UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU
Wydział Anglistyki

Poznań, 18.08.2026

dr hab. Anna Balas, prof. UAM
Zakład Współczesnego Języka Angielskiego i Wielojęzyczności
Wydział Anglistyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

Recenzja rozprawy doktorskiej pana Stevena Jarosza pt. „A neural network approach to modeling the acoustic effects of phonological neighborhood density and word frequency in Polish speech" napisanej pod kierunkiem profesora doktora habilitowanego Arkadiusza Rojczyka (Instytut Językoznawstwa, Wydział Humanistyczny, Uniwersytet Śląski) i profesor Margherity Dore (Department of European, American and Intercultural Studies, Sapienza University of Rome,)

Przedstawiona do recenzji rozprawa liczy 141 stron. Składa się ze wstępu, sześciu rozdziałów, bibliografii oraz załącznika z listą bodźców. Pierwsze dwa rozdziały stanowią krytyczny wstęp teoretyczny i zajmują 57 stron. Pozostała część pracy poświęcona jest części empirycznej.

Celem rozprawy jest zbadanie, czy polskie samogłoski są podatne na efekty gęstości sąsiedztwa fonologicznego i częstotliwości słów, a następnie sprawdzenie, czy systemy syntezy i rozpoznawania mowy są zdolne do reprodukowania i dostrzegania systematycznego zróżnicowania artykulacyjnego samogłosek, obecnego w mowie produkowanej przez ludzi.

We wstępie Autor zarysowuje tło badawcze, opisuje cel rozprawy, identyfikuje główne luki w aktualnym stanie wiedzy, przedstawia pytania badawcze, zawęża tematykę oraz kreśli plan pracy. Rozdział pierwszy, poświęcony fonetycznym aspektom pracy, nakreśla ramy koncepcji językoznawczej, omawia polski system samogłoskowy oraz czynniki psycholingwistyczne wpływające na jakość samogłosek: gęstość sąsiedztwa fonologicznego (phonological neighborhood density) oraz frekwencję słowa (word frequency) oraz charakteryzuje korpusy językowe. Rozdział drugi jest poświęcony modelowaniu mowy przez sztuczną inteligencję. Po omówieniu podstaw sztucznej inteligencji i sieci neuronowych, Autor przedstawia główne zasady działania sieci neuronowych, współczesne zasady syntezy mowy i rozpoznawania mowy w odniesieniu do sztucznej inteligencji i modelowania języka. W rozdziale trzecim, poświęconym metodologii, najpierw omówieni są uczestnicy badania, wykorzystane modele syntezy mowy: Amazon Polly i ElevenLabs oraz model rozpoznawania mowy – wav2vec 2.0, a następnie opisane są procedury zbierania danych.

Na analizę danych składają się: pomiarów odległości euklidesowej w przestrzeni F1-F2 pomiędzy środkami samogłosek, analiza porównawcza samogłosek z łatwych i trudnych kontekstów, która mierzy odseparowanie kategorii od siebie, oraz analiza wewnątrz samogłoskowa, mierząca precyzję danej kategorii oraz kąt odchylenia pomiędzy wektorami radialnymi obliczonymi dla centroidów samogłosek w łatwym i trudnym kontekście. W kolejnym podrozdziale Autor omawia przetwarzanie danych z automatycznego rozpoznawania mowy: z końcowej warstwy transformera, aby ocenić, w jakim stopniu kodowanie przez sieć kontrastu trudnego/łatwego przypomina ludzką reprezentację akustyczną oraz z warstwy wyjściowej, żeby sprawdzić, czy poziom pewności sieci różni się w przypadku samogłosek trudnych i łatwych. Ostatnia część tego rozdziału jest poświęcona zasadom, według których przeprowadzono analizę statystyczną. W rozdziale piątym Autor prezentuje wyniki trzech komplementarnych badań, które przeprowadził, żeby zbadać akustyczne konsekwencje trudności leksykalnej w języku polskim. Wyniki wskazują na systematyczne zróżnicowanie artykulacyne obecne u użytkowników języka polskiego, ale nie w przypadku zsyntetyzowanej mowy. Natomiast model rozpoznawania mowy był wyczulony na systematyczne zróżnicowanie artykulacyjne samogłosek w ludzkiej mowie. Te wyniki pozwoliły Autorowi na odpowiedzenie na szereg pytań z pogranicza fonetyki akustycznej, psycholingwistyki i sztucznej inteligencji. Rozdział szósty jest poświęcony praktycznym zastosowaniom wyników w rozpoznawaniu mowy wyprodukowanej przez syntetyzator mowy, ulepszaniu naturalności syntezy mowy oraz kwestiom etycznym, poprawie jakości nagrań odtwarzanych w przyspieszonym tempie oraz w nauczaniu wymowy drugiego języka.

