

Warszawa, 30 grudnia 2024

prof. dr hab. inż. Jarosław Arabas
Politechnika Warszawska
Wydział Elektroniki i Technik Informacyjnych
Instytut Informatyki

WYDZIAŁ AUTOMATYKI,
ROBOTYKI I I ELEKTROTECHNIKI
POLITECHNIKA POZNAŃSKA

15 -01- 2025

WPŁYNĘŁO

RECENZJA

wniosku habilitacyjnego

dr. inż. Szymona Drgasa

PRZEWODNICZĄCY RADY DISCYPLINY
Automatyka, Elektronika, Elektrotechnika
i Technologie Kosmiczne

prof. dr hab. inż. Wojciech Szelaąg

I. ŻYCIORYS ZAWODOWY HABILITANTA

Szczegółowy życiorys zawodowy Habilitanta jest zawarty w autoreferacie, w przygotowanej przez niego dokumentacji. Wynika z niego, że w roku 2005 Habilitant uzyskał dwa dyplomy magistra: na kierunku Fizyka, specjalizacja akustyka, zaś w roku 2006 na kierunku Informatyka Stosowana na Uniwersytecie im. Adama Mickiewicza w Poznaniu, na Wydziale Fizyki. W roku 2013 uzyskał on stopień doktora w dyscyplinie automatyka i robotyka, na Wydziale Informatyki Politechniki Poznańskiej. Rozprawa doktorska (wyróżniona) nosiła tytuł „Automatic Speaker Recognition Based on Multilevel Analysis of Speech Signals”. Spełnione jest zatem wymaganie z rozdziału 3, art. 219, ust.1, pkt 1 ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce”.

Habilitant jest zatrudniony na Politechnice Poznańskiej od roku 2006, obecnie pracuje jako adiunkt na Wydziale Automatyki, Robotyki i Elektrotechniki.

Ten życiorys wskazuje na ciekawą ewolucję formacji intelektualnej Habilitanta – od nauk ścisłych i solidnych podstaw matematycznych procesów fizycznych, do nauk technicznych i zastosowania uzyskanej wiedzy fizycznej.

Habilitant ma w swoim dorobku 3 krótkie pobyty naukowe (w sumie ok. 2 miesiące): w Tampere University of Technology oraz w Carl von Ossietzky Universitaet, które zaowocowały m.in. powstaniem publikacji wraz ze współautorami z zagranicy [A2, A4, A6, A10]. Spełnione jest zatem wymaganie z rozdziału 3, art. 219, ust.1, pkt 3 ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce”, gdyż pobyty wraz ze wspólnymi publikacjami stanowią dowód „istotnej aktywności naukowej realizowanej w więcej niż jednej uczelni”.

Na szczególne podkreślenie i docenienie zasługuje współpraca Habilitanta z firmą StethoMe, która projektuje i rozwija stetoskop działający w oparciu o sztuczną inteligencję (StethoMe AI), analizujący i wykrywający nieprawidłowe dźwięki osłuchowe w układzie oddechowym. W ramach współpracy, jako inżynier projektu, Habilitant zajmował się doбором i opracowaniem algorytmów do przetwarzania i ekstrakcji informacji, a wyniki prac upowszechniał w formie publikacji.

Habilitant miał więc okazję wyjść poza mury Alma Mater zarówno w obrębie środowiska naukowego, jak i w otoczeniu gospodarczym. Świadczy to o powiązaniu jego badań naukowych z potrzebami realnego świata, a jednocześnie o jego twórczej obecności w międzynarodowym otoczeniu naukowym.

II. OCENA OSIĄGNIĘĆ NAUKOWYCH

Wśród swoich osiągnięć naukowych Habilitant wyróżnia cykl powiązanych tematycznie publikacji, dla którego przyjmuje on roboczy tytuł „Interpretowalne i niskokosztowe metody poprawy jakości mowy”. Cykl ten zawiera 13 pozycji, wśród których są (numeracja zaczerpnięta z wykazu dorobku):

- 7 artykułów w czasopismach „tradycyjnych”: Speech Communication [A1], IEEE/ACM Transactions on Audio, Speech, and Language Processing [A2, A6], Computer Speech & Language [A4], Journal of Speech, Language, and Hearing Research [A5], Multimedia Tools and Applications [A7, A8]
- 2 artykuły w MDPI Sensors [A3, A9]
- 4 referaty w materiałach konferencji międzynarodowych: 12th International Conference, LVA/ICA [A10], Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA) [A11, A13], IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) [A12].

