Prof. dr hab. inż. Daoud Robert Iskander
Katedra Inżynierii Biomedycznej
Wydział Podstawowych Problemów Techniki
Politechnika Wrocławska
Wybrzeże St. Wyspiańskiego 27
50-370 Wrocław

# R E C E N Z J A
pracy doktorskiej mgr inż. Agaty Sage
p.t. „Opracowanie metodyki przetwarzania sygnałów akustycznych i danych obrazowych dla celów komputerowego wspomagania diagnostyki logopedycznej z wykorzystaniem technik sztucznej inteligencji"

Recenzja została opracowana na podstawie uchwały nr 41/2024 Rady Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej z dnia 11 lipca 2024 r. oraz listu Przewodniczącego Rady, prof. dr. hab. inż. Ewy Piętki.

## 1. Układ i treść rozprawy

Praca doktorska, napisana w języku polskim, posiada 195 stron i składa się z dziewięciu rozdziałów oraz streszczeniami w języku polskim i angielskim. We wprowadzeniu do pracy, doktorantka szczegółowo opisuje złożony proces diagnostyki logopedycznej i zapoznaje czytelnika ze sposobami artykulacji zgłosek dentalizowanych (tzw. sybilantów) oraz sposobu ich zapisu. To bardzo wartościowa część pracy pozwalająca czytelnikowi na swobodne „poruszanie" się po zagadnieniach omawianych w dalszej części pracy. Rozdział drugi zwięźle opisuje zakres pracy oraz przedstawia graficznie (rys. 2.1) metodykę pracy. W rozdziale trzecim, doktorantka opisuje urządzenia i protokoły badawcze użyte podczas badań logopedycznych. Rozdziały czwarty i piąty poświęcone są, odpowiednio, analizie danych obrazowych (video) oraz akustycznych (audio). Elementy analizy statystycznej danych pochodzących z sygnałów audio z danymi z zarejestrowanych sekwencji video są opisane w rozdziale szóstym. Kolejno, doktorantka przedstawia wyniki pracy w rozdziale siódmym, wraz ze stosowną analiza statystyczną. Dyskusja i zwięzłe podsumowanie pracy są umieszczone, odpowiednio, w rozdziałach ósmym i dziewiątym. Wstępna część pracy zawiera typowe dla prac naukowych spisy rysunków, tabel, skrótów i oznaczeń. Końcowa część pracy zawiera wyczerpującą bibliografię (183 pozycje) oraz dwa dodatki ze szczegółowymi wynikami analiz. Praca napisana jest w profesjonalnym środowisku edytorskim. Od strony redakcyjnej to jedna z najlepszych prac, jakie recenzowałem.

## 2. Uwagi merytoryczne

### *2.1. Rozdział pierwszy*

Jak wspomniałem powyżej, to bardzo pożyteczny rozdział wprowadzający czytelnika w dyscyplinę logopedii, pozwalający na lepsze zrozumienie stosowanych pojęć, szczególnie zapisów symboli IPA. W rozdziale doktorantka odnosi się głównie do prac naukowych prowadzonych przez polskich naukowców. Nasuwa się tu zatem pytanie, na ile specyfika danego języka jest ważna w kontekście automatyzacji wspomagania diagnostyki logopedycznej. Czy opisane w dalszych częściach pracy techniki przetwarzania audio i video mogą być łatwo dostosowane do specyfiki innego języka niż polski?

### *2.2. Rozdział drugi*

Nie mam uwag merytorycznych do tej części pracy. Doktorantka stawia interesującą tezę główną, w której zakłada statystycznie istotne różnice w cechach sygnałów audio oraz video pomiędzy normatywnymi i nienormatywnymi grupami badanych dzieci. Dodatkowo, doktorantka formułuje dwie tezy pomocnicze, dotyczące wiarogodnej segmentacji obrazów i skutecznej ekstrakcji i analizy cech artykulacyjnych. Postawione tezy pozwoliły doktorantce na trafne sformułowanie celu pracy. W dalszej części rozdziału, doktorantka syntetycznie przedstawiła szkic metodyki, pozwalający na lepsze zrozumienie synergii pomiędzy sygnałami audio i video wykorzystanej w dalszej części pracy.

