prof. dr hab. inż. Paweł Strumiłło
Instytut Elektroniki
Politechniki Łódzkiej

WYDZIAŁ ELEKTRONIKI
TELEKOMUNIKACJI I INFORMATYKI
Politechniki Gdańskiej
RPH 1081/2023
Wpłynęło dnia 09.01.2023

3 stycznia 2023 r.

RECENZJA ROZPRAWY DOKTORSKIEJ
DLA RADY DYSCYPLINY NAUKOWEJ INFOMATYKA TECHNICZNA
I TELEKOMUNIKACJA POLITECHNIKI GDAŃSKIEJ

Tytuł rozprawy: Automatyczna detekcja błędów wymowy z wykorzystaniem głębokiego uczenia maszynowego w celu wsparcia nauki języka

Autor rozprawy: mgr inż. Daniel Korzekwa

Promotorzy: prof. dr hab. inż. Bożena Kostek

I. Jakie zagadnienie naukowe jest rozpatrzone w pracy /teza rozprawy/ i czy zostało ono trafnie i jasno sformułowane?

Przedłożona do recenzji rozprawa doktorska dotyczy zagadnienia wykrywania błędów w wymowie i treningu wymowy (ang. Computer-Assisted Pronouncation Training – CAPT) u osób uczących się języka angielskiego jako języka obcego z zastosowaniem technik komputerowych (ang. Computer Assisted Language Learning – CALL). W szczególności w języku angielskim występują elementy wymowy nieistniejące w innych językach, reguły wymowy w tym języku są trudne do przyswojenia oraz występuje duża zmienność w prawidłowej wymowie. Z tego też powodu, jak zaznacza Autor we wstępie pracy, istniejące dotychczas systemy wykrywania błędów wymowy nie są wystarczająco dokładne. Istnieje zatem potrzeba prowadzenia badań nad udoskonalaniem komputerowych technik automatycznego wykrywania błędów wymowy. W związku z tak określonym zagadnieniem naukowym podstawowa teza rozprawy została sformułowana w następującym brzmieniu:

Możliwe jest zwiększenie dokładności metod uczenia głębokiego do wykrywania błędów wymowy w nierodzimej mowie angielskiej poprzez zastosowanie syntetycznego generowania mowy i bezpośredniej detekcji błędów typu end-to-end, które zmniejszają zapotrzebowanie na nagrania i fonetyczną transkrypcję mowy.

Postawiono również w pracy dodatkową tezę badawczą, w której postuluje się możliwość zastosowania wypracowanych technik komputerowych do wykrywania tzw. mowy dyzartrycznej (mowy z zaburzeniami motorycznym aparatu mowy):

Metody uczenia głębokiego służące do wykrywania błędów wymowy w nierodzimej mowie angielskiej można przenieść na pokrewne zadania wykrywania i rekonstrukcji mowy dyzartrycznej.

Tezę tę należy uznać za interesujące rozszerzenie obszaru badań i pokazania dużego potencjału technik uczenia maszynowego.

Uważam, że zagadnienie naukowe podjęte w pracy zostało konkretnie i jasno sformułowane.

II. Czy autor rozwiązał postawiony problem i czy użył do tego właściwych metod dowodząc, że posiadł umiejętności związane z metodyką i metodologią prowadzenia badań naukowych?

Podstawy teoretyczne i metodologiczne technik obliczeniowych opracowanych i zastosowanych w rozprawie omówiono szczegółowo w 2. rozdziale rozprawy. Opis ten obejmuje techniki uczenia maszynowego ze szczególnym uwzględnieniem tzw. uczenia głębokiego oraz definicje miar dokładności wykrywania błędów wymowy i miary oceny błędów wymowy w mowie syntezowanej.

Kolejne dwa rozdziały pracy tj. rozdziały 3. i 4. poświęcono wykazaniu odpowiednio tezy 1. i 2. rozprawy.