Struktura rozprawy w pełni odpowiada specyfice podjętego problemu badawczego. Autor utrzymuje porządek i równowagę pomiędzy zagadnieniami fonetycznymi a tymi związanymi z syntezą i rozpoznawaniem mowy. Struktura jest przejrzysta, a wywód prowadzony jest bardzo logicznie i konsekwentnie.

Na uznanie zasługuje imponująca, krytycznie dobrana, najnowsza literatura przedmiotu. Autor przytacza zawsze najnowsze źródła, bez zbędnych pozycji luźno powiązanych z tematem, za to wyszukując zawsze te potrzebne, dzięki którym możliwe jest zarysowanie tła teoretycznego i zadanie pytań badawczych, opracowanie metodologii badań czy zinterpretowanie wyników.

Rozprawa została starannie i konsekwentnie przygotowana pod względem formalno-edytorskim. Autor należycie stosuje techniki pisania pracy naukowej, zachowując dbałość o odpowiednie użycie technik typograficznych. Znakomicie radzi sobie z wizualizacją zjawisk, hierarchii i danych empirycznych, zachowując konsekwencję w stosowaniu tabel i wykresów. Język rozprawy jest dojrzały i właściwy dla pracy naukowej. Lista pozycji bibliograficznych jest kompletna i sporządzona w konsekwentny sposób. Mam jedynie drobne uwagi edytorskie. Symbol przedniej środkowej samogłoski powinien być ujednolicony (/e/ lub / ɛ/). Na stronie siódmej samogłoski nosowe mogłyby wystąpić w nawiasach kątowych. Tytuł podrozdziału z dołu strony 65 powinien być przeniesiony na kolejną. Na stronie 82 zamiast „makes" powinno być „make".

W rozdziałach teoretycznych zakres tematyczny omawianych kwestii jest jak najbardziej odpowiedni do postawionych celów. Doktorant znakomicie potrafi wybrać, co należy omówić i z jaką szczegółowością należy to uczynić. W sposób zwięzły i klarowny opisuje złożone kwestie teoretyczne, obecne możliwości technologiczne, zasady działania systemów syntezy i rozpoznawania mowy. Podaje przykłady. Dba także, żeby jego interdyscyplinarna praca była zrozumiała dla odbiorców z obu dziedzin: definiuje kluczowe terminy i objaśnia zależności w fonetyce i psycholingwistyce, przedstawia podstawy sieci neuronowych i głębokiego uczenia oraz pokazuje rozwój syntezy i rozpoznawania mowy aż do stanu obecnego. Znajomość fonetyki oraz syntezy i rozpoznawania mowy pozwoliła Autorowi sformułować ciekawe pytania badawcze, które są jednocześnie bardzo aktualne.

Zastosowana metodologia badań, obejmująca porównanie kilku wariantów analizy akustycznej nagrań pochodzących zarówno od uczestników badania, jak i z syntezy mowy, a także ocena rozpoznawania mowy na podstawie materiałów pozyskanych od uczestników, jest w pełni adekwatna do celów sformułowanych przez Doktoranta. Badanie samogłosek zostało zaprojektowane i przeprowadzone z dużą starannością, zgodnie z zasadami przeprowadzania eksperymentów fonetycznych z udziałem ludzi oraz badań porównawczych nad mową naturalną i syntetyzowaną. Proces badawczy został przedstawiony w sposób niezwykle skrupulatny i profesjonalny. Na uznanie zasługuje dbałość o zebranie odpowiednich danych, właściwe metody pomiarów i ich merytoryczne uzasadnienie oraz wieloaspektowa analiza wyników. Warto również podkreślić umiejętność pisania skryptów w R i Praatcie, programowania w Pythonie oraz formułowania wniosków na podstawie przeprowadzonych zaawansowanych analiz statystycznych. Zastosowane analizy akustyczne i statystyczne pozwoliły udzielić odpowiedzi na postawione pytania badawcze. Na szczególne podkreślenie zasługuje nowatorskie podejście do badania rozpoznawania mowy z wykorzystaniem wewnętrznych aktywacji i miar pewności modelu.

