RADA NAUKOWA DYSCYPLINY
INFORMATYKA TECHNICZNA I TELEKOMUNIKACJA POLITECHNIKI WARSZAWSKIEJ

zaprasza na

PUBLICZNĄ OBRONĘ ROZPRAWY DOKTORSKIEJ

mgr. inż. Kacpra Radzikowskiego

która odbędzie się w dniu **20 czerwca 2022 roku o godzinie 11:00** w trybie hybrydowym*

Temat rozprawy doktorskiej:

„ A study on speech recognition and correction for non-native English speakers ”

Promotor: dr hab. inż. Robert Nowak - Politechnika Warszawska

Recenzenci: prof. dr hab. inż. Bożena Kostek - Politechnika Gdańska

prof. dr hab. inż. Krzysztof Ślot - Politechnika Łódzka

* Obrona odbędzie się w Sali nr 116* oraz zdalnie na platformie MS Teams. Osoby zainteresowane uczestnictwem w obronie proszone są o zgłoszenie chęci uczestnictwa w formie elektronicznej na adres sekretarza komisji: dr hab. inż. Piotr Gawrysiak, prof. uczelni – email : piotr.gawrysiak@pw.edu.pl do dnia 17 czerwca 2022 do godz. 17:00.

Z rozprawą doktorską i recenzjami można zapoznać się w Czytelni Biblioteki Głównej Politechniki Warszawskiej, Warszawa, Plac Politechniki 1.

Streszczenie rozprawy doktorskiej i recenzje są zamieszczone na stronie internetowej: https://bip.pw.edu.pl/Postepowania-w-sprawie-nadania-stopnia-naukowego/Doktoraty/Wszczete-do-30-kwietnia-2019-r/Dyscyplina-informatyka-techniczna-i-telekomunikacja-dziedzina-nauk-inzynieryjno-technicznych/mgr-inz.-Kacper-Radzikowski.

Przewodniczący Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej

dr hab. inż. Jarosław Arabas

Abstract

The educational services market is currently undergoing dynamic and diverse changes, mostly due to the advancement of the Internet and convenient communication methods. E-learning is a whole new style of learning that involves the use of electronic media, most notably the Internet. Furthermore, it allows students to learn from people from many cultures, countries, and geographic zones. In such situations, a discourse (i.e. lesson) between several individuals of different nationalities is possible. It is obvious that e-learning is only one of multiple examples of situations where participants represent many different nationalities and linguistic backgrounds. Others might include work and business related conference calls between employees of international companies, or any other interactions where the commonly used language is not the first language for the participating speakers.

Such circumstances need the use of a common language that is not the native tongue of at least one person to the conversation. Communication between two or more persons using a common language necessitates that all participants be proficient in that language. As a result, it may be useful for all parties involved in the encounter to employ a variety of accessible tools to improve communication and mutual understanding. Automatic speech recognition (ASR) techniques, for example, can be used to provide real-time closed captioning for the interaction. It is relatively simple to develop a support software for an e-learning platform that could use any existing cloud-based ASR system. However, the effectiveness of cloud ASR systems varies significantly depending on whether the sound samples represent the speech of a native or non-native speaker. In essence, the non-native speech recognition problem relates to the situation where the speaker's mother tongue is different from the language he speaks at the time of the speech recognition process. For example, we might consider a Japanese person speaking English and using the English language ASR system.

Supervised learning techniques are commonly used in traditional approaches to developing speech recognition classifiers. While this approach is ideal for recognizing speech in the majority of the world's languages, it will not produce good classifiers for non-native speakers.

The fundamental reason for this is a lack of large labeled datasets of non-native speech to be used as a training set in a supervised learning system. ASR technologies are becoming more accurate and depending on the benchmark used, can achieve up to 90-95 percent accuracy. However, such high levels of accuracy can only be achieved when the system is used to recognize native speakers' speech (e.g. English language for North American people).

The rationale for the ASR's lower accuracy is the presence of patterns connected to the speaker's

mother tongue that can impact the grammar and accent traits of the second language. In order to solve the aforementioned issues it is possible to focus on the problem either from the perspective of optimizing the process of creating ASR models adapted for non-native speech or approach the problem by trying to modify (convert) the speech (accent) which is problematic for the existing ASR systems to process correctly. The current rate of globalization requires the effective recognition of non-native speakers, who today account for the vast majority of users. The research presented in this dissertation focuses on automatic speech recognition for non-native speakers. This thesis represents the research studying the aforementioned issues and is organized into five chapters.

Chapter 1 describes in detail the problem of non-native speech recognition, especially in the context of non-native speakers. It describes in detail the main reasons behind the differences between speech recognition systems utilized by native speakers and the same process involving non-native speakers. In this chapter we also underline the several dimensions at which the native and non-native speech might differ and explain the rationale behind this logic. Furthermore we provide the motivation for conducting this research. We also state the purpose of this dissertation and represent it as a problem in the machine learning domain, and as such we approached it using developed algorithms. We outline the dissertation's layout and describe the content of each chapter.

Chapter 2 provides a summary of extensive research conducted on the problem of speech recognition for both native and non-native speakers. There have been multiple approaches evaluated by researchers that come from a wide variety of national and linguistic backgrounds. The research done by them involves multiple languages and nationalities of the non-native speakers. In this chapter we explain that while the research that had been done in the past yields an incremental progress, the scalable solution to the non-native speech recognition problem is still to be found and developed.

Chapter 3 describes a research conducted in order to address the problem of non-native speech data scarcity from the point of view of creating an ASR adapted for non-nartive speakers, from the ground up. The methodology designed and evaluated within the scope of this chapter represents an approach called dual supervised learning. The idea presents a setup of slightly modified actor-critic model adapted for training speech recognition models. We designed a custom feedback loop including a language model and speech model acting as critics, and speech-to-text with speech synthesis models becoming actors. The method focuses on training and updating the state of actor models while preserving the state of the critic models. Critic models had been trained before using a set of unlabelled non-native speech samples, that usually come in large amounts. Actor models are trained using the aforementioned feedback loop setup. The experiments conducted within the scope of this

chapter prove that it is possible to leverage the much greater availability of unlabelled datasets in order to create an ASR system adapted for non-native speech. In this chapter we also described a method for creating a conversion methodology between sentences produced by non-native speakers of any particular language in his or her second language and corresponding sentences as if they were formed by native speakers of the language. The methodology plays a supplementary role to the dual supervised learning method described above. The approach uses a sequence-to-sequence encoder-decoder setup which realizes the conversion between sentence pairs. It plays the role of a language model which has been widely utilized in ASR techniques and as such, is typically employed right after the ASR output was received for a non-native speech sample. The language model becomes a support mechanism which is adapted and trained for examples of sentences produced by non-native speakers of a particular language. We evaluated the idea using an experimental setup designed to compare the model trained using dual supervised learning to the baseline model trained using a traditional supervised approach. The experiment yielded a 3 % increase in the model accuracy when trained with a dual supervised approach compared to the model trained with the traditional way.