Rozdział 3. pracy przedstawiono w układzie, w którym zamieszczono obszerne streszczenia osiągnięć badawczych Autora opublikowanych we współautorskich referatach i artykułach naukowych, które scharakteryzowano poniżej:

Ad. *Weakly-Supervised Word-Level Pronunciation Error Detection in Non-Native English Speech*
Zaproponowano oryginalny tzw. słabo nadzorowany model do wykrywania błędnej wymowy na poziomie słów w nierodzimej (L2) mowie angielskiej. Do trenowania tego modelu nie wykorzystano transkrypcji fonetycznej mowy, a jedynie zaznaczanie błędnie wymówionych słów. Zatem model jest trenowany tylko na podstawie sygnału o błędach wymowy na poziomie słów. Zaproponowano wielozadaniową konfigurację trenowania modelu, w której szacowane jest prawdopodobieństwo błędnej wymowy na poziomie słowa oraz wykorzystano wyniki rozpoznawania fonemów dla fonetycznie transkrybowanej mowy L1 (mowa rodzima), która jest łatwo dostępna i może być automatycznie etykietowana. W porównaniu z istniejącymi metodami, poprawiono dokładność wykrywania błędów wymowy na poziomie słów wg miary AUC o 30% dla korpusu GUT L2 polskich mówców oraz o 21.5% dla korpusu L2 niemieckich i włoskich mówców.

Ad. *Mispronunciation Detection in Non-Native (L2) English with Uncertainty Modeling*
Szeroko stosowaną metodą automatycznego wykrywania błędnej wymowy jest porównanie fonemów wypowiadanych przez ucznia do wymowy mówcy rodzimego i porównywanie ich z oczekiwaną wymową rodzimego mówcy. Wadą takiego podejścia badawczego jest duży wskaźnik fałszywych alarmów z uwagi na wiele możliwych sposobów wypowiedzi. Zaproponowano nowatorskie podejście do tego problemu, w którym uwzględniono: 1) niepewność automatycznego rozpoznawania fonemów oraz 2) fakt istnienia wielu poprawnych sposobów wymowy. Uzyskany model poddano weryfikacji dla nierodzimej mowy angielskiej dla niemieckich, włoskich i polskich mówców. Uzyskano precyzję wykrywania błędnej wymowy większą o 18% w porównaniu z tradycyjnie stosowanym podejściem.

Ad. *Detection of Lexical Stress Errors in Non-Native (L2) English with Data Augmentation and Attention*
Opracowano nowe metody wykrywania błędów wymowy, które umożliwiły zwiększenie dokładności wykrywania błędów leksykalnych w nierodzimej mowie angielskiej. Zaproponowano model głębokiego uczenia oparty na modelu uwagi (ang. *Attention-Based Classification Model*) oraz augmentację danych z wykorzystaniem techniki Text-To-Speech (TTS). Model uwagi polega na mapowaniu cech sygnału mowy z poziomu ramek i fonemów na poziom sylab. Jest to nowe

podejście w odróżnieniu od klasycznych technik, w których cechy sygnału mowy są wyodrębniane dla ustalonych fragmentów sygnału mowy. Zastosowanie tego nowego podejścia obliczeniowego pozwala osiągnąć 95% miary precyzji i 49% odtwarzalności podczas wykrywania niepoprawnie akcentowanych słów w nierodzimej mowie angielskiej L2 u mówców słowiańskich i z rejonu Wybrzeża Bałtyku.

Ad. *Computer-assisted pronunciation training – Speech synthesis is almost all you need*
W związku z coraz szerzej stosowanymi technikami ucznia głębokiego występuje potrzeba zapewnienia dużej liczby przykładów poprawnie i błędnie wymawianej mowy. Ograniczenie to jest głównym powodem trudności w opracowaniu skutecznych metod CAPT wykrywania błędów wymowy. W pracy zaproponowano innowacyjne techniki z zastosowaniem konwersji fonemów na fonemy (P2P), tekstu na mowę (T2S) oraz mowy na mowę (S2S), służące do wytwarzania poprawnie i błędnie wymawianej mowy syntezowanej. Umożliwiły one budowanie bogatego zestawu danych treningowych. Skuteczność tych technik weryfikowano w zadaniach wykrywania błędów wymowy i stresu leksykalnego w nierodzimej mowie angielskiej pozyskanej od mówców z Włoch i Polski. Najlepsza z zaproponowanych technik, tj. S2S poprawia dokładność wykrywania błędów wymowy dla miary AUC (pole pod krzywą operacyjno-charakterystyczną) o 41% (z 0,528 do 0,749) w porównaniu do innych technik udokumentowanych w literaturze.

