Kraków, 5 października 2024 r.

dr hab. inż. Konrad Kowalczyk, prof. uczelni
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
email: konrad.kowalczyk@agh.edu.pl

RECENZJA

Rozprawy doktorskiej mgr inż. Agaty Sage pt.:

„Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych dla celów komputerowego wspomagania diagnostyki logopedycznej z wykorzystaniem technik sztucznej inteligencji"

Promotor: dr hab. inż. Paweł Badura, prof. PŚ

Zagadnienie naukowe, główne cele i tezy pracy doktorskiej

Przedstawiona do recenzji praca doktorska mgr inż. Agaty Sage pt. „Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych dla celów komputerowego wspomagania diagnostyki logopedycznej z wykorzystaniem technik sztucznej inteligencji" dotyczy zastosowania komputerowych technik uczenia maszynowego w wykrywaniu wad wymowy u dzieci. W pracy skupiono się na istotnym z punktu widzenia rozwoju dziecka problemie niepoprawnej realizacji głosek dentalizowanych zwanym potocznie seplenieniem, który jest specjalnym rodzajem dyslalii. W szczególności, rozprawa doktorska skupia się na analizie zależności pomiędzy wybranymi parametrami wyznaczonymi z reprezentacji sygnału wizyjnego i akustycznego a wybranymi cechami artykulacyjnymi, w kontekście diagnostyki wad wymowy u dzieci przedszkolnych.

Głównym wskazanym celem rozprawy jest: „Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych z wykorzystaniem metod sztucznej inteligencji".

W rozprawie sformułowano i udowodniono tezę główną, postawiono też dwie tezy pomocnicze.

Teza główna: „Istnieją istotne statystycznie różnice w cechach sygnałów akustycznych i danych obrazowych prezentujących mowę dzieci z różnymi (normatywnymi i nienormatywnymi) cechami mowy."

Teza pomocnicza nr 1: Możliwa jest wiarygodna segmentacja wybranych artykulatorów w obrazach twarzy z wykorzystaniem metod sztucznej inteligencji.

Teza pomocnicza nr 2: Ekstrakcja i analiza cech obrazowych 2D i 3D oraz parametrów akustycznych pozwala na określenie różnic między grupami w wybranych cechach artykulacyjnych.

Zakres rozprawy

Recenzowana praca doktorska została napisana w języku polskim, składa się z 9 rozdziałów głównych, obejmujących w sumie 121 stron, bibliografii zawierającej 183 źródła, dwóch dodatków przedstawiających szczegółowo wyniki uzyskane w ramach pracy doktorskiej oraz streszczenia w języku polskim i angielskim. Praca doktorska zawiera imponującą listę 73 tabel, 35 rysunków oraz wyczerpujący spis skrótów i oznaczeń.

Rozdział 1 wprowadza czytelnika w tematykę zaburzeń wymowy u dzieci oraz podstaw diagnostyki logopedycznej. Następnie omówiona została artykulacja głosek dentalizowanych wraz z podziałem w zależności od miejsca i sposobu artykulacji, wskazane zostały również główne czynniki i formy nienormatywnej realizacji sybilantów. W dalszej części rozdziału dokonano krótkiego przeglądu literatury dotyczącego systemów rejestracji danych pomiarowych oraz technik komputerowych powiązanych z zakresem tematyki rozprawy.

Rozdział 2 przedstawia zakres pracy doktorskiej, w tym podaje cel główny, tezę główną oraz dwie tezy pomocnicze. Po krótce prezentuje też przyjętą metodykę oraz układ rozprawy.

Rozdział 3 opisuje kontekst badań przedstawionych w rozprawie, w tym krótkie omówienie projektu NCN Sonata Bis, w ramach którego prowadzone były prace badawcze, opis urządzenia pomiarowego - maski akustyczno-wizyjnej wyposażonej w dwie kamery i piętnaście mikrofonów - jak również omówienie protokołu rejestracji danych, protokołu badania logopedycznego oraz sposób tworzenia etykiet eksperckich dla detekcji i segmentacji danych wizyjnych, segmentacji danych akustycznych i opisów logopedycznych dotyczących cech artykulacyjnych wymowy dziecka.