Chapter 4 represents research conducted on methodologies for real-time modification of accent for non-native speech samples. The chapter studies the performance of already existing ASR solutions when applied to non-native speech. We presented the problems that often affect the speech-to-text systems upon being used by a non-native speaker. To overcome the problems we designed the method to modify the non-native speech samples on-the-fly, as a data adaptation technique with the purpose of increasing the ASR performance with respect to non-native speech. The chapter presents the methodology which is based on the neural style transfer approach used in the context of speech and sound, instead of graphics and image. The approach involves creating an algorithm based on convolutional networks to extract certain information related to style (accent) and content (phonemes, words). The style transfer idea allows us to extract such information from non-native and native speech samples as well as a sample created by our speech style converter. Next step involves designing the loss function that minimizes the style difference between these samples, thus allowing the converted sample to resemble the native speech style to a higher extent. The audio style transfer solution was evaluated with the experimental setup that involved a Google Cloud Speech to Text service as a baseline ASR model. We evaluated its accuracy using the dataset of Japanese students speaking English. It was compared to the experiment which involved executing style transfer right before the converted samples were fed into the Google Cloud ASR. The comparison yielded a 40 %

relative improvement. The idea presented in the chapter provides the possibility of using already productionalized ASR systems for recognition of non-native speech not constrained by nationality and mother tongue.

Chapter 5 provides a brief description of all the research done within the range of this dissertation as well as the main contributions. In this chapter we provide brief information related to research problems presented in the thesis. We describe the significance of the research method for creating ASR models adapted for non-native speech, as well as the method of real-time speech conversion, in the context of increasing the ASR accuracy. It also highlights the issues that are still yet to be solved and discusses directions for future research on the topic.

Łódź, 14.03.2022

Prof. dr hab. inż. **Krzysztof Ślot**
Instytut Informatyki Stosowanej
Politechnika Łódzka

Recenzja rozprawy doktorskiej

Pana magistra inż.
Kacpra Radzikowskiego
pt.

A study on speech recognition and correction for non-native English speakers

1. Tematyka i cele rozprawy

Przedstawiona do recenzji rozprawa doktorska dotyczy problematyki automatycznego rozpoznawania mowy (ang. *Automatic Speech Recognition*, ASR), przy czym szczególnym wątkiem badawczym tej niezwykle obszernej dziedziny, którego dotyczą prace badawcze Autora, jest rozpoznawanie mowy wypowiadanej przez osoby, dla których język wypowiedzi nie jest językiem ojczystym. Zagadnienie podjęte przez Doktoranta jest zarówno **istotne** jak i niezwykle **aktualne**. Postępująca globalizacja i związane z tym procesy społeczne, kulturowe i ekonomiczne wymagają umiejętności sprawnego komunikowania się z osobami z innych kręgów językowych, co w ostatnich dziesięcioleciach doprowadziło do powszechnej akceptacji jako narzędzi komunikacji jedynie kilku wybranych języków (przede wszystkim, języka angielskiego). Ten stan rzeczy okazał się poważnym wyzwaniem dla metod automatycznego rozpoznawania mowy, które stanowią istotny element wspierający postęp techniczny i cywilizacyjny, a które okazują się być niewystarczająco skuteczne w konfrontacji z mową pozyskiwaną od mówców ‘nienatywnych’.

Celem badań Doktoranta była próba opracowania metod rozpoznawania mowy pozwalających na skuteczne poradzenie sobie z przedstawionym, trudnym problemem. Skala złożoności zadania, wynikająca z wielu czynników: ogromnej różnorodności sposobów artykulacji i konstruowania wypowiedzi, często znacznie odbiegających od istniejącego dla danego języka kanonu, naleciałości i nawyków fonetycznych ukształtowanych przez inny język ojczysty, ubóstwa leksykalnego i gramatycznego, jest na tyle duża, że jak dotąd nie zostały opracowane żadne skuteczne metody jego rozwiązania. Dlatego też zadanie, którego rozwiązania podjął się Doktorant jest nietrywialne, nakreślony cel pracy – ambitny, a uzyskanie znaczących efektów prac może stanowić osiągnięcie o randze spełniającej kryteria merytoryczne stawiane przed rozprawą doktorską i stanowić

zauważalny wkład do dyscypliny naukowej **informatyka techniczna i telekomunikacja**, w której bez wątpienia lokuje się obszar badań podjętych przez Doktoranta.

2. Struktura i tezy rozprawy

Metodyka postępowania, przyjęta przez Doktoranta dla osiągnięcia postawionych celów rozprawy nie budzi żadnych zastrzeżeń. We wstępnej części rozprawy, Doktorant kompetentnie podsumowuje stan wiedzy w zakresie istniejących propozycji metod poprawy skuteczności rozpoznawania mowy nienatywnej oraz wskazuje podstawowe przyczyny utrudniające właściwe działanie standardowych narzędzi rozpoznawania mowy. Z uwagi na oczywistą trudność precyzyjnej identyfikacji czynników decydujących o odmienności mowy wypowiadanej przez osoby nienatywne, Doktorant zwraca się w kierunku wykorzystania uczenia maszynowego, jako jedynej sensownej metodyki adekwatnej dla możliwości rozwiązania problemu. Jednocześnie, dokonuje trafnej diagnozy jednego z głównych powodów trudności wdrożenia konwencjonalnych metod uczenia maszynowego, jakim jest brak wystarczająco obszernego, etykietowanego materiału eksperymentalnego, niezbędnego dla budowy skutecznych modeli rozpoznawania. W konsekwencji, uwaga Doktoranta skupia się na jedynym realnym alternatywnym źródle informacji - bazach mowy pozbawionych etykiet, których zgromadzenie w odpowiednio dużych wolumenach nie nastręcza problemu, ale których skuteczne wykorzystanie wymaga opracowania zaawansowanych algorytmów analizy.

Prace nad poszukiwaniem metod rozpoznawania mowy dla języka nie będącego językiem ojczystym mówcy, zostały przez Doktoranta przeprowadzone w trzech różnych kierunkach, odpowiadających trzem zasadniczym celom naukowym, które wytyczył dla swoich badań, i które można uznać za tezy przedstawionej rozprawy. Pierwszy z nich dotyczy opracowania metody pozwalającej na budowę modeli algorytmicznych ukierunkowanych na analizę i rozpoznawanie wypowiedzi mówców nienatywnych. Drugi cel to próba sformułowania metodyki modyfikacji zarejestrowanej mowy, dokonywanej na poziomie fonetycznym i zmierzającej do dostosowania struktury oraz brzmienia wypowiedzi do kanonów rozważanego języka, pozwalająca na wykorzystanie do rozpoznawania algorytmów opracowanych dla mówców natywnych. Wreszcie, ostatni cel badawczy sformułowany w rozprawie to próba weryfikacji możliwości zbudowania algorytmu dokonującego korekty tekstu będącego efektem analizy wypowiedzi, dostosowującej go do wymogów składni i gramatyki języka docelowego, pozwalającej na dodatkowe zwiększenie poprawności rozpoznawania. Celem prac w każdym z wymienionych wątków badawczych było uzyskanie rozwiązań o możliwie najbardziej ogólnym charakterze, pozwalającym na realizacje algorytmów w odniesieniu do dowolnych języków wypowiedzi.