W rozdziale 4. pracy podobnie jak w rozdziale 3. zamieszczono obszerne omówienie referatu odnoszącego się do zadania badawczego związanego z drugą tezą pracy, tj. wykrywaniem i rekonstrukcją mowy dyzartrycznej.

Ad. *Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech*
Zaproponowano oryginalną metodę trenowania modelu neuronowego za pomocą techniki jednoczesnego trenowania zadań wykrywania dyzartrii i rekonstrukcji mowy. Zastosowany model enkodera-kodera pozwala rozłożyć na cechy mowy czynniki, które można poddać łatwej interpretacji w odniesieniu do mowy dyzartrycznej (m.in. zrozumiałość mowy i jej płynność). Pokazano, że zastosowana wielozadaniowa technika uczenia modelu i reprezentacja widma mowy w skali melowej pozwala uzyskać zasadniczą poprawę dokładności wykrywania mowy dyzartrycznej. Praca ta może mieć konkretne zastosowania w rehabilitacji pacjentów dotkniętych dyzartrią.

Podsumowując ww. osiągnięcia badawcze zebrane w 3. i 4. rozdziale rozprawy należy stwierdzić, że Autor zastosował różnorodne metody przetwarzania i reprezentacji sygnałów cyfrowych, analizy statystycznej oraz najnowsze techniki uczenia maszynowego z wykorzystaniem uczenia głębokiego. Wykazał się pogłębioną znajomością tematyki badań i zaproponował innowacyjne metody generowania mowy syntezowanej (poprawnie i błędnie wymawianej) i ich augmentacji. Opracowane modele poddał weryfikacji dla danych pozyskanych od mówców, dla których język angielski jest językiem obcym.

Biorąc pod uwagę skuteczność opracowanych oryginalnych modeli wykrywania błędów wymowy dla nierodzimej mowy angielskiej należy jednoznacznie stwierdzić o udowodnieniu przez Autora tez pracy, a mianowicie:

- wykazano, że zastosowane technik bezpośredniego wykrywania błędów wymowy oraz zamiany mowy na mowę syntezowaną z błędami wymowy umożliwiło poprawę miar dokładności rozpoznawania błędów wymowy przy jednoczesnym zmniejszeniu

zapotrzebowania na liczbę nagrań wykorzystywanych do trenowania modeli CAPT, co dowodzi pierwszej i zasadniczej tezy pracy,

- wykazano, że metody uczenia głębokiego umożliwiają wykrywanie i rekonstrukcję mowy dyzartrycznej z dużą skutecznością, tym samym udowodniono drugą, pomocniczą tezę rozprawy.

III. Czy tematyka rozprawy jest aktualna i dostatecznie ważna?

Z uwagi na podkreśloną w pracy małą dokładność istniejących metod wykrywania błędów wymowy oraz powszechność języka angielskiego jako języka nierodzimego tematyka rozprawy jest ważna i aktualna. Dla przykładu w Chinach liczbę osób uczących się języka angielskiego szacuje się na ok. 400 mln.

IV. Na czym polega oryginalny dorobek autora i jakie jest jego znaczenie poznawcze lub przydatność praktyczna dla nauki bądź techniki?

Do zasadniczych osiągnięć naukowych Autora rozprawy stanowiących istotny wkład do dyscypliny Informatyka Techniczna i Telekomunikacja zaliczam:

1. Opracowanie nowej techniki bezpośredniego wykrywania błędów na poziomie wyrazów bez konieczności wykonywania transkrypcji fonetycznej mowy, która umożliwia zwiększenie dokładności wykrywania błędów o ok. 30% w porównaniu do dotychczas opracowanych metod. Dodatkowa zaletą niniejszego podejścia jest zmniejszenie zapotrzebowania na liczbę nagrań i fonetyczną transkrypcje mowy wykorzystywanych w budowaniu modeli komputerowych CAPT.
2. Zaproponowanie oryginalnego podejścia do wykrywania błędów wymowy na podstawie tzw. mechanizmu uwagi, który pozwala zdekomponować schemat uczenia sieci na niezależne zadania.
3. Uwzględnienie w procesie wykrywania błędów wymowy różnych dialektów języka, w których zdania mogą być wymawiane wg różnych akcentów leksykalnych oraz składać się z różnej sekwencji fonemów i czasu ich trwania. Podejście to umożliwiło poprawę precyzji rozpoznawania o ok. 18% w porównaniu do opracowanych dotychczas metod.
4. Zdefiniowanie problemu wykrywania błędów wymowy jako problemu syntezy błędnie wymawianej mowy. Również ta technika umożliwiła poprawę dokładności wykrywania błędów wg metryki AUC z wartości 0,53 do 0,75 oraz zmniejszyła zapotrzebowanie na liczbę danych wymaganych do trenowania modeli.