Cykl ten zawiera prace opublikowane w dość długim okresie: od roku 2015 [A7, A10], poprzez lata 2017 [A6], 2018 [A13], 2019 [A12], 2020 [A11], 2021 [A4, A5], 2022 [A8, A9], 2023 [A2, A3] aż do roku 2024 [A1].

Prace ujętych cyklu są jedynie fragmentem większej całości. W całym dorobku publikacyjnym Habilitanta, powstałym po doktoracie, mieści się 9 artykułów w czasopismach międzynarodowych, 2 artykuły w czasopismach krajowych, 7 referatów w materiałach konferencji międzynarodowych lub anglojęzycznych. Z wyjątkiem dwóch jednoautorskich publikacji czasopiśmiennych, Habilitant publikuje prace w niewielkich zespołach współautorskich (najczęściej z jednym współautorem, najwięcej z czterema współautorami).

Tematyka publikowanych prac mieści się w większości w nurcie zastosowań metod przetwarzania sygnałów do różnorodnych zadań związanych z mową człowieka, a zwłaszcza z zagadnieniami ukierunkowanymi na realizację systemu wzmacniania sygnału mowy na potrzeby osób z wadami słuchu. Oprócz tego głównego nurtu, Habilitant ma w swoim dorobku prace dotyczące innych aspektów przetwarzania sygnałów, ze szczególnym uwzględnieniem mowy ludzkiej (patologia głosu) i dźwięków osłuchowych. Dlatego właściwe jest wskazanie dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne jako kontekstu, w którym ten dorobek naukowy podlega recenzji. Jednocześnie na podkreślenie zasługuje fakt, że Habilitant umiejętnie wykorzystuje techniki maszynowego uczenia, właściwe dyscyplinie Informatyka Techniczna i Telekomunikacja, a cel główny prowadzonych badań, związany z poprawą słyszalności sygnału mowy, mieści się w obszarze nauk medycznych (protetyki słuchu).

Pomysł na poprawę słyszalności sygnału mowy, któremu towarzyszą zakłócenia, Habilitant oparł na założeniu, że właściwy sygnał mowy, występujący wraz z innymi dźwiękami tła, powinien zostać rozpoznany, odseparowany od zakłóceń, wzmacniony, i ewentualnie powtórnie zmieszany z tłem. W prezentowanych badaniach został dodatkowo wykorzystany fakt, że sygnał mowy jest związany z wypowiedzią w języku, co oznacza, że znajomość zasad formowania dźwięków tego języka stanowi kontekst ułatwiający filtrację zakłóceń. Elementy toru przetwarzania sygnału postanowił Habilitant zrealizować za pomocą wybranych technik maszynowego uczenia – sieci neuronowych głębokich, splotowych, rekurencyjnych lub z przesuwym oknem i atencją.

Fazy rozwoju badań zgodnie z tym schematem dokumentuje cykl wskazany przez Habilitanta. Powiązania pomiędzy składnikami tego cyklu zostały przekonująco opisane w rozdziale 4.1.3 autoreferatu.

Praca [A10] opisuje metodę weryfikacji mówcy na podstawie nieujemnego rozplotu macierzowego, zastosowanego do spektrogramu magnitudowego. Zakłada się, że w spektrogramie obecna jest ważona mieszanina wzorców, które są elementami słownika mogą odpowiadać np. fonemom. Dodatkowo zakłada się, że istnieje słownik zawierający cechy wspólne dla wielu mówców, a każdy z mówców ma własny charakterystyczny słownik, którego elementy są wynikiem indywidualnej, liniowej modyfikacji wspólnego słownika. Weryfikacja mówcy w nagraniu testowym następuje po wstępnym nauczaniu modelu „indywidualnego słownika” mówcy, które polega na doborze parametrów opisujących odmiennosć w stosunku do „słownika wspólnego” poprzez minimalizację względem tych parametrów wartości dywergencji Kullbacka-Leiblera. Używana jest metoda



najszybszego spadku. W fazie testowej proces minimalizacji jest powtarzany tak, jakby tworzony był wzorzec „indywidualnego słownika” kolejnego mówcy. Weryfikacja mówcy polega na porównaniu (poprzez miarę podobieństwa cosinusowego) wektorów wartości parametrów „indywidualnych słowników” uzyskanych dla mówcy z nagrania testowego oraz dla mówców z nagrań trenujących.

