Łódź, 12 maja 2020 r.

prof. dr hab. inż. Michał Strzelecki
Instytut Elektroniki Politechniki Łódzkiej
ul. Wólczańska 211/215
90-924 Łódź

Recenzja rozprawy doktorskiej mgr. inż. Michała Kręcichwosta **„Analiza przestrzennych modeli akustycznych głosek dentalizowanych w diagnostyce sygmatyzmu"**,
promotor rozprawy: dr hab. inż. Paweł Badura, prof. uczelni,
promotor pomocniczy: dr inż. Joanna Czajkowska

Podstawą niniejszej recenzji jest pismo dziekana Wydziału Inżynierii Biomedycznej Politechniki Śląskiej, prof. dr. hab. inż. Marka Gzika z dnia 9.04.2020 r. informujące o powołaniu mnie przez Radę Dyscypliny Inżynieria Biomedyczna na recenzenta w postępowaniu doktorskim mgr. inż. Michała Kręcichwosta prowadzonym w dyscyplinie inżynieria biomedyczna.

W ciągu ostatnich lat logopedzi obserwują wzrost różnych zaburzeń w rozwoju mowy u dzieci (wg niektórych szacunków problem ten dotyczy aż 75% polskich dzieci w wieku szkolnym). Wady wymowy, nie korygowane w odpowiedni sposób, mogą z czasem przełożyć się na problemy komunikacyjne oraz w efekcie na trudności już dorosłych osób w prawidłowym funkcjonowaniu w społeczeństwie. Recenzowana rozprawa dotyczy budowy i analizy modeli akustycznych wybranej klasy głosek w celu wspomagania diagnostyki sygmatyzmu (seplenienia) – wady polegającej na nieprawidłowej wymowie jednego, dwu lub wszystkich trzech szeregów: s, z, c, dz; sz, ż, cz, dż i ś, ź, ć, dź. Uważam, że tematyka badawcza przedstawiona w rozprawie jest bardzo istotna i aktualna, szczególnie ze względu na możliwość jej zastosowania przez logopedów w swojej praktyce terapeutycznej. Opracowywanie komputerowych metod analizy sygnałów akustycznych do wspomagania diagnostyki i terapii wad wymowy jest niezwykle istotne, ponieważ metody takie są obiektywne i powtarzalne, ponadto istnieje bardzo niewiele rozwiązań w tym zakresie dedykowanych dla osób posługujących się językiem polskim.

Główny cel rozprawy został sformułowany jednoznacznie, dotyczy on opracowania przestrzennych modeli akustycznych głosek dentalizowanych i ich wykorzystania do klasyfikacji normatywnych oraz nienormatywnych realizacji głoskowych fonemów dentalizowanych. Również poprawnie określono cel utylitarny, związany z zaprojektowaniem i skonstruowaniem stanowiska pomiarowego pozwalającego na przestrzenną akwizycję sygnału mowy (choć wyjaśnienie znaczenia „przestrzenny" pojawia się dopiero podczas lektury dalszych części rozprawy).

W rozprawie sformułowano również ogólną tezę badawczą:

Wykorzystanie metodyki przestrzennego przetwarzania sygnału mowy umożliwia rozpoznawanie normatywnej oraz nienormatywnej głoskowej realizacji fonemów dentalizowanych.

oraz szczegółową tezą badawczą:

Skwantyfikowany opis przestrzennych modeli akustycznych głosek dentalizowanych umożliwia rozpoznawanie normatywnej oraz nienormatywnej głoskowej realizacji analizowanych fonemów.

Teza ogólna z nadmiarem spełnia cechę ogólności, podobnie równie ogólna jest teza szczegółowa. Nie odniesiono się do zastosowanego rodzaju skwantyfikowanego opisu analizowanych modeli akustycznych ani do innych metod wykorzystanych w rozprawie, które umożliwiają klasyfikację normatywnej i nienormatywnej realizacji badanych fonemów. Tak postawiona teza z reguły jest prawdziwa, ponieważ w zbiorze wszystkich możliwych ilościowych ujęć modeli analizowanej klasy głosek zawsze znajdzie się taki opis, który pozwoli na poprawne rozróżnienie wymawianych głosek. Moim zdaniem teza powinna odnosić się do konkretnych metod i narzędzi (m.in. głębokich sieci neuronowych) wykorzystywanych w rozprawie do rozwiązania przedstawionego problemu naukowego rozpoznawania i klasyfikacji wybranych głosek.