3. Merytoryczna ocena pracy

Koncepcje zaproponowane przez Doktoranta w ramach realizacji prac są, w odniesieniu do dwóch pierwszych wymienionych wątków badawczych, czyli budowy systemu ASR wyspecjalizowanego w analizie mowy nienatywnej i opracowania metody fonetycznej transformacji wypowiedzi, bardzo ciekawe, oryginalne i merytorycznie znaczące. Podejście zastosowane w odniesieniu do realizacji trzeciego postawionego celu – opracowania algorytmu korekcji wyników rozpoznawania modułu ASR, mają w mojej opinii, raczej szablonowy charakter i dlatego w dalszej części recenzji zostanie mu poświęcone znacznie mniej uwagi. Punktem wyjścia dla sformułowanych przez Doktoranta

propozycji jest bardzo dobra identyfikacja właściwych, zaawansowanych i stosunkowo jeszcze świeżych koncepcji uczenia maszynowego, stanowiących bazę dla przedstawionych przez Niego, bardzo pomysłowych metod. Na uznanie zasługuje intuicja badawcza, której efektem jest formułowanie adekwatnych kryteriów sterujących przebiegiem uczenia rozważanych algorytmów oraz rzetelność i rozległość procedur eksperymentalnej weryfikacji hipotez, obejmująca trafny wybór materiału doświadczalnego, a także stosowanie właściwych miar i kryteriów oceny uzyskiwanych wyników, decydujące o wiarygodności procedury testowania i istotności otrzymanych rezultatów. Przedstawiona w dalszej części ocena dokonań Doktoranta została podzielona na dwie części, odzwierciedlające układ treści przedstawianych w pracy.

3.1 Metoda budowy modeli ASR dostosowanych do analizy mowy nienatywnej

Podstawowym ograniczeniem dla możliwości utworzenia modelu ASR dla mowy wypowiadanej w języku innym niż język ojczysty mówcy jest brak dostępności odpowiednio obszernego, etykietowanego materiału eksperymentalnego. Zapewnienie odpowiedniej złożoności modelu rozpoznawania mowy, niezbędnego z punktu widzenia złożoności rozwiązywanego problemu, wymaga dostarczenia podczas procesu uczenia wystarczającej wiedzy o problemie, co stoi w sprzeczności z istniejącym ograniczeniem. Dlatego, jedynym sposobem osiągnięcia zakładanego celu jest opracowanie procedury korzystającej z nieetykietowanych baz danych, co jest zadaniem nietrywialnym, wymagającym kreatywności i doskonałej orientacji w zaawansowanych koncepcjach uczenia.

Zaproponowane przez Doktoranta rozwiązanie – metoda budowy modeli ASR dla analizy mowy nienatywnej, stanowi twórcze rozwinięcie koncepcji Podwójnego Uczenia Nadzorowanego (Double Supervised Learning - DSL), adaptowane w sposób umożliwiający osiągnięcie założonego celu z wykorzystaniem wyłącznie dwóch nieetykietowanych źródeł danych: pozbawionych transkrypcji nagrań mowy oraz pozbawionych akustycznej ilustracji tekstów. Przedstawiony przez Doktoranta, niezwykle pomysłowy sposób rozwiązania zadania to wykorzystanie dwóch komplementarnych modeli transformacji danych: tekstowych w akustyczne (który będę określać dalej jako *Text-to-Speech*: T2S) i akustycznych w tekstowe (*Speech-to-Text*: S2T), pozwalające na realizację procedury uczenia nadzorowanego w konwencji autoenkodowania, korzystającego z danych generowanych przez dwa dodatkowe modele: generacji tekstu (dokonywanej przez wstępnie wytrenowany model języka) i generacji wypowiedzi (dokonywanej przez wstępnie wytrenowany model akustyczny).

Obydwa modele generacyjne: językowy i akustyczny są budowane na bazie nienadzorowanych, łatwo dostępnych (a więc, obszernych) zbiorów danych, przy czym ich dodatkową funkcjonalnością jest możliwość oceny poprawności podawanych na ich wejścia próbek tekstu i mowy. Dla budowy modelu języka Doktorant wykorzystuje korpus współczesnego języka angielskiego (COCA), stanowiący dostatecznie obszerną bazę informacji niezbędnej do zapewnienia poprawnej reprezentacji wiedzy, zaś dla budowy modelu akustycznego, równie bogaty, nieetykietowany zbiór nagrań języka angielskiego wypowiadanego przez mówców polsko- lub japońsko-języcznych. Procedura uczenia systemu ASR dla mowy nienatywnej, który jest tożsamy z modułem S2T przedstawionej komplementarnej pary, obejmuje również uczenie modelu T2S i jest dokonywana w pętli obejmującej: naprzemienną generację próbek albo przez generator języka, albo przez generator

akustyczny, przetworzenie tych próbek przez obydwa uczone modele i dokonanie odpowiednich korekt na podstawie odpowiednio zdefiniowanych funkcji straty. W dwuetapowym przekształceniu danych dokonywane są dwie oceny poprawności ich realizacji – cząstkowa (w jakim stopniu wygenerowany tekst / wypowiedź odpowiada modelowi języka / modelowi akustycznemu) i końcowa (zgodność zrekonstruowanej próbki z oczekiwaną, czyli wygenerowaną), co skłania Autora do przyjęcia interpretacji procesu treningu jako metody uczenia ze wzmocnieniem.

Dopełnieniem pomysłowej koncepcji treningu modelu rozpoznawania mowy jest nie budzący zastrzeżeń wybór implementacji jej komponentów składowych, przy czym oprócz narzucających się dla realizacji zadań cząstkowych neuronowych architektur rekurencyjnych (podstawowej RNN i LSTM), Doktorant dodatkowo wprowadza metodę n-gramów jako możliwy, alternatywny względem sieci neuronowych, sposób implementacji modułu modelowania języka. Jednocześnie, ponieważ moduł T2S stanowi w zaproponowanym algorytmie komponent o charakterze pomocniczym, Doktorant buduje go w oparciu o sprawdzoną architekturę Wavenet.