Rozszerzenie tych wyników na zagadnienie identyfikacji i separacji mówców opisuje praca [A4]. Faza treningowa przebiega w podobny sposób, natomiast w fazie testowej zakłada się, że dopasowywane są równocześnie dwa wektory parametrów opisujących „indywidualne słowniki”, przy czym powinny one być możliwie najmniej podobne (w sensie miary cosinusowej).

Uczenie słownika, polegające na wyznaczeniu wzorców w spektrogramach magnitudowych, odpowiadających fonemom, może być procesem utrudnionym ze względu na niekiedy podobne spektrogramy fonemów. Habilitant wskazuje w punkcie 4.1.6 na kłopoty w formułowaniu słownika objawiające się mało charakterystycznymi wzorcami, będącymi wynikiem zastosowania metod z prac [A10, A4]. Rozwiązaniem tego problemu jest wymuszenie, aby w procesie optymalizacji wartości parametrów opisujących słownik narzucić ograniczenie, że wzorzec ze słownika jest albo obecny, albo nie występuje, podobnie jak uproszczenie przyjęte w nutowym zapisie muzyki. Praca [A6] podaje algorytm uczenia słownika wraz z jego wyprowadzeniem matematycznym, wskazuje również na korzystne zmiany w procesie rozpoznawania mówcy oraz separacji mówców.

Równolegle do nurtu prac, w którym Habilitant rozwijał metodę rozpoznawania mówcy na podstawie podobieństwa sposobu różnicowania „indywidualnych słowników” i „słownika wspólnego” oraz optymalizacji parametrów mieszaniny atomów słownika, rozwijał on metody wykorzystujące techniki bazujące na różnorodnych sieciach neuronowych.

Pierwszą pracą z tego nurtu jest [A13] gdzie Autor demonstruje sposób użycia architektury U-net – model dyfuzyjnej sieci neuronowej, stosowanej oryginalnie do odszumiania obrazu. W tej architekturze oryginalny obraz jest przetwarzany przez warstwy splotowe oraz zmniejszające rozdzielczość, po czym następują warstwy zwiększające rozdzielczość oraz splotowe. Cechą charakterystyczną sieci są „skróty”, które dokonują złączenia obrazów na poszczególnych fazach zmniejszania rozdzielczości z odpowiadającymi im obrazami powstającymi w fazach zwiększania rozdzielczości. W efekcie „użyteczny” obraz jest wzmacniany, a zakłócenia – redukowane. Habilitant przystosował tę architekturę sieci do przetwarzania krótkoczasowych transformat Fouriera. Uczenie sieci polega na minimalizacji kwadratowej funkcji straty metodą Adam. Habilitant zaproponował w pracy [A13] zastosowanie rekurencyjnej warstwy GRU (Gated Recurrent Unit) w warstwie operującej na najmocniej zredukowanej rozdzielczości.

W związku z pojawieniem się rekurencyjnej wersji architektury U-net, w której procesy splatania i redukcji rozdzielczości, bazujące na całości obrazu, są zastępowane rekurencyjnymi procesami łączącymi oba zadania, w pracy [A12] Habilitant zaproponował nową wersję architektury z [A13] w której zastąpił oryginalne warstwy splotowe warstwami rekurencyjnymi GRU w dziedzinie czasu albo w dziedzinie częstotliwości. W ten sposób proces splatania w dziedzinie czasu lub częstotliwości z oryginalnego rozwiązania (który ogranicza wpływ próbek do okna o ograniczonej długości) został zastąpiony procesem rekurencyjnego przetwarzania w czasie lub częstotliwości (pozwalającego na zwiększenie powiązania między wynikami przetwarzania a próbkami oddalonymi mocniej niż wielkość okna). Weryfikacja empiryczna wskazuje na większą efektywność tego podejścia w porównaniu z [A13] w procesie odszumiania sygnału mowy. Pogłębione badania metod z publikacji [A12] i [A13] zawiera artykuł [A8], w którym rozważa się koncepcję *Wide Context Unit* jako jednostki uogólniającą przetwarzanie poprzez splot lub warstwę rekurencyjną. Badany jest wpływ różnych sposobów realizacji WCU na efektywność architektury U-net.