Rozdział 4 przedstawia analizę danych wizyjnych, rozpoczynając od omówienia zebranej bazy danych składającej się z ponad 17 tysięcy obrazów pochodzących z nagrań wideo

76 dzieci oraz jej podziału na trzy główne podzbiory. Następnie szczegółowo omówiono przetwarzanie wstępne, sposoby augmentacji danych oraz zastosowaną metodę detekcji artykulatorów przy pomocy sieci YOLO, wskazując na wybór wariantu jak i hiperparametrów modelu. Najwięcej uwagi poświecono zastosowanej w pracy doktorskiej metodzie segmentacji artykulatorów (tj. warg, ust, zębów i języka) opartej o znaną sieć DeepLab v3+. W dwuetapowym treningu Doktorantka najpierw zastosowała uczenie słabo nadzorowane przy użyciu licznego zbioru z niedokładnymi obrysami, a następnie dokonała dostrojenia sieci przy pomocy niewielkiego zbioru z ręcznymi obrysami ekspertów. W końcowej części tego rozdziału opisano zbiór rozlicznych cech geometrycznych w przestrzeni dwu- i trzywymiarowej oraz cech związanych z teksturą, co w przypadku wykorzystania obrazów pochodzących z dwóch kamer, dało w sumie aż 396 wybranych cech wizyjnych.

Rozdział 5 traktuje o analizie sygnału akustycznego pochodzącego z jednego z mikrofonów, w tym o przetwarzaniu wstępnym oraz ekstrakcji cech akustycznych spektralnych, cepstralnych oraz wyznaczonych bezpośrednio z przebiegu czasowego sygnału.

Rozdział 6 przedstawia dwuetapową analizę statystyczną zależności pomiędzy wybranymi 472 parametrami wizyjnymi i akustycznymi, a cechami artykulacyjnymi dla 12 głosek dentalizowanych.

Rozdział 7 zawiera dokładny opis przeprowadzonych eksperymentów oraz prezentuje wyniki detekcji artykulatorów w obrazach twarzy, segmentacji artykulatorów oraz analizy statystycznej dla wybranych cech artykulacyjnych, z podziałem na szeregi: syczący, szumiący i ciszący. Szczegółowe wyniki tej analizy przedstawiono w tabelach zamieszczonych w dodatkach A i B.

Rozdział 8 przedstawia dyskusję na temat detekcji i segmentacji artykulatorów w obrazach twarzy oraz dyskusję na temat zaobserwowanych zależności pomiędzy cechami wyekstrahowanymi z obrazu i sygnału akustycznego a cechami artykulacyjnymi.

Rozdział 9 zawiera krótkie podsumowanie wskazujące na najważniejsze osiągnięcia rozprawy w odniesieniu do głównego celu badawczego, tezy głównej oraz tez pomocniczych.

OCENA MERYTORYCZNA PRACY

Analiza źródeł

Bibliografia zawiera 183 starannie wyselekcjonowane źródła, z uwzględnieniem istotnych pozycji książkowych, artykułów w czasopismach naukowych oraz materiałów konferencyjnych. Wybrane źródła pozwalają na poznanie kontekstu prowadzonych badań,

adekwatne do tematu rozprawy omówienie stanu wiedzy w zakresie wspierania komputerowej diagnostyki logopedycznej oraz detekcji i segmentacji obiektów w obrazach.

Strona edycyjna pracy

Praca doktorska jest świetnie przygotowana od strony redakcyjnej. Na szczególną pochwałę zasługuje strona edytorska pracy doktorskiej, która przygotowana została z niespotykaną wręcz starannością i dokładnością. Bardzo pozytywnie odbieram trud przygotowania licznych schematów blokowych prezentujących w sposób graficzny poszczególne elementy pracy i relacje między nimi. Na każdym etapie zapoznawania się z pracą odczuwalne jest, iż Doktorantka przygotowała rozprawę z myślą o czytelniku. Niewątpliwy trud włożony w tak staranne przygotowanie rozprawy doktorskiej pod względem językowym i edycyjnym zasługuje na najwyższe uznanie.