Weryfikacja eksperymentalna zaproponowanej metody jest również przeprowadzona w sposób nie budzący żadnych zastrzeżeń. W celu umożliwienia oceny przydatności swojej koncepcji Doktorant określa metodę referencyjną klasyfikacji, którą jest rekurencyjna sieć neuronowa trenowana w schemacie nadzorowanym. Celem eksperymentów jest nie tylko określenie podstawowych parametrów służących ocenie zaproponowanego algorytmu klasyfikacji, ale również weryfikacja dodatkowego pomysłu poprawy skuteczności jego działania poprzez wprowadzenie dwóch wariantów wstępnego przygotowania (pre-treningu) modułów T2S i S2T. Pierwszy z nich, określony przez Doktoranta jako ‘miękki start’ to inicjalizacja parametrów algorytmu dokonywana na podstawie niewielkich (a więc, dostępnych) etykietowanych baz danych. W drugim scenariuszu, inicjalizacja jest dokonywana tylko dla modułu klasyfikatora (S2T), co Doktorant określa terminem startu ‘pół-miękkiego’.

Uzyskane przez Doktoranta wyniki potwierdzają nie tylko słuszność hipotezy o możliwości zbudowania systemu rozpoznawania mowy wypowiadanej w języku, który nie jest językiem ojczystym osób mówiących, ale pokazują, że osiągana poprawność klasyfikacji przewyższa wyniki możliwe do uzyskania dla metody referencyjnej. Ten wynik, w świetle zupełnie innych uwarunkowań dla budowy modelu ASR dla obydwu przypadków, gdzie proponowana przez Doktoranta metoda jest tworzona na bazie zbiorów nieetykietowanych, podczas gdy druga metoda wymaga takiego zbioru, ma w mojej opinii niezmiernie istotne znaczenie i stanowi dowód na możliwość skutecznego rozwiązania postawionego problemu.

Na dodatkowe uznanie zasługuje podjęta przez Doktoranta próba uzupełnienia opracowanej metody o dodatkową procedurę zmierzającą do uzyskania dalszej poprawy skuteczności rozpoznawania mowy, stanowiąca realizację trzeciego z zapowiedzianych celów badawczych rozprawy. Istotą zaproponowanej koncepcji jest wprowadzanie do wyników rozpoznawania otrzymanych za pomocą omówionego wcześniej algorytmu, korekt wynikających z wiedzy o różnicach między gramatykami i konwencjami języka wypowiedzi i natywnego języka mówcy, nabytej w drodze zastosowania procedur uczenia maszynowego. Jako metodę realizacji korekt Doktorant wykorzystuje dwustopniową strukturę transformera (koder-dekoder), trenowaną na zbiorze par tekstów zawierających wyrażenia niepoprawne i skorygowane. Efektem wprowadzenia zaproponowanej

transformacji jest uzyskanie dodatkowej poprawy rozpoznawania, czyli dalsza poprawa właściwości opracowanego przez Doktoranta modelu ASR.

Podsumowując ocenę dokonań Doktoranta w pierwszym wątku przedstawionych przez Niego prac, chcę stwierdzić, że są one zdecydowanie **znaczące** i w mojej opinii, stanowią **zauważalny wkład** do nauki w skali światowej.

3.2 Metoda korygowania fonetycznych błędów artykulacji wypowiedzi

O ile pierwszym obszarem prac Doktoranta było zbudowanie kompleksowego algorytmu automatycznego rozpoznawania mowy, o tyle drugi zasadniczy wątek prac przedstawiony w rozprawie dotyczy badań prowadzonych nad modyfikacją postaci danych wejściowych – mowy wypowiadanej przez osobę nienatywną, dopasowującej sposób artykulacji wypowiedzi do kanonu obowiązującego dla rozważanego języka. Zadanie, określone przez Doktoranta jako modyfikacja akcentu wypowiedzi, jest realizowane w dziedzinie spektrogramu, a istotą realizacji zadania jest opracowanie algorytmu dokonującego przekształcenia spektrogramu według ‘wyuczonego’ dla danej grupy etnicznej mówców schematu. Narzędziem realizacji zadania jest albo koncepcja Autoenkodera albo metoda ‘transferu stylu’, a podstawą dla treningu metod jest zbiór danych złożony z par nagrań, zawierających te same zdania, wypowiadane przez mówców natywnych i mówców z rozważanej, nienatywnej grupy etnicznej. Należy w tym miejscu zaznaczyć, że utworzenie takiego zbioru danych nie jest procesem kosztownym, a więc zgromadzenie odpowiedniej ilości materiału eksperymentalnego jest jak najbardziej realne.

Pierwsza z zaproponowanych strategii wprowadzania korekt ujawnia doskonałą intuicję Doktoranta, który dostrzega podobieństwo podjętego zadania z ‘generycznym’ dla koncepcji Autoenkodera problemem ‘odszumiania’ danych wejściowych. W rozważanym przez Doktoranta zagadnieniu, dane wejściowe to wypowiedzi zaburzone niewłaściwym sposobem akcentowania, a więc efektem redukcji lub eliminacji ‘szumu’ powinna być generacja wypowiedzi z poprawną artykulacją (próbka referencyjna mowy wypowiadanej przez natywnego mówcę), usuwającą zniekształcenie obecne w próbce wejściowej (wypowiadanej przez osobę nienatywną).

Druga z rozważonych przez Autora metod jest bardziej złożona, a jej istotą jest próba modyfikacji sposobu artykulacji mowy, bazująca na założeniu możliwości dokonania separacji między komponentami ‘treści’ i ‘stylu’ wypowiedzi, wykorzystywanej w koncepcji tzw. transferu stylu. Podstawę dla możliwości dokonania takiej separacji stanowi założenie, że informacja o treści danych jest wydzielana na etapie filtracji dokonywanej w początkowych warstwach konwolucyjnych sieci. Dlatego też, wyniki oceny przetwarzania danych przedstawiających tą samą treść ale w różniącej się formie, uzyskiwane na wyjściach tych warstw powinny być identyczne. Ta obserwacja jest wyrażona ilościowo jako pierwszy składnik funkcji straty stosowanej do treningu Autoenkodera (równanie 10). Zabiegiem stosowanym dla identyfikacji komponentu stylu zawartego w danych jest przyjęcie, że miarą indywidualnego sposobu prezentacji treści jest statystyczna (pozbawiona informacji o przestrzennym rozkładzie) różnica wyników przetwarzania danych wejściowych przez dalsze warstwy konwolucyjne (równanie 11). Bazując na założeniu możliwości separacji stylu i treści, Doktorant proponuje procedurę, która korzystając z par wypowiedzi: referencyjnej (osoby natywnej) i badanej (osoby nienatywnej) dokonuje ekstrakcji odpowiednich

komponentów, a następnie przekształca, korzystając z modułu Autoenkodera, spektrogram wypowiedzi osoby nienatywnej w sposób redukujący obydwa komponenty funkcji straty (a więc, dąży do zachowania tej samej treści w wypowiedziach referencyjnej i analizowanej oraz upodabnia styl obydwu wypowiedzi).

