem nazewnictwa z j. angielskiego, np. sformułowanie na str. 50 „co powinno uruchomić procedury failsafe” da się zapisać całkowicie w j. polskim. Podobnie „logarytmicznego sweepu” na str. 57 powinno być „przemiatania logarytmicznego”. Str. 71 „wektory embeddingów twarzy, bounding boxy” jest to slang informatyczny, patrz Słownik terminów informatycznych i komputerowych ang. bounding box – pole ograniczenia. W pracy naukowej warto konsekwentnie stosować jeden wariant zapisu i to w dodatku uznany w słownikach języka polskiego, lub słownikach technicznych.
- 4) W tytułach podrozdziałów występuje niejednolita kapitalizacja, np. „Lokalizacja Źródeł Dźwięku” oraz „Zarządzanie kontekstem i płynność Interakcji”. W języku polskim nie ma potrzeby rozpoczynania każdego rzeczownika wielką literą.
- 5) W podpisach pod rysunkami i tabelami warto zachować jednolity standard. W pracy pojawiają się zapisy „Tabela” i „Tablica”, a źródło rysunku bywa opisane jako „Internet”. W rozprawie doktorskiej należy podać pełne źródło, adres, datę dostępu lub informację, że rysunek jest opracowaniem własnym.
- 6) Zwraca uwagę duża liczba rysunków, wykresów i diagramów. Część z nich jest bardzo wartościowa, ale niektóre mają niską czytelność po wydruku, zwłaszcza tam, gdzie występują małe opisy osi, symbole greckie i zagęszczone adnotacje. Warto rozważyć powiększenie najważniejszych artefaktów wynikowych albo przeniesienie części materiału szczegółowego do załączników. Ponadto warto ujednolicić wersje językowe rysunków, skoro dysertacja jest w j. polskim to rysunki autorskie też winny być w j. polskim.
- 7) Praca jest obszerna i miejscami powtarza podobne uzasadnienia dotyczące roli SNR, RT, WER, fallback i failsafe. W kilku miejscach można skrócić opis, pozostawiając pełną wersję w rozdziałach modelowych lub w załącznikach.
- 8) Warto poprawić interpunkcję przy wtrąceniach i wyliczeniach, w szczególności przed konstrukcjami „który”, „w którym”, „co”, „ponieważ”. Poprawi to czytelność, zwłaszcza w rozdziałach 9-14, w których tekst ma charakter techniczno-sprawozdawczy.
- 9) Podrozdziały w rozdziale 9.2 są jednozdaniowe, nie powinny być podrozdziałami.

Uwagi merytoryczne

- 1) Hipoteza badawcza została sformułowana dość szeroko: „możliwe jest wykorzystanie autonomicznego systemu robotycznego do jednoczesnej obsługi konwersacji z wieloma rozmówcami w warunkach rzeczywistych”. Czy doktorant może doprecyzować, jakie minimalne warunki graniczne muszą być spełnione, aby hipotezę uznać za prawdziwą, np. liczba rozmówców, minimalna separacja kątowa, odległość od robota, poziom hałasu, dopuszczalne nakładanie się wypowiedzi i parametry toru wizyjnego?
- 2) W pracy porównano wariant *baseline* i wariant *full*. Czy doktorant może szerzej uzasadnić, czy wariant *baseline* jest reprezentatywny dla aktualnych rozwiązań komercyjnych i akademickich?
- 3) Wyniki eksperymentu rzeczywistego są przekonujące, ale opis wskazuje na dwie tury pomiarowe i około 60 sekund interakcji dla każdego wariantu. Czy doktorant może określić, jaka była liczba powtórzeń, liczba niezależnych uczestników oraz zmienność wyników między próbami? Warto rozważyć podanie przedziałów ufności lub przynajmniej miar rozproszenia dla kluczowych KPI.
- 4) Przyjęty próg $RT \leq 600$ ms jest logicznie uzasadniony tempem konwersacji, ale warto mocniej powiązać go z percepcją użytkownika i literaturą HRI. Czy próg 600 ms jest uniwersalny dla różnych języków, typów wypowiedzi i użytkowników, czy stanowi wartość dobraną do zaprojektowanego scenariusza?
- 5) Próg $WER \leq 20\%$ został przyjęty jako granica akceptowalna dla dalszego przetwarzania NLP/LLM. Czy doktorant może doprecyzować, jak ten próg przekłada się na skuteczność rozpoznawania intencji i slotów? W pracy pojawia się poprawność semantyczna, ale warto wyraźniej pokazać relację między WER, dokładnością semantyczną (ang. semantic accuracy) oraz końcową jakością decyzji dialogowej.
- 6) Parametr DER został w pracy potraktowany jako wskaźnik poprawnego przypisania mowy do kontekstu rozmówcy. Warto doprecyzować, czy jest to klasyczny Diarization Error Rate, czy też miara skonstruowana na potrzeby systemu. Jeżeli jest to miara, należy jednoznacznie podać jej definicję, sposób wyznaczania i zasady liczenia błędów.
- 7) Zaproponowana warstwa VCF jest jednym z najciekawszych elementów rozprawy. Warto jednak mocniej oddzielić jej wkład algorytmiczny od wkładu integracyjnego. Czy VCF wprowadza nową metodę asocjacji i utrzymania tożsamości rozmówcy, czy jest przede wszystkim architektoniczną warstwą łączącą znane mechanizmy detekcji twarzy, głosu, diaryzacji i świadomości celu?
- 8) W rozdziałach porównawczych Smart Beamforming jest zestawiany z HARK oraz mechanizmami platformy Pepper. Czy możliwe byłoby podanie bardziej