Z kolei w pracy [A11] rozważana była alternatywna forma wprowadzenia rekurencyjnego przetwarzania do architektury [A13]. W tej formie, architektura U-net posiada wejście i wyjście „użytkowe” oraz wejście i wyjście dodatkowe, działające w sposób analogiczny do stanu w sieciach rekurencyjnych. Tak zmodyfikowana architektura U-net jest w procesie uczenia pobudzana wejściem

i wyjściem w postaci tensora wynikającego z przetworzenia sygnału zaszumionego i czystego, a stan jest inicjowany tensorem zerowym. Pozwala się na wielokrotną zmianę stanu. Autorzy rozważają scenariusz uczenia mający analogię do wstecznej propagacji w czasie oraz scenariusz, w którym zmiana wag następuje pod warunkiem pobudzenia konkretną wartością „stanu”. Pomysł wprowadzony w pracy [A11] jest rozwinięty w artykule [A9], w którym Autorzy realizują badanie porównawcze trzech różnych architektur przetwarzania, pracujących w identycznym układzie rekurencyjnego wywołania jak w [A9]. Porównywana jest sieć U-net (podobna jak w pracy [A13], bez rekurencji), sieć rekurencyjna ResLSTM, (wykorzystująca warstwy LSTM), a także architektura transformera. Dwie ostatnie architektury są, jak się wydaje, pomysłami własnymi Autorów, gdyż nie są cytowane żadne wcześniejsze prace. Z zamieszczonych w artykule wyników nie wynika jasno zwycięska architektura, natomiast widać, że pomysł wielokrotnego rekurencyjnego wywołania modelu w procesie uczenia korzystnie wpływa na wynik uczenia, niezależnie od wybranej architektury realizującej przetwarzanie.

W wyróżnionym cyklu publikacji można wskazać dwie prace, które w bardziej szczegółowy sposób odnoszą się do modelowania ubytku słuchu – są to prace [A1] i [A2]. Artykuł [A2] dotyczy analizy sytuacji w przypadku, gdy założony jest pewien konkretny model ubytku słuchu, a rolą systemu poprawy jakości sygnału mowy jest takie przetworzenie sygnału, aby uzyskać efekt kompensacji tego ubytku. Wprowadzony jest model tzw. *Dynamic Processing Network* (DPN), inspirowany prostymi sieciami neuronowymi, polegający na wykonaniu mieszaniny nieliniowo przetworzonych wyjść elementów banku wstępnie przygotowanych filtrów. Opcjonalnie w artykule dopuszcza się dodatkowe przetworzenie wyniku filtracji za pomocą spłotowej sieci neuronowej. Uczenie DPN polega na doborze współczynników wzmocnienia wyjść filtrów w taki sposób, aby dla zoptymalizować jakość filtracji sygnały mowy na wyjściu założonego modelu ubytku słuchu. W artykule podkreślona jest niewielka obliczeniowość zaproponowanego modelu DPN, pozwalająca na szybkie i nie wymagające dużego sprzętu przetwarzanie sygnału mowy, a także jakość uzyskanego wyniku przewyższająca sieci spłotowe, użyte do porównania.

Z kolei w artykule [A1] Habilitant koncentruje uwagę na architekturze służącej do ewaluacji modeli poprawy zrozumiałości mowy, zakładając modelowanie wady słuchu w każdym z uszu osobno. Powstałą metodykę nazywa FT-GESTOI i wskazuje na możliwość jej zastosowania do uczenia modeli poprawy jakości mowy jako funkcji straty. Funkcja ta jest różniczkowalna, co pozwala na jej sprzęgnięcie z modelami maszynowego uczenia, w większości bazującymi na algorytmie stochastycznego gradientu. Autor demonstruje na przykładach empirycznych, że wprowadzona przezeń metodyka pozwala na uzyskanie lepszych parametrów modeli niż przy użyciu dotychczas stosowanych funkcji straty.

Dodatkowo, w przedstawionym przez Habilitanta cyklu prac, znajduje się pozycja [A5], w której analizowana jest odpowiedź neurologiczna (aktywność EEG) na pobudzenia akustyczne, a także pozycja [A7], w której analizowane są możliwości rozpoznawania mówcy na podstawie cech sygnału mowy, z wykorzystaniem maszyny wektorów nośnych (SVM). Obie prace stanowią ciekawe pozycje literaturowe, ale w mojej ocenie nie wnoszą wiele do cyklu pozostałych zaprezentowanych przez Habilitanta prac. Co więcej, zgodnie ze stopką wydawcy, praca [A7] została opublikowana online w dniu 25 maja 2023, zatem w roku nadania stopnia doktora (co miało miejsce 26 lutego 2023) i w dość dużej zbieżności z tematem rozprawy doktorskiej.