Wypowiedzi skorygowane z użyciem zaproponowanej przez Doktoranta procedury są następnie przedmiotem rozpoznawania w systemie ASR, wytrenowanym w sposób ‘konwencjonalny’, a więc na podstawie danych zawierających wypowiedzi natywnych mówców języka. Doktorant rozważył w swoich pracach dwa warianty procedury rozpoznawania: pierwszy bazujący na analizie zmodyfikowanego spektrogramu i wykorzystujący odpowiednio zaprojektowany klasyfikator konwolucyjny, oraz drugi, bazujący na przekształceniu zmodyfikowanego spektrogramu do postaci czasowej i użycie klasyfikatora rekurencyjnego. O ile pierwszy ze wspomnianych sposobów został w pracy opisany w sposób jasny, o tyle takiej jasności nie mam w odniesieniu do drugiego z nich – nie potrafię powiedzieć, jak Doktorant przekształcił spektrogram w sygnał mowy (odwrotna transformacja Fouriera wymaga wiedzy o fazach prążków widma, a te nie są przekształcane).

Podsumowanie eksperymentów ewaluacji zaproponowanych metod korekcji stylu wypowiedzi, dokonane z użyciem właściwych do tego celu miar ilościowych wskazuje na bardzo znaczącą poprawę dokładności rozpoznawania, uzyskiwaną dzięki zastosowaniu obydwu metod. Jednocześnie otrzymane wyniki jednoznacznie identyfikują transfer stylu, jako strategię modyfikacji sposobu artykulacji wypowiedzi dającą lepsze efekty. Bardzo interesującą właściwością metody transferu stylu, na którą zwraca uwagę Doktorant, jest możliwość konfrontacji w trakcie procedury treningowej próbek różnych wypowiedzi, co znacznie wzbogaca liczbę przykładów używanych do treningu. W konsekwencji może to oznaczać, że lepsze wyniki uzyskiwane dla metody transferu stylu są albo wynikiem zastosowania znacznie bardziej ogólnego podejścia do analizy nagrań (‘styl’ i ‘zawartość’ są pewnymi pośrednio analizowanymi komponentami), ale też, może wynikać ze wskazanego zwiększenia rozmiarów zbioru danych treningowych (co, w obliczu szczupłości baz możliwych do zgromadzenia dla większości problemów praktycznych, i tak świadczy na korzyść metody transferu stylu).

Podsumowując ocenę prac Doktoranta w obszarze dotyczącym modyfikacji stylu artykulacji, chcę jednoznacznie stwierdzić, że podobnie jak w odniesieniu do poprzedniego wątku, są one **oryginalne, wartościowe** i stanowią **interesujący** i w moim przekonaniu, **zauważalny wkład** do nauki.

4. Mankamenty tekstu

Rozprawa jest napisana w sposób bardzo jasny, ma poprawny i logiczny układ i zawiera jedynie nieliczne usterki edycyjne lub fragmenty, które wymagają dodatkowego doprecyzowania lub komentarza i które zostaną wskazane w dalszej części niniejszego rozdziału.

Zanim przejdę do ich prezentacji, chciałbym jednak zwrócić uwagę na jedyną dostrzeżoną przeze mnie słabość przedstawionego tekstu, która dotyczy prezentacji kontekstu realizowanych prac. Przedstawiony przegląd stanu wiedzy jest obszerny. Autor przyjmuje w nim konwencję zwięzłej prezentacji głównych treści wybranych przez siebe pozycji, przy czym kryterium wyboru jest

prawdopodobnie istotność, według oceny Autora, wpływu prezentowanych prac na rozwój dziedziny. Przyjęcie takiej konstrukcji jest dla Autora dość wygodne, bo dzięki temu Doktorant unika konieczności dokonania trudnej systematyzacji podejść i metod a więc wskazywania, które wątki stanowią rozwinięcia wcześniejszych pomysłów, a które są zupełnie nowymi propozycjami. Mankamentem przyjętego podejścia jest oczywiście subiektywność wyboru – oprócz pozycji o niepodważalnie fundamentalnym charakterze, Doktorant w przytoczonych opisach omawia sporo prac o znaczeniu, w mojej opinii, niewielkim (co zresztą też jest poglądem subiektywnym). Jednak moje główne zastrzeżenie w odniesieniu do przeglądu stanu wiedzy w obszarze automatycznego rozpoznawania mowy dotyczy pominięcia rozwiązań zaproponowanych w ostatnich latach. Przedstawione przez Autora pozycje są datowane najpóźniej na 2014 rok, co uwzględniając długość cyklu wydawniczego oznacza, że Autor praktycznie wcale nie odnosi się do stanu dziedziny jaki ma miejsce w chwili obecnej. Postępy w tym obszarze, które za sprawą pojawienia się głębokich sieci neuronowych dokonały się w ostatnich ośmiu latach, nie objętych przeglądem są kolosalne (tak jak i w innych obszarach analizy danych). Chcę jednocześnie podkreślić, że wspomniana uwaga dotyczy wyłącznie edytorskiego mankamentu przedstawionego tekstu, a nie kompetencji Doktoranta, który w swoich pracach korzysta z najbardziej aktualnych i zaawansowanych koncepcji istniejących w obszarze tworzenia algorytmów ASR.

Dostrzeżone przez mnie w tekście, nieliczne fragmenty i zdania, których zrozumienie jest trudne lub niemożliwe są następujące:

Str. 29: The skilled speakers with advances semantics and syntax, and phonology related to their language, despite the diversity that is always present in native speech. - nie wiem, co Autor chciał tu powiedzieć.

str. 30: „As the vectors of phonemes existing for every new language that is new, may be listed based on current linguistic information or the extension of the corpora with its phonemic representation, not existing in languages defined before inside the system, could be defined. „ - niestety, tego zdania też nie rozumiem – co jest podmiotem, który ‘mógłby być zdefiniowany’?

Str 69: „As the loss network model for automatic speech recognition tasks, properties of convolutional and recurrent layers were combined, where the former layers become, in fact, employed as feature extractors”. Zdanie jest trudne do zrozumienia: wynika z niego, że modelem są właściwości ..

Prezentacja wyników metody uczenia modelu ASR na rys. 5 jest w mojej opinii niefortunna – wykresy ilustrujące przebieg procesu uczenia modelu powinny być przedstawione dla przypadku, który daje najlepsze efekty (te pokazane prawdopodobnie odpowiadają pierwszemu lub trzeciemu scenariuszowi treningu - są gorsze niż wyniki uzyskiwane dla sieci referencyjnej).

Przedstawione na rys.9 odniesienia do fragmentów pracy dotyczących komponentów przedstawianej metody są niewłaściwe, przy czym wydaje się, że błąd polega między innymi na użyciu niewłaściwej pierwszej cyfry identyfikowanych podrozdziałów (nie 2 tylko 4).