Ostatni rozdział uważam za szczególnie cenny. Autor wylicza sporo zastosowań wyników swoich badań. Zauważa, że wyniki jego badań zachęcają, żeby widzieć zastosowanie systematycznego artykulacyjnego zróżnicowania jako cechy typowej dla mowy ludzkiej do wykrywania mowy będącej wynikiem automatycznej syntezy. Wyniki są także podpowiedzią, jak zoptymalizować syntezę mowy w taki sposób, żeby efekt był bardziej naturalny, jak poprawić zrozumiałość nagrań odtwarzanych w przyspieszonym tempie, a także jak usprawnić formowanie odpowiednich kategorii fonetycznych w drugim języku. Każde z tych zastosowań jest wnikliwie omówione. Doktorant udowodnił, że nie tylko potrafi przeprowadzać interdyscyplinarne badania, ale także ma szerokie horyzonty, wie, w jaki sposób można wykorzystać wyniki badań przy dzisiejszym stanie rozwoju technologicznego i że nieobce są mu kwestie etyczne.

Gorąco zachęcam Autora rozprawy do jej opublikowania. Rozważyłabym Cambridge Elements dla publikacji o charakterze naukowym, ale też czasopisma lub strony internetowe popularyzujące naukę. Poniżej przedstawiam kilka uwag do rozważenia w przypadku zgłoszenia pracy

do publikacji. Nie ujmują one wartości przedstawionej do oceny pracy, lecz są zaproszeniem do dyskusji.

Zastanawiam się, dlaczego Autor wybrał tytuł „A neural network approach to modeling the acoustic effects of phonological neighborhood density and word frequency in Polish speech". Praca bada, czy zaobserwowane u ludzi akustyczne efekty gęstości sąsiedztwa fonologicznego i częstotliwości słów mają swoje odzwierciedlenie w mowie zsyntetyzowanej i w automatycznym rozpoznawaniu mowy. Moim zdaniem tytuł zdaje się jednak sugerować, że wspomniane efekty akustyczne są w jakiś sposób modelowane przez sieci neuronowe lub że można je wspomóc. Jak pokazują wyniki, w syntezie mowy efektów akustycznych gęstości sąsiedztwa fonologicznego i częstotliwości słów nie znaleziono.

W pierwszym rozdziale Autor udowadnia, że jest świadomy wyników badań nad angielskimi samogłoskami oraz różnic pomiędzy polskimi i angielskimi samogłoskami, ale nie jestem pewna, czy konieczne było poświęcanie podrozdziałów zagadnieniom dynamiki widmowej (vowel inherent spectra change) i cichego centrum (silent center), skoro w żaden sposób nie są one uwzględnione w projekcie badania polskich samogłosek w pracy. Natomiast być może podrozdział 1.1.2 Neighborhood Activation Model zyskałby na wartości, gdyby przytoczone zostały badania testujące model aktywacji sąsiedztwa, tak żeby przybliżyć stosowaną w nich metodologię.

W pracy znajdują się opisy Narodowego Korpusu Języka Polskiego (Przepiórkowski 2012) i korpusu SUBTLEX, które nie zostają wykorzystane w badaniu, oczywiście słusznie, ponieważ, jak Autor pokazuje, nie zawierają danych na temat gęstości sąsiedztwa fonologicznego. Natomiast sądzę, że przydałby się trochę bardziej wyczerpujący opis bazy danych i kalkulatora miar sąsiedztwa Jiwar (Alzahrani 2025).

W sekcjach dotyczących modelowania mowy z pomocą sztucznej inteligencji dobrze by było może dodać więcej odnośników bibliograficznych, np. na stronie 27.

W podrozdziale 3.1 Data Stimuli mogło się znaleźć więcej danych dotyczących wyznaczania gęstości sąsiedztwa fonologicznego i częstotliwości słów.