### *2.3. Rozdział trzeci*

Jak wcześniej wspomniałem, rozdział ten jest poświęcony materiałom badawczym (w tym uczestniczącym w badaniach dzieciom), urządzeniom pomiarowym oraz protokołom akwizycji danych. Ważnym wątkiem pracy było utworzenie adekwatnej bazy danych dla języka polskiego, na co doktorantka zwraca szczególną uwagę aż do końca sowich prac badawczych wykonywanych w ramach doktoratu. Z pracy wynika, że w tej części pracy była zaangażowana grupa specjalistów logopedów (tu domyślam się z czysto logistycznych powodów przeprowadzenia badań na znacznej grupie osób). Niemniej, pod koniec rozdziału, doktorantka stwierdza, że korzystano z diagnoz tylko jednego eksperta. Pytanie więc jakie można zadać w tym miejscu jest: jakich kryteriów użyto do wybrania z pośród grupy współpracujących logopedów jednego eksperta? To ważne, gdyż na podstawie jej/jego wiedzy ustalono GT (*ground truth*)?

Z czystej ciekawości mam też pytanie dotyczące ewentualnego wpływu wieku dziecka lub jego płci na badania. Czy można dzieci w wieku pięciu lat łączyć z grupą dzieci w wieku lat ośmiu? Czy pięcioletni chłopcy mają tę samą biegłość w mówieniu co ich rówieśniczki? I jeśli nie, czy ten czynnik może być w jakiś sposób brany pod uwagę w kontekście interpretacji wyników pracy. Chętnie usłyszałbym odpowiedzi na te pytania podczas obrony.

*2.4. Rozdział czwarty*

W rozdziale czwartym doktorantka opisuje metody przetwarzania oraz analizy danych obrazowych (video). W pracy nie analizowano sygnału video, tylko losowe wybrane klatki z nagrań. W tytule rozdziału, doktorantka trafnie użyła frazy „analiza danych video". Zdaję sobie sprawę z niemożliwości eksperckiego opisania każdej z klatek nagrania, ale nasuwa się tu pytanie: w jaki sposób ta losowość może wpływać na wyniki uczenia sieci? W części dotyczącej detekcji artykulatorów przy użyciu sieci YOLO, doktorantka opisuje potrzebę skalowania wcześniej wyznaczanych ramek o rozmiarach 300×400 (Y×X) do rozmiaru mozaik (320×320). Rozumiem więc, że dla składowej pionowej następuje tu skalowanie (rozciąganie) ze współczynnikiem 1,066 podczas gdy składowa pozioma obrazu zostaje poddana skalowaniu (ściskaniu) ze współczynnikiem 0,8. Czy nie lepiej było dokonać kompresji tylko składowej X lub użyć jednakowej kompresji ramki (tj. 0,8 dla osi X oraz Y) a następnie losowo wpisywać ją do obszaru mozaiki? A ogólniej mówiąc: czy rodzaj interpolacji dwuliniowej może wpłynąć na wyniki segmentacji? I jeszcze jedno pytanie: dlaczego rotacja obrazów (tak do kilku stopni, powiązanych z ruchem głowy) nie była brana w tym miejscu pod uwagę w procesie augmentacji? W przeciwieństwie, dla sieci DeepLab 3+ (Tab. 4.4) rotacja do ±36° była rozważana.

Warto podkreślić, że jestem pod wrażeniem ilości cech, opisanych w Tabelach 4.5 – 4.12, jakie doktorantka brała pod uwagę w procesie ekstrakcji obiektów (w sumie prawie 200 niepowtarzających się cech wizualnych). To kolejny przykład z jak wielką starannością i sumiennością doktorantka podeszła do pracy.

*2.5. Rozdział piąty*

Analizę sygnału audio doktorantka ograniczyła do zbadania sekwencji trwania danej głoski dentalizowanej. Z równania (5.2) wynika, że długość użytego segmentu (ramki) wynosi 1470 próbek, co przy częstotliwości próbkowania równej 44,1 kHz wynosi ~33 ms. Nasuwają się tu pytania natury technicznej: ile przeciętnie trwa wypowiedzenie przez dziecka danej zgłoski dentalizowanej? Czy 33-milisekundowe-ramki można było wydłużyć i wciąż założyć (pseudo) stacjonarność sygnału? A ogólniej mówiąc: czy nie można byłoby zastosować testu (np. KPSS) celem wyznaczenia jak najdłuższej ramki, dla której hipoteza o stacjonarności sygnału nie byłaby odrzucona (i przy okazji zastosować nakładkowanie)?