5. Wniosek końcowy

Podsumowując przedstawioną recenzję, chcę jednoznacznie stwierdzić, że Doktorant osiągnął wszystkie postawione przed sobą, ambitne cele badawcze i opracował funkcjonalne metody

pozwalające na skuteczną realizację zadania rozpoznawania mowy wypowiadanej przez osoby nie będące natywnymi mówcami danego języka. Opracowane przez Doktoranta algorytmy są oryginalne, pomysłowe i stanowią cenny wkład naukowy do dziedziny automatycznego rozpoznawania mowy, o istotnym znaczeniu w skali światowej i, jestem przekonany, dużym potencjale oddziaływania.

Konkludując recenzję, chciałbym stwierdzić, że przedłożona do recenzji rozprawa doktorska Pana magistra inżyniera Kacpra Radzikowskiego pt. „A study on speech recognition and correction for non-native English speakers” **spełnia** w moim przekonaniu z dużym nadmiarem wymagania określone w odnośnej ustawie o stopniach i tytule naukowym i tym samym **wnioskuję o dopuszczenie Autora rozprawy do publicznej obrony**.

Dodatkowo, uwzględniając bardzo wysoki poziom merytoryczny zaproponowanych przez Doktoranta rozwiązań chciałbym zwrócić się z wnioskiem o wyróżnienie rozprawy.

Signed by /
Podpisano przez:

Krzysztof Maciej
Ślot
Politechnika Łódzka

Date / Data: 2022-
03-14 11:57

Prof. dr hab. inż. Bożena Kostek, czł. koresp. PAN
Politechnika Gdańska,
Wydział Elektroniki, Telekomunikacji i Informatyki
Lab. Akustyki Fonicznej

28.02.2022 r.

Opinia nt. rozprawy doktorskiej mgra inż. Kacpra Radzikowskiego

pt.: „**A study on Speech Recognition and Correction for Non-native English Speakers**", wykonanej pod kierunkiem dra hab. inż. Roberta Nowaka, prof. PW oraz prof. Osamu Yoshie.

1. **Jakie zagadnienie naukowe/badawcze jest rozpatrywane w pracy (cel i teza rozprawy) i czy zostało ono dostatecznie sformułowane przez autora**

Przedmiotem recenzji jest rozprawa doktorska mgra inż. Kacpra Radzikowskiego pt.: „**A study on Speech Recognition and Correction for Non-native English Speakers".** Recenzowana rozprawa doktorska ma charakter eksperymentalny, składa się z pięciu zasadniczych rozdziałów: wprowadzenia, przeglądu prac w kontekście automatycznego rozpoznawania mowy (ARM, ang. Automatic Speech Recognition, ASR – tym akronimem będę się dalej posługiwać) oraz adaptacji systemów w przypadku osób, dla których język nie jest natywny (tzw. L2 *speakers*), metodologię wykorzystującą podwójne nadzorowane uczenie się, eksperymentów prowadzonych w kontekście modyfikacji akcentu w czasie rzeczywistym dla nienatywnych próbek mowy (tzw. transfer stylu). W pracy zawarto również wnioski, bibliografię oraz dwa dodatki (ogólnie dostępne zbiory próbek mowy oraz lista publikacji autora i współautorów). W pracy znajdują się również streszczenia w j. angielskim i polskim. Brakuje natomiast wykazu wykaz oznaczeń i wielkości matematycznych. Rozprawa obejmuje 92 stron tekstu wraz z dodatkami.

W Ustawie w odniesieniu do rozpraw doktorskich jest sformułowanie: „Rozprawę doktorską może stanowić praca pisemna, w tym monografia naukowa, zbiór opublikowanych i powiązanych tematycznie artykułów naukowych, praca projektowa, konstrukcyjna, technologiczna, wdrożeniowa lub artystyczna, a także samodzielna i wyodrębniona część pracy zbiorowej (art. 187.1.3)". Zgodnie z przepisami wymaga się złożenia oświadczeń: „W przypadku gdy rozprawę doktorską stanowi samodzielna i wyodrębniona część pracy zbiorowej, kandydat przedkłada oświadczenia wszystkich jej współautorów określające indywidualny wkład każdego z nich w jej powstanie. W przypadku gdy praca zbiorowa ma więcej niż pięciu współautorów, kandydat przedkłada oświadczenie określające jego indywidualny wkład w powstanie tej pracy oraz oświadczenia co najmniej czterech pozostałych współautorów."

We Wprowadzeniu, doktorant na str. 18 podaje, że w rozdziale 3 w części wykorzystano materiał opublikowany w artykule: K. Radzikowski, L. Wang, O. Yoshie, R. Nowak, *Dual supervised learning for non-native speech recognition*, EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:3, https://doi.org/10.1186/s13636-018-0146-4. Rozdział 4 obejmuje badania zawarte całościowo w artykule: K. Radzikowski, L. Wang, O. Yoshie, R. Nowak, *Accent modification for speech recognition of non-native speakers using neural style transfer*, EURASIP Journal on Audio, Speech, and Music Processing (2021) 2021:11 https://doi.org/10.1186/s13636-021-00199-3. Końcowe zdanie tego akapitu ze str. 18 zawiera informację, że niektóre z modeli, algorytmów i narzędzi zostały przygotowane przez doktoranta i zostały zaprezentowane w sześciu innych współautorskich pracach. W tym przypadku brakuje jednak oświadczeń współautorów, a – co bardziej istotne – uszczegółowienia, jakie są to modele, algorytmy i narzędzia. W rozdziale 1.3 można przywołać te informacje (ostatni akapit tego podrozdziału).

We Wprowadzeniu K. Radzikowski podaje również przedmiot badań dotyczący automatycznego rozpoznawania mowy, motywację stanowiącą genezę rozprawy w kontekście rozpoznawania mowy nienatywnej, cel i zawartość pracy.

Podany cel rozprawy uwzględnia dwa aspekty rozpoznawania mowy nienatywnej i odnosi się do uzyskania większej wynikowej skuteczności systemu ASR poprzez implementację dodatkowych algorytmów, a w szczególności:

- stworzenia skalowanej metodologii do wykorzystania systemu ASR;

- przygotowania algorytmu adaptacyjnego w zastosowaniu do próbek mowy w celu zwiększenia skuteczności;

- stworzenia dodatkowego algorytmu reprezentującego model języka nienatywnego (w stosunku do j. angielskiego).

Drugi cel stanowi odniesienie do skalowalności tworzonych systemów ASR mowy nienatywnej, tj., zaproponowania metodologii, która pozwalałaby na jej zastosowanie w przypadku innych języków i innego zestawu nienatywnych mówców.