Zanonimizowane nagrania, skrypty w Praatcie i R mogłyby być udostępnione w OSF.

Chciałabym też zadać trzy pytania.

1) Skoro syntetyzowane bodźce samogłoskowe musiały zostać nagrane w zdaniach, to dlaczego ludzie nagrywali tylko pojedyncze słowa? W jaki sposób te dwa sposoby pozyskiwania danych mogły wpłynąć na wyniki badań?
2) Do pomiarów dystansów pomiędzy wariantami jednej samogłoski użyto dystansu euklidesowego. Jakie wady/zalety miałoby użycie wyniku Pillai lub dystansu Mahalanobisa?
3) Czy w świetle uzyskanych wyników, zasadne wydaje się używanie Polly lub ElevenLabs w nauczaniu języków obcych, a w szczególności ich wymowy?

Moja ogólna ocena pracy jest bardzo wysoka. Rozprawa doktorska przedłożona przez Pana Stevena Jarosza z naddatkiem spełnia wymogi art. 187. p. 1—2 ustawy Prawo o szkolnictwie wyższym i nauce z dnia 20 lipca 2018 roku, zgodnie z którymi rozprawa doktorska powinna prezentować „ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach oraz umiejętność samodzielnego prowadzenia pracy naukowej (...)", a jej przedmiotem ma być „oryginalne rozwiązanie problemu naukowego (...)". Doktorant zaprezentował dużą wiedzę i znakomity warsztat prowadzenia eksperymentalnych badań fonetycznych oraz w zakresie automatycznej syntezy i rozpoznawania mowy. Obrał ambitne cele oraz dobrał do nich zaawansowane metody analizy fonetycznej i statystycznej. Z całym przekonaniem stwierdzam, że praca stanowi oryginalny i znaczący wkład w rozwój badań nad fonetycznymi aspektami języka polskiego, syntetyzowanej mowy, a także nad rolą detali fonetycznych w automatycznym rozpoznawaniu mowy oraz wnoszę o dopuszczenie do publicznej obrony. Stwierdzam ponadto, że rozprawa doktorska Pana Stevena Jarosza powinna być wyróżniona i nagrodzona. Wybitna wartość naukowa rozprawy wyraża się w jej interdyscyplinarnym charakterze oraz we wzorcowym zastosowaniu wieloaspektowej analizy mowy ludzkiej i syntetyzowanej, a także nowatorskim podejściu do badania automatycznego rozpoznawania mowy i udzielaniu kompleksowych odpowiedzi na aktualne pytania związane z rozwojem sztucznej inteligencji. Wyniki są bardzo ciekawe, a konkluzje mają potencjał, jak sam Autor wylicza, do zastosowania w ulepszaniu naturalności mowy zsyntetyzowanej, w obróbce dźwięku oraz w nauczaniu wymowy języka polskiego/języków obcych. Praca porusza też ważne kwestie etyczne związane z rozwojem mowy produkowanej lub rozpoznawanej przez sztuczną inteligencję. Profesjonalna dyskusja nad fonetycznymi aspektami w technologii mowy jest nam bardzo potrzebna.

Anna Balas

recenzja_Jarosz.pdf

recenzja_Jarosz.pdf

Dokument został podpisany elektronicznie zgodnie z Rozporządzeniem Unii Europejskiej eIDAS. Podpisy elektroniczne, integralność i autentyczność dokumentu zostały zabezpieczona pieczęcią elektroniczną. Dokładna data złożenia podpisu na dokumencie przez każdą ze stron została oznaczona kwalifikowanym znacznikiem czasu.

Podpisujący:

PODPIS ZAAWANSOWANY

ANNA AGNIESZKA BALAS
01:26 | 18.08.2026 (UTC+02.00) |

eIDAS

Usługa została wykonana zgodnie z Rozporządzeniem Parlamentu Europejskiego i Rady (UE) nr 910/2014 z dnia 23 lipca 2014 r. w sprawie identyfikacji elektronicznej i usług zaufania w odniesieniu do transakcji elektronicznych na rynku wewnętrznym oraz uchylające dyrektywę 1999/93/WE (eIDAS).