Zaprezentowane powyżej subiektywne omówienie prac wskazuje na ciekawą ewolucję metod przetwarzania sygnału mowy, jaką można zaobserwować w rozwoju naukowym Habilitanta: począwszy od podejścia „tradycyjnego”, opartego na dedykowanych modelach analizy sygnału, poprzez podejście związane z maszynowym uczeniem, oparte na przystosowanych do analizy sygnału mowy architekturach sieci neuronowych, na powrót do modeli „tradycyjnych”. Wydaje się że motywację takiej ewolucji Habilitant wyraził w roboczym tytule cyklu, wskazując na „niskokosztowe metody poprawy jakości mowy”. Ten niski koszt został wyraźnie wskazany w najświeższych pracach, zwłaszcza [A2], ale przejście przez fazę dużych i bardzo zasobochłonnych modeli maszynowego uczenia stanowi dobre uzasadnienie powrotu do podejścia opartego w większym stopniu na wiedzy

dziedzinowej. Z kolei element tytułu „interpretowalne [...] metody poprawy jakości mowy” stanowi motywację wiążącą się z przeważającą większością cyklu, ale chyba największego znaczenia nabiera w związku z pracą [A1].

Przedstawiony dorobek oceniam, siłą rzeczy, przyjmując perspektywę inżyniera informatyka, a zatem przede wszystkim patrząc na aspekty związane z maszynowym uczeniem. Uważam, że dobór architektur sieci neuronowych oraz ich przystosowanie do zadania przetwarzania dźwięku świadczą o dużej biegłości Habilitanta w posługiwaniu się warsztatem badacza sztucznej inteligencji. Patrząc nieco bardziej krytycznie, brakowało mi w przejranych tekstach pogłębionej analizy porównawczej możliwych do wykorzystania architektur sieci neuronowych oraz zasad doboru ich hiperparametrów.

Z inżynierskiego punktu widzenia żałuję, że opracowane metody pozostały wyłącznie w fazie publikacji obiecujących wyników. Badania prowadzone przez Habilitanta wydają się mieć duży potencjał aplikacyjny, m.in. w obszarze wsparcia osób z ubytkami słuchu, i pożądane byłoby nie zmarnowanie tego potencjału.

Dorobek publikacyjny Habilitanta jest widoczny w skali międzynarodowej. Wg danych przygotowanych przez bibliotekę Politechniki Poznańskiej, w bazach Scopus oraz WoS jego teksty były cytowane ponad 100-krotnie (bez autocytowań). Indeks Hirscha wynosi 7 (WoS), 8 (Scopus). Indeks Hirsha w google scholar w dniu opracowania recenzji wynosił 11.

Biorąc pod uwagę całość osiągnięć naukowych przedstawionych do oceny, uważam że:

1. publikacje Habilitanta przedstawiają wyniki wartościowe i na poziomie światowym,
2. nastąpiła staranna pozytywna weryfikacja przez środowisko naukowe poprzez proces recenzji w wiodących czasopismach i konferencjach międzynarodowych,
3. przedstawione jako cykl powiązanych prac publikacje są istotnie ze sobą powiązane.

Tym samym jestem w pełni przekonany, że przedstawiony dorobek **stanowi istotny wkład** w rozwój dyscypliny Automatyka, Elektronika, Elektrotechnika i Technologie Kosmiczne. Spełnione jest zatem wymaganie z rozdziału 3, art. 219, ust.1, pkt 2 oraz 2b ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce”.

III. WNIOSEK KOŃCOWY

Recenzowany wniosek **oceniam pozytywnie**. Zgodnie z wymaganiem z rozdziału 3, art. 219, ust.1, pkt 8 ustawy z dnia 20 lipca 2018 r. „Prawo o szkolnictwie wyższym i nauce” stwierdzam, że osiągnięcia naukowe Habilitanta odpowiadają wymaganiom określonym w art. 219 ust. 1 pkt 2.

Janusz Aube