Brakuje jednak sformułowania typowej tezy (tez) dla rozprawy doktorskiej we Wprowadzeniu, w której cel byłby określony w postaci hipotezy badawczej (hipotez) do udowodnienia w stosunku do stanu wiedzy (odniesienie jakościowe), z podaniem wskaźników (ilościowe), itd. **Hipoteza badawcza pojawia się dopiero w rozdziale 3.1.**

2. **Czy w rozprawie przeprowadzono w sposób właściwy analizę źródeł, w tym, literatury światowej, stanu wiedzy i zastosowań w przemyśle**

Bibliografia zawarta w rozprawie jest dość skąpa, liczy 68 pozycji, w tym 8 współautorskich publikacji autora rozprawy. Spis źródeł został podany w kolejności

cytowania, a nie alfabetycznej, co nie ułatwia analizy odniesienia do stanu wiedzy zawartej w tych publikacjach. Należy jednak zauważyć, że w rozdziale 1 i 2, pojawiają się odniesienia do źródeł, które nie zostały zamieszczone w Bibliografii. Niepokojące jest odwołanie do roku 2015 i wcześniejszych lat w przypadku konferencji INTERSPEECH, najbardziej prestiżowego forum prezentującego aktualne doniesienia nt. stanu wiedzy w obszarze systemów ASR, zaś w przypadku konf. ICASSP – do roku 2018. W tym kontekście aktualne są jedynie publikacje własne autora. W związku z powyższym, można przyjąć, że przywołane źródła nie odnoszą się w pełni do stanu wiedzy, co w jakimś stopniu warunkuje postawiony cel. Uzasadnienie dla powyższego wniosku można też zauważyć, analizując rys. 1 (str. 27), przedstawiający wielkość baz mowy w skali czasu wykorzystywanych w systemach ASR (ostatnie naniesione chronologicznie daty dotyczą roku 2015). Ponadto podany przez autora rozprawy zakres skuteczności w przypadku systemów ASR (mowa nienatywna, akcent), tj. 50-60% nie odpowiada stanowi wiedzy w tym zakresie.

Brakuje również odniesienia do zastosowań w przemyśle, należałoby przywołać systemy, tzw. *Language Support in Voice Assistants*, jak np. SIRI (Apple), ALEXA (Amazon), Google assistant, rozwiązania firm Microsoft, Alibaba, NVIDIA, itp. Odniesienia do technologii pojawią się we Wprowadzeniu, ale mają one charakter rysu historycznego.

3. **Czy autor rozwiązał postawione zagadnienia, czy użył właściwej do tego metody i czy przyjęte założenia są uzasadnione**

W przyjętej hipotezie badawczej (rozdział 3.1) doktorant odnosi się stworzenia metody, która wykorzystuje zbiory danych bez adnotacji, tj. próbki sygnału mowy nienatywnej bez adnotacji tekstowej oraz korpusu mowy bez odniesienia do sygnału mowy. Przyjęta metodologia wykorzystuje w części model DSL (Dual Supervised Learning), tj. podwójne (jednoczesne) nadzorowane uczenie zaproponowane pierwotnie przez Xia i współautorów w zastosowaniu do jednoczesnego treningu modelu w przypadku dwóch kategorii (tłumaczenie tekstu dwóch korpusów językowych w różnych konfiguracjach języków, rozpoznawanie obrazów, rozpoznawanie emocji, generowanie zdań nacechowanych emocjami). Zadaniem pierwszego modelu jest wygenerowanie zdania w formie tekstu oraz estymacja prawdopodobieństwa, z jakim wygenerowane zdanie jest naturalne w stosunku do zbudowanego modelu języka. Drugi model ma za zadanie wygenerować wypowiedź syntetyczną, która powinna odpowiadać przyjętemu modelowi. Na tym etapie autor nie wykorzystuje jeszcze metody DSL. Następnie autor rozprawy wprowadza dwa kolejne modele, jeden odpowiedzialny za rozpoznawanie fonemów z danej wypowiedzi, drugi odpowiada za syntezę mowy na podstawie zapisu tekstowego. Sądzę, że dla czytelności przedstawienia tej metodologii należałoby podać schemat blokowy zaproponowanej całościowej metody, a nie tylko przedstawić ją w postaci

pętli sprzężenia zwrotnego dla modeli języka i syntezy i opisu tekstowego, zwłaszcza, że w tej kolejnej fazie wprowadzony zostaje enkoder-dekoder wykorzystujący sieci neuronowe rekurencyjne w architekturze głębokiej (RNN). W metodologii w fazie testowania przyjęto dodatkowe metody/modele, jak np. RNN z długoterminową pamięcią krótkoterminową LSTM (Long Short-Term Memory) czy model N-gram (w postaci modelu 3-gram). W rozdziale 3.3.1 należałoby podać jednak dokładniejsze uzasadnienie dla przyjętych metod testowania, w szczególności dotyczy to metody RNN+LSTM; takie uzasadnienie można znaleźć dopiero na stronie 49 i na kolejnych stronach.

Ze względu na schemat działania metody, zasadne było przyjęcie funkcji straty, metoda treningu sekwencji, gdy nie ma możliwości dopasowania ramek (aligment) – *Connectionist Temporal Classification* (CTC) *loss function*. Ponadto przyjęty został błąd odpowiadający różnicy jednej litery (znaku?). Skuteczność przedstawiona dla różnych konfiguracji językowych mieści się w granicach 81,04% do 87,24%. Ważna jest tabela 6, która pokazuje czas potrzebny do wytrenowania poszczególnych modeli. Kolejne etapy eksperymentu dotyczące modelowania lingwistyczne, chociaż przyniosły dobrą skuteczność są mniej czytelne – ponownie – przydałoby się podać schemat blokowy eksperymentów dla badanych wątków. Dokładniejsze wyjaśnienie tego podejścia znajduje się w podsumowaniu tego rozdziału.

Tabela 7 zawiera wyniki związane z treningiem modelu językowego, przyjęto miarę CER (czy na pewno powinno pojawić się w tytule Tab. 7 słowo „training"?). Uwaga: definicja miary CER pojawia się we wzorze 13 w rozdziale 4. Wyniki 10-krotnej walidacji krzyżowej dla przyjętej miary są wysokie, ale brak jest porównania z literaturą.

Rozdział 4 obejmuje dwie metody modyfikacji akcentu wypowiedzi osoby nienatywnej w celu poprawy dokładności systemu ASR. Pierwsza metoda: trenowany autoenkoder z wypowiedzią osoby nienatywnej na wejściu i zdaniem (ten sam tekst) rodowitego mówcy na wyjściu. Metoda ta powinna działać poprawnie, jeśli byłaby trenowana na danych podawanych równolegle – tj. ta sama osoba (lub zdanie syntetyczne (text-to-speech, TTS) mówi to samo zdanie w dwóch akcentach. Czy nie było tego typu problemu w tym podejściu? W opisie tej metody brakuje mi właśnie odniesienia, w jaki sposób autoenkoder radzi sobie z problemem, gdy wypowiedź na wejściu ma inną długość niż wypowiedź na wyjściu

Drugi wątek badawczy w rozdziale 4. odnosi się do „transferu stylu". Założono, że „styl", akcent osoby, dla której język jest językiem rodzimym, jest podawany na wejście sieci neuronowej (nazwanej *loss network*), również w tym przypadku działanie algorytmów obejmuje reprezentację 2D (a nie obraz), czyli spektrogram. Czy w zaprojektowanej metodzie znajduje się mechanizm, który rozróżnia akcent nienatywny od natywnego?

