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."

KOC

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.