- ilościowego porównania, np. na wspólnym scenariuszu testowym, albo przynajmniej tabelarycznego zestawienia cech z zaznaczeniem, które elementy zostały faktycznie przetestowane, a które wynikają z analizy dokumentacji?
- 9) Praca pokazuje duże znaczenie degradacji toru wizyjnego. Czy doktorant może dokładniej wyjaśnić, jak system rozstrzyga konflikt między informacją wizyjną a audio-only/BSS, zwłaszcza w sytuacjach, gdy oba tory wskazują różne pozycje rozmówcy albo mają podobną niepewność?
 - 10) W rozdziale dotyczącym stanowisk badawczych warto mocniej doprecyzować parametry własnej macierzy mikrofonowej 2x4, geometrię rozmieszczenia mikrofonów, kalibrację poziomów, charakterystykę pomieszczenia oraz warunki akustyczne. Jest to ważne z punktu widzenia powtarzalności badań.
 - 11) Wdrożenie w recepcji firmy Weegree jest znaczącym atutem pracy. Czy doktorant może przedstawić, jak długo system pracował w warunkach naturalnej eksploatacji, ilu użytkowników obejmowały testy i jakie były kryteria sukcesu wdrożeniowego poza metrykami laboratoryjnymi?
 - 12) Warto szerzej omówić przypadki niepowodzeń wariantu *full*. W pracy pokazano znaczną poprawę KPI, ale dla dalszego rozwoju naukowego i wdrożeniowego szczególnie cenne byłyby przykłady sytuacji, w których mechanizmy fallback/failsafe nie wystarczyły lub wymagały ręcznego strojenia.
 - 13) Warto jednoznacznie wskazać, które elementy opisane w pseudokodach zostały faktycznie zaimplementowane w prototypie ver. 4, które były użyte w symulacji, a które stanowią jedynie projekt lub kierunek dalszych prac. W obecnej postaci granica między implementacją, modelem i koncepcją nie zawsze jest wystarczająco czytelna.
 - 14) Należy ostrożniej formułować twierdzenia o „warunkach rzeczywistych”. Test w recepcji firmy Weegree jest wartościowy wdrożeniowo, ale dla uogólnienia wyników potrzebne są dane z dłuższej eksploatacji, większej liczby użytkowników, różnych akcentów, płci, poziomów hałasu, konfiguracji przestrzennych i typów wypowiedzi.
 - 15) Warto uzupełnić pracę o analizę złożoności obliczeniowej i opóźnień poszczególnych modułów. ODAS wprost akcentuje problem kosztu obliczeniowego i pracy na sprzęcie wbudowanym. W recenzowanej pracy także pojawiają się wątki czasu rzeczywistego, ale brakuje syntetycznego profilu CPU/GPU, pamięci, opóźnień modułowych i marginesów czasowych.

Konkluzja

Całość przygotowanej rozprawy doktorskiej wskazuje jednoznacznie na opanowanie warsztatu naukowego przez doktoranta. Przedstawiona rozprawa zawiera interesujące wyniki badań teoretycznych, eksperymentalnych oraz wdrożeniowych. Autor zidentyfikował istotny problem praktyczny i naukowy, zaproponował architekturę jego rozwiązania, zbudował prototypy, przeprowadził walidację oraz

przedstawił wyniki porównawcze wskazujące na skuteczność opracowanego podejścia. Doktorant wykazał się znajomością zagadnień związanych z przetwarzaniem sygnałów, robotyką, akustyką, synchronizacją danych wieloczuJNIKOWYCH, rozpoznawaniem mowy i sterowaniem interakcją człowiek-robot. Autor rozwiązał samodzielnie oryginalny problem naukowy i wdrożeniowy z dyscypliny automatyka, elektronika, elektrotechnika i technologie kosmiczne oraz potrafił przekonująco zaprezentować uzyskane rezultaty.

W mojej ocenie rozprawa doktorska mgr inż. Bartłomieja Klina pt. „Zwiększanie efektywności komunikacji człowiek-komputer (human-to-machine) w języku naturalnym w autonomicznych systemach robotycznych poprzez akwizycję i deskrypcję zagregowanej informacji wielokanałowej” spełnia wymagania stawiane rozprawom doktorskim, w szczególności prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie, potwierdza umiejętność samodzielnego prowadzenia pracy naukowej oraz zawiera oryginalne rozwiązanie problemu naukowego i aplikacyjnego.

Zgodnie z zaprezentowaną powyżej oceną merytoryczną rozprawy Pana mgr, inż. Bartłomieja Klina stwierdzam, że opiniowana praca spełnia warunki i wymagania stawiane rozprawom doktorskim określone w art. 187 ustawy z dnia 20 lipca 2018 r. Prawo o szkolnictwie wyższym i nauce (tj. Dz U. z 2023 r., poz. 742 z późn. zm.).

Dlatego też stawiam wniosek o dopuszczenie rozprawy doktorskiej mgr inż. Bartłomieja Klina do publicznej obrony zgodnie z procedurą określoną przez Radę Dyscypliny Naukowej Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne Politechniki Opolskiej.

D. Czerwiński

dr hab. inż. Dariusz Czerwiński, profesor uczelni