Dobór metod i narzędzi jest poprawny, natomiast słabsze jest uzasadnienie przyjętych schematów działania metod. W przyjętym układzie rozprawy wydaje się, że jest miejsce na rozszerzenie wyjaśnień w stosunku do materiału zawartego w publikacjach.

Może warto byłoby pokazać takie schematy blokowe w czasie prezentacji rozprawy doktorskiej. Byłoby też wskazane pokazanie ich w Odpowiedzi na recenzję, tj. podanie szczegółów struktury zaprojektowanego algorytmu oraz pełniejszej metodologii działania systemu.

Odniesienie do tego czy cel, czyli hipoteza badawcza została osiągnięta zawarta jest w wynikach rozdziału 4. Metoda transferu stylu pozwoliła na uzyskanie większej skuteczności.

Końcowe uwagi: interesujące są wyniki uzyskane w kontekście j. japońskiego i polskiego (przyjęta miara pokazuje podobne ilościowe wyniki). Należałoby jednak przeprowadzić dyskusję dotyczącą czy taki wynik był spodziewany czy nie (są to bardzo różniące się języki).

Czy to oznacza, że skonstruowany model jest w tym sensie skalowalny?

4. **Na czym polega oryginalność rozprawy, co stanowi samodzielny i oryginalny dorobek autora, jaka jest pozycja rozprawy w stosunku do stanu wiedzy i poziomu techniki reprezentowanych przez literaturę światową**

Należy zaznaczyć, że tematyka rozprawy doktorskiej p. K. Radzikowskiego jest aktualna i obecnie intensywnie rozwijana, stanowi więc przyczynek w tym temacie. Osiągnięciem autora rozprawy są przygotowane struktury algorytmów, jak również opracowanie metodologii, nawet jeśli wymaga ona dodatkowych wyjaśnień. Jest to niewątpliwie samodzielny i oryginalny wkład autora.

Nawet, jeśli tytuł rozdziału 4 (...on-the-fly...) wydaje się na obecnym etapie nadmiarowy, zwłaszcza w kontekście wniosku, który podał autor rozprawy w rozdziale 5 (str. 77), to widać, że przyjęte założenia stanowią na pewno słuszny kierunek działania– z jednej strony – w celu zwiększenia zasobów mowy nienatywnej L2 poprzez syntezę akcentu czy transferu stylu, zaś – z drugiej do stworzenia tego typu systemów rozpoznawania mowy nienatywnej.

5. **Czy autor wykazał umiejętność poprawnego i przekonywującego przedstawienia uzyskanych przez siebie wyników (zwięzłość, jasność, poprawność redakcyjna rozprawy)?**

Autor pracy przedstawił poprawnie punkt wyjścia badań, który stanowi genezę, uzasadnienie tematyki oraz motywację, nawet jeśli stan wiedzy nie w pełni odpowiada aktualnym badaniom w literaturze (i technologii). Autor przedstawia ciąg

logiczny przywołanych w pracy badań w sposób jasny, nawet, jeśli brakuje szczegółów dotyczących metodologii.

Rozprawa doktorska jest przygotowana poprawnie od strony redakcyjnej – co prawda – przyjęło się numerowanie rys., tabel i równań w obrębie poszczególnych rozdziałów. Ponieważ jednak praca nie jest duża objętościowo, to można przyjąć, że nie wpływa to na czytelność układu pracy. Ponadto w pracach technicznych nie używa się 1 os. l. pojedynczej, rozumiem jednak, że autor chciał w ten sposób podkreślić wkład własny w przypadku materiału stanowiącego wkład do publikacji współautorskich.

Edycja pracy jest staranna, tj. rysunki obrazujące wyniki analiz są czytelne (chociaż brakuje szczegółów), to samo dotyczy tabel, język rozprawy (j. angielski) jest poprawny, zdarzają się tylko drobne literówki czy nie zawsze właściwe zastosowane przedimki (określniki). W pracy można znaleźć jednak usterki redakcyjne, np. odwołanie do tabeli „poniżej" (str. 45, rozdz. 3.3.5 przy czym tabela, a właściwie tabele znajdują się na kolejnej stronie czy kolejne odwołanie do tabeli bez numeru – też „poniżej" – tab. 6, str. 46). Nie są to usterki istotne z punktu widzenia oceny merytorycznej pracy, ale niewątpliwie łatwiej tę pracę by się analizowało, gdyby autor uważniej stosował właściwe odniesienia do wszystkich elementów pracy. Ponadto autor rozprawy stosuje cytowania w postaci numerów w nawiasach kwadratowych, jak również w postaci nazwiska i roku w nawiasach okrągłych i wreszcie odnosi się do źródeł, nie zamieszczając tych źródeł w Bibliografii.

6. **Jaka jest przydatność rozprawy dla nauk inżynieryjno-technicznych?**

Jak wspomniałam wcześniej, tematyka systemów ASR jest ważna i istotna, w szczególności dotycząca wykorzystania automatycznej syntezy mowy w celu zwiększenia zasobów do uczenia w rzeczywistych systemach komunikacji człowiek-komputer. Wskazuje to na potrzebę tworzenia algorytmów (uczenie głębokie), które mogą tworzyć w sposób automatyczny (nienadzorowany) bazy zawierające wypowiedzi „naśladujące" wymowę nienatywną. Prowadzone badania wpisują się w dyscyplinę informatyka techniczna i telekomunikacja (wg nowej klasyfikacji).

Podsumowanie

W podsumowaniu chciałbym zauważyć, że przedłożona mi do recenzji rozprawa p. Kacpra Radzikowskiego wymaga w zasadzie uzupełnienia ze względu na nie w pełni satysfakcjonującą konstrukcję rozprawy doktorskiej (brak szczegółów eksperymentów), wskazane usterki redakcyjne oraz brak wszystkich oświadczeń współautorów publikacji. Ze względu jednak na liczne publikacje w wysoko punktowanych czasopismach wnoszę **o dopuszczenie rozprawy doktorskiej p. mgra inż. Kacpra Radzikowskiego do publicznej obrony.**

Dlatego w końcowym wniosku stwierdzam, że rozprawa doktorska **spełnia wymagania** stawiane rozprawom doktorskim w aktualnie obowiązującej Ustawie.