Ww. osiągnięcia badawcze Autora rozprawy zostały udokumentowane we współautorskich publikacjach naukowych z zasadniczym wkładem merytorycznym Autora (scharakteryzowanym szczegółowo w rozdz. 1. pracy). Dorobek ten jest bardzo bogaty jak na wymagania stawiane kandydatom do stopnia naukowego doktora w dziedzinie nauk inżynieryjno-technicznych. Sześć prac jest bezpośrednio związanych z głównymi osiągnięciami badawczymi Autora rozprawy, z których cztery to referaty konferencyjne a dwie to publikacje w czasopismach naukowych o dużym współczynniku wpływu IF.

Wyniki niniejszej rozprawy mogą znaleźć praktyczne zastosowanie w rozwoju i wdrożeniu komputerowych metod CALL do wykrywania błędów wymowy i nauce poprawnej wymowy u osób uczących się języka angielskiego jako języka obcego. Innym obszarem zastosowań jest rehabilitacja pacjentów z zaburzeniami mowy.

V. Czy rozprawa świadczy o dostatecznej wiedzy autora, wiedzy na zaawansowanych poziomie, o charakterze podstawowym dla dziedziny nauk technicznych oraz o charakterze szczegółowym, odpowiadającej obszarowi prowadzonych badań naukowych

Niniejsza praca doktorska ma charakter interdyscyplinarny, gdyż oprócz opracowania i stosowania odpowiednich technik uczenia maszynowego Autor pogłębił wiedzę z zakresu anatomii aparatu mowy i fizjologii mowy. Wykazał się przede wszystkim pogłębioną wiedzą i umiejętnościami uczenia maszynowego. Zamieścił w pracy obszerną sekcję poświęconą zagadnieniom budowania trenowania i zastosowań sieci neuronowych o różnych architekturach (sieci Bayesowskich i sieci głębokich, w tym rekurencyjnych). Również na odpowiednim poziomie szczegółowości dokonał przeglądu modeli probabilistycznych w tym procesów Gaussowskich i ukrytych modeli Markowa. Opisał też podstawowe miary oceny jakości decyzji, np. klasyfikacyjnej.

Warto zaznaczyć, że Autor biegle posługuje się terminologią angielską w obszarze prowadzonych badań (w tym również polską jak można wnioskować na podstawie zamieszczonego obszernego streszczenia pracy w języku polskim).

VI. Czy rozprawa obejmuje najnowsze osiągnięcia nauki i świadczy o znajomości współczesnej literatury z dyscypliny naukowej, której dotyczy?

W pracy nie ma oddzielnej sekcji poświęconej przeglądowi literatury tematu. Autor omówił doniesienia literaturowe w autorski sposób polegający na wskazaniu głównych problemów badawczych obejmujących tematykę pracy, do których należą m.in. trudności związane z: transkrypcją mowy obcej, dokładnym dopasowaniem fonemów, oceną istotności błędów, różnymi poprawnymi sposobami wymowy wyrazów i zdań, ograniczeniem dostępności przykładów mowy nierodzimej zawierającej błędy. Autor pokazał odwołania do pozycji literatury z ostatnich kilku lat, które podejmują ww. problemy badawcze. Autor wskazał również istotne cele (także z cytowaniem innych prac) jakie stawia się technikom uczenia głębokiego w rozważanej tematyce, tj. konieczność wykrywania zarówno błędów wymowy fonemów jak i błędów akcentu leksykalnego oraz możliwość rozpoznawania fonemów, co pozwoli zabezpieczyć techniki uczenia przed nadmiernym dopasowaniem do danych.

Obszerny i wielowątkowy przegląd aktualnej literatury zamieszczono również w obszernych opisach współautorskich opublikowanych pracach naukowych poświęconych wykrywaniu błędów w mowie nierodzimej.