Metodyka badawcza

Metodyka badawcza stosowana przez Doktorantkę jest właściwa i wskazuje na znajomość zarówno problemów jak i samej metodyki prowadzenia badań naukowych w obszarze przetwarzania obrazów metodami głębokiego uczenia maszynowego oraz umiejętności przeprowadzenia analizy statystycznej danych. Przeprowadzone eksperymenty zostały zaplanowane i przeprowadzone zgodnie z dobrymi praktykami i standardami. Metodyka zbierania danych, ich wstępna obróbka i anotacja uzyskana w sposób automatyczny lub przy pomocy ekspertów jak również podział na zbiory treningowe i testowe należy uznać za prawidłowy. Sposoby treningu stosowanych modeli neuronowych, w szczególności modelu do segmentacji obiektów z wykorzystaniem uczenia słabo nadzorowanego, a następnie dostrajania w sposób nadzorowany, wskazują na dobry warsztat badawczy Doktorantki w stosowaniu metod głębokiego uczenia maszynowego. Wybór miar ewaluacyjnych dla poszczególnych elementów analizy obrazów oraz wybór przeprowadzonych testów statystycznych w kolejnych etapach analizy jednoznacznie wskazują na bogatą wiedzę i umiejętność stosowania właściwych technik analizy danych.

Oryginalność rozwiązania problemu naukowego i uzyskanych wyników badań

Oryginalne rozwiązania problemów badawczych przedstawione zostały w rozdziałach 4 - 6, a ich wyniki i główne wnioski z analizy zostały przedstawione w ramach dyskusji w rozdziale 8. Głównym autorskim osiągnięciem Doktorantki jest zastosowanie dwuetapowego treningu metody segmentacji artykulatorów (tj. warg, ust, zębów i języka) polegającego w pierwszym

etapie na treningu słabo nadzorowanym przy użyciu obrazów ze zgrubnymi obrysami interesujących obiektów, a w drugim etapie na dostrojeniu sieci w ramach treningu nadzorowanego przy użyciu niewielkiego zbioru danych anotowanego przez ekspertów. Proces segmentacji został poprzedzony przez detekcję artykulatorów znaną siecią YOLO. Autorskim wkładem Doktorantki jest niewątpliwie analiza zależności pomiędzy opisem logopedycznym odnoszącym się do cech artykulacyjnych a licznych wyselekcjonowanych przez Doktorantkę cech pochodzących z obrazów uzyskanych niezależnie przez dwie kamery oraz z sygnału mikrofonowego. Prawdopodobnie jest to jedna z nielicznych prac analizująca relacje pomiędzy aż tak duża liczbą cech ekstrahowanych z sygnałów wizyjnych i akustycznych (w sumie aż 472 parametry) w kontekście normatywnej i nienormatywnej wymowy w języku polskim. Rozbudowana wieloetapowa analiza statystyczna stanowi mocny punkt pracy doktorskiej, a konsekwencja doboru odpowiednich testów i ich sukcesywnego stosowania pozwoliła na uzyskanie tak licznych wyników i wniosków dotyczących badanych zależności.

Wyniki prezentowanych w rozprawie badań zostały opublikowane m.in. w dwóch artykułach naukowych wydanych przez wydawnictwo MDPI (czasopismo Sensors i Applied Sciences), których Doktorantka jest pierwszym autorem, oraz jako współautor w referatach zaprezentowanych na znaczących konferencjach międzynarodowych typu Interspeech.

Główne uwagi i pytania dotyczące recenzowanej rozprawy

Po przeczytaniu rozprawy doktorskiej, nasuwa się kilka następujących uwag i pytań:

1) Praca doktorska kończy się dyskusją na temat wyników analizy zależności pomiędzy parametrami wizyjno-akustycznymi a cechami artykulacyjnymi wskazanymi przez logopedów. W rozprawie bardzo brakuje opracowania końcowego klasyfikatora lub propozycji systemu, który umożliwiałby wykorzystanie wyników przedstawionej w pracy analizy w celu wsparcia diagnostyki logopedycznej wad wymowy. Czytelnik ma wrażenie, że praca niespodziewanie kończy się bez ostatecznego wyboru cech i propozycji modelu lub systemu opartego o najbardziej istotne z nich, który umożliwiałby w pełni wykorzystanie otrzymanych wyników.