W równaniu 5.7 (Tab. 5.1), jak ma się maksymalne opóźnienie o wartości 40 próbek dla częstotliwości 25 kHz do opisu w pomocy Matlab do funkcji `harmonicRatio` gdzie mowa jest o maksymalnej wartości 40 ms dla 25 Hz? Czy w pracy rozważano sygnały i częstotliwości 25 kHz?

Ilość cech wziętych pod uwagę w analizie sygnałów audio (Tabele 5.1 – 5.3) jest równie imponująca co ta brana pod uwagę w analizie danych video.

*2.6. Rozdział szósty*

Rozdział szósty doktorantka poświęca statystycznym metodom analizy danych, zarówno video jak i audio. Doktorantka wybrała „bezpieczny" wariant zestawów testów ze wskazaniem na metody nieparametryczne. Takie podejście (włącznie z dość „ortodoksyjną" korekcją Bonferroniego) nie budzi żadnych zastrzeżeń formalnych. Warto tu jednak podkreślić, że parametryczny test ANOVA, który z zasady jest dużo silniejszy od testu Kruskala-Wallisa, jest również testem odpornym, dla którego równość wariancji czy też odstępstwo od hipotezy normalności nie są przeszkodami do jego użycia. Całkowicie zgadzam się z doktorantka, ze w przypadku 10-cio krotnych różnic w wariancjach poszczególnych zmiennych losowych należy rozważyć podejścia nieparametryczne.

*2.7. Rozdział siódmy*

Tu doktorantka syntetycznie i klarownie zawarła wyniki swojej pracy. Nie mam merytorycznych uwag do tej części pracy.

*2.8. Rozdział ósmy*

Rozdział ósmy poświęcony jest dyskusji otrzymanych w pracy wyników. Tu też warto podkreślić, że dyskusja jest rzeczowa, zasadna i trafna, a bardzo ważnym jej elementem, na jaki zwraca uwagę doktorantka, są ograniczenia zaproponowanych metod analizy danych video oraz audio.

*2.9. Rozdział dziewiąty*

Pracę zamyka podsumowanie, w którym to doktorantka powraca do celu pracy wyartykułowanego w rozdziale drugim. Jeśli chodzi o plany na przyszłość i kierunki rozwoju proponowanych technik, doktorantka podkreśla złożoność i wieloaspektowość systemów wspomagania diagnostyki logopedycznej, zaznaczając, że wyniki pracy stanowią jedynie część prac naukowych, jakie należy jeszcze wykonać, aby stworzyć w pełni funkcjonalny i klinicznie użyteczny system diagnostyczny.

## 3. Uwagi końcowe

Rozprawa opisuje bardzo ważne i znaczące zagadnienia dotyczące tworzenia automatycznych systemów wspomagania diagnostyki logopedycznej opartych na analizie danych z sekwencji video oraz sygnałów audio, wraz ze wspomaganiem za pomocą metod uczenia maszynowego. Warte podkreślenia jest samodzielne opracowaniu eksperymentów, przeprowadzanie znacznej ilości badań eksperymentalnych i klinicznych, tworzenie bazy danych oraz wnikliwa oraz rygorystyczna analiza wyników przy użyciu zaawanasowanych narzędzi przetwarzania sygnałów.

Przedstawione powyżej w mojej recenzji uwagi merytoryczne nie są krytyczne i nie umniejszają wysokiej jakości naukowej materiału przedstawionego w rozprawie doktorskiej.

Podsumowując, z pełnym przekonaniem stwierdzam, że recenzowana przeze mnie praca spełnia, z wyraźnym nadmiarem, ustawowe wymagania stawiane pracom doktorskim (art. 187 ust 1 i 2 Ustawy z dnia 20 lipca 2018 r. *Prawo o szkolnictwie wyższym i nauce*) i wnoszę o dopuszczenie mgr. inż. Agaty Sage do dalszych etapów przewodu doktorskiego. Biorąc pod uwagę wysoką rangę opublikowanych artykułów oraz wysoką jakość samej rozprawy wnoszę o jej wyróżnienie.

Wrocław, 1 października 2024 r.