VII. Jakie są wady i słabe strony rozprawy?

Trudno doszukać się wad lub słabych stron rozprawy. Praca jest napisana bardzo dobrym językiem angielskim zawiera obszerne streszczenie rozprawy w języku polskim bardzo dobrze podsumowujące istotę osiągnięć badawczych Autora rozprawy.

Moje drobne uwagi i sugestie do pracy są następujące:

Tytuł rozprawy:

Zauważono, że tytuł rozprawy w języku angielskim nie jest precyzyjnie przetłumaczony na język polski. W polskim tytule zabrakło informacji, że praca dotyczy wykrywania błędów w mowie nierodzimej angielskiej. Polski tytuł rozprawy mógłby być zapisany np. w brzmieniu:

Automatyczna detekcja błędów wymowy w nierodzimej mowie angielskiej z wykorzystaniem głębokiego uczenia maszynowego

Na stronicy 22. na początku sekcji 2.1.2 zauważono zdanie: *Energy is defined as the variance of speech signal*, które nie jest prawdziwe. Podobnie występujące w następnych zdaniach stwierdzania: *Energy is usually expressed in decibels (dB)...*, a także *Energy in dB is called loudness* nie są prawdziwe. Proszę Autora o poprawne definicje wymienionych pojęć.

VIII. Do której z następujących kategorii Recenzent zalicza rozprawę?

Rozprawę zaliczam do kategorii wyraźnie wykraczającej ponad poziom przeciętny i spełniającą z nadmiarem wymagania stawiane rozprawom doktorskim w dziedzinie nauk inżynieryjno-technicznych.

Zgłaszam rozprawę do wyróżnienia. Praca spełnia wymaganie formalne uchwalone przez Radę Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Gdańskiej stawiane rozprawom doktorskim. Autor rozprawy opublikował dwa współautorskie artykuły bezpośrednio związane z rozprawą doktorską w czasopismach naukowych pt. *Speech Communication Journal* i *Journal of the Acoustic Society of America*. Czasopisma te są ujęte w *Wykazie czasopism naukowych i recenzowanych materiałów konferencji międzynarodowych wraz z przypisaną liczbą punktów* odpowiednio 140 i 100 pkt. MEiN. W obydwu pracach Kandydat jest pierwszym autorem i ma zasadniczy koncepcyjny i merytoryczny wkład do udokumentowanych w tych pracach badań. Mój wniosek o wyróżnienie rozprawy uzasadniam, istotnym wkładem badawczym do dyscypliny naukowej Informatyka Techniczna i Telekomunikacja w zakresie analizy sygnałów mowy w kontekście wykrawania błędów wymowy i nauki prawidłowej wymowy w nierodzimej mowie angielskiej, a mianowicie:

- nowatorskim charakterem pracy, w której zaproponowano i poddano weryfikacji nowe metody wykrywania błędów wymowy jako zadanie generowania mowy syntezowanej, w tym opracowanie nowej techniki bezpośredniego wykrywania błędów typy end-to-end co umożliwia znaczące zmniejszenie zapotrzebowania na liczbę danych treningowych do budowania modeli neuronowych oraz umożliwia zwiększenie dokładności wykrywania błędów w nierodzimej mowie angielskiej,
- możliwością praktycznego zastosowania opracowanych metod uczenia głębokiego do wykrywania mowy dyzartrycznej.

Wniosek końcowy

Stwierdzam, że przedłożona do recenzji rozprawa doktorska spełnia z nadmiarem wymagania Ustawy o stopniach naukowych i tytule naukowym z 14 marca 2003 roku, Dziennik Ustaw Nr 65, poz. 595 z późn. zm. oraz Rozporządzeniem Ministra Nauki i Szkolnictwa Wyższego z dnia 19 stycznia 2018 roku w sprawie szczegółowego trybu i warunków przeprowadzania czynności w przewodach doktorskich, postępowaniu habilitacyjnym oraz w postępowaniu o nadanie tytułu profesora. Wnioskuję do Rady Dyscypliny Naukowej Informatyka Techniczna i Telekomunikacja Politechniki Gdańskiej o dopuszczenie rozprawy doktorskiej mgr inż. Daniela Korzekwy do publicznej obrony.

Paweł Strumiłło