2) W rozprawie przedstawiono szczegółową analizę statystyczną zależności pomiędzy wybranymi cechami wizyjno-akustycznymi (198 parametrów wizyjnych pochodzących z obrazów dwóch kamer oraz 76 parametrów akustycznych) a cechami artykulacyjnymi na podstawie zebranej bazy danych. Czy przedstawione zależności pomiędzy poszczególnymi parametrami a cechami artykulacyjnymi mogą być uznane za zależności generalne czy są one właściwe dla analizowanych danych? W konsekwencji, czy wnioski płynące z tej analizy są

właściwe jedynie dla zebranej w ramach projektu bazy danych czy są zasadne w ogólnej diagnostyce wymowy?

3) Zdecydowanie brakuje porównania przeprowadzonej analizy zależności oraz walidacji wniosków wynikających z tej analizy na innych zbiorach danych z mową nienormatywną. Analiza ta mogłaby dotyczyć jedynie podzbioru cech, które w rozprawia analizowane są rozłącznie. Czy Doktorantka przeprowadziła analizę dla wszystkich lub podzbioru opisanych w pracy doktorskiej parametrów wizyjnych i/lub akustycznych na innych zbiorach danych?

4) Stosowane w rozprawie cechy wizyjne ekstrahowane były niezależnie z obrazu pochodzącego z dwóch kamer. Analizowana była też zależność pomiędzy odpowiadającymi sobie cechami pochodzącymi z obu źródeł. Czy Doktorantka rozważała zastosowanie stereowizji lub prowadziła eksperymenty związane z ekstrakcją cech pochodzących z obu obrazów jednocześnie?

5) Ekstrakcja cech akustycznych dokonana została bezpośrednio z reprezentacji sygnału akustycznego w dziedzinie czasu, krótkoczasowej reprezentacji spektrum lub cepstrum, bez stosowania wspomnianej w głównym celu pracy „sztucznej inteligencji". Jest to podejście odmienne niż w przypadku ekstrakcji cech z sygnałów wizyjnych. Czy Doktorantka rozważała lub stosowała alternatywne reprezentacje sygnału akustycznego zawierającego mowę, na przykład reprezentację otrzymaną przy pomocy dużych modeli neuronowych pretrenowanych w sposób samo-nadzworowany (ang. self-supervised learning)?

6) Niektóre (trójwymiarowe) cechy wizyjne uzyskano z następujących po sobie ramek wideo. Czy w kontekście badania zaburzeń wymowy, stosownym byłoby uwzględnienie zmienności reprezentacji sygnału akustycznego w czasie?

7) Akwizycja danych dokonana była przy pomocy urządzenia pomiarowego zawierającego 2 kamery i 15 mikrofonów. Dlaczego w pracy doktorskiej wykorzystany został jedynie jeden sygnał mikrofonowy i czy Doktorantka analizowała, podobnie do materiału wizyjnego, zależności pomiędzy cechami pochodzącymi z różnych mikrofonów?

8) Posiadając wiedzę wynikającą z przeprowadzonej analizy statystycznej cech dla szeregów syczących, szumiących i ciszących: Jakie cechy i/lub reprezentacje zaproponowałaby Doktorantka do budowy klasyfikatora umożliwiającego w sposób generalny rozróżnienie cech artykulacyjnych lub wad wymowy? Czy byłyby to głównie cechy akustyczne określane w rozpowie jako szumowe oraz parametry związane z kształtem języka, jak to wynika z analizy, czy też należałoby stosować inną reprezentację? Jeśli tak to jaką?

Wnioski końcowe

Podsumowując stwierdzam, że w przedłożonej do recenzji rozprawie doktorskiej Pani mgr inż. Agaty Sage pt. „Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych dla celów komputerowego wspomagania diagnostyki logopedycznej z wykorzystaniem technik sztucznej inteligencji", Doktorantka wykazała się znajomością zagadnień związanych z analizą statystyczną i eksploracją danych oraz detekcją i segmentacją obiektów w obrazach metodami uczenia głębokiego. Opisany w rozprawie doktorskiej autorski wkład w dziedzinę naukową jak również dorobek publikacyjny oceniam jako znaczący. W mojej opinii, przedstawiona praca spełnia warunki stawiane rozprawom doktorskim w aktualnie obowiązującej ustawie o stopniach i tytule naukowym. W związku z tym wnioskuję do Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej o dopuszczenie Pani mgr inż. Agaty Sage do dalszych etapów przewodu doktorskiego.