Rozprawa liczy 172 strony i została podzielona na 7 rozdziałów. Rozdział 1 stanowi wprowadzenie w tematykę rozprawy, opisuje cele badawcze oraz zawiera postawione tezy. Opisano tam również obecny stan wiedzy w zakresie istniejących systemów wspierania diagnozy i terapii zaburzeń mowy. Omówiono główne wady wymowy, w tym sygmatyzm oraz charakterystykę widmową i spektroskopową głosek dentalizowanych. Nieprawidłowa wymowa takich głosek jest głównie wynikiem seplenienia, a wykrywanie takich zaburzeń stanowi główny obszar zainteresowań badawczych Autora pracy. Bardzo ciekawy rozdział 2 opisuje projekt prototypu urządzenia do akwizycji danych – maski akustycznej - umożliwiającego wielokanałową, przestrzenną i powtarzalną rejestrację sygnału mowy. Z przeprowadzonego przeglądu literatury wynika, że nie istnieją takie gotowe rozwiązania, zwłaszcza które mogłyby być wykorzystywane w pracy z dziećmi. Opisano szczegółowo projekt urządzenia, prawidłowo dobierając poszczególne elementy elektroniczne (matrycę składającą się z 15 mikrofonów elektretowych, układy przedwmaczniacza i wzmacniacza), co świadczy o posiadaniu przez Doktoranta wysokich kompetencji w projektowaniu tego typu urządzeń. Bardzo starannie i szczegółowo przeprowadzone testy maski akustycznej wykazały, że spełnia ona wszystkie założone wymagania projektowe. Potwierdzono poprawne działanie przetworników akustycznych oraz dobrą powtarzalność odpowiedzi mikrofonów na bodźce dźwiękowe. Przeprowadzono również analizy statystyczne odpowiedzi sygnałów z poszczególnych mikrofonów, wykazując brak istotnych różnic pomiędzy nimi. Badania maski akustycznej oraz przeprowadzone testy użytkowe potwierdziły jej poprawne działanie i możliwość wykorzystania do akwizycji sygnałów akustycznych. Co bardzo istotne, podczas projektowania urządzenia zadbano o jego prawidłową ergonomię, tak aby było możliwe wykorzystanie maski do pracy z dziećmi. Rozdział trzeci opisuje bazę mowy, która zawiera użyty w rozprawie materiał badawczy. Do stworzenia tej bazy, we współpracy z logopedami, wykorzystano nagrania pochodzące od 6-7. letnich dzieci. Nagrania zawierały zarówno poprawnie (normatywnie) jak i nienormatywnie artykułowane fonemy /s/ oraz /ʃ/, wraz z wyróżnieniem 3 klas patologii: interdentalnej, addentalnej i dentalnej. Dla każdej z klas

zarejestrowano od 80 do 217 słów (w zależności od głoski, łącznie ponad 1300 nagrań) co jest wystarczającą liczbą dla uzyskania statystycznie wiarygodnych wyników badań. W tabelach 3.3 zabrakło jednak podania liczby mówców oraz zarejestrowanych słów dla addentalnej realizacji analizowanych głosek.

W rozdziale 4, najistotniejszym rozdziale rozprawy, przedstawiono metodę przetwarzania i analizy zarejestrowanego sygnału mowy w celu detekcji błędnie wymawianych głosek oraz określenia związanej z tym patologii. Przeanalizowano właściwości akustyczne skonstruowanej matrycy mikrofonów (ułożonych w postaci 5 trójelementowych zestawów), opracowano metodę ich agregacji i synchronizacji uzyskując 5 niezależnych sygnałów bez przesunięć fazowych przy zachowaniu odpowiedniej wartości SNR. Uzyskane sygnały poddano również procesowi filtracji i normalizacji uwzględniając charakterystyki częstotliwościowe analizowanych sygnałów, a następnie równomiernej segmentacji w celu uzyskania quasi stacjonarnych fragmentów (ramek) o długościach 10-50 ms. Następnie sygnał poddano filtracji pasmowoprzepustowej za pomocą zestawu 64 filtrów w zakresie 1-22 kHz. Dla każdego zakresu filtracji wyznaczono energię widma, których zbiór wykorzystano jako parametry opisujące poszczególną ramkę. Poza energiami banku filtrów (FBE) wyznaczono również ich pierwsze i drugie pochodne dla oceny dynamiki sygnałów audio. Skonstruowano w ten sposób zestawy 4-ro wymiarowych danych (uwzględniające informacje czasowo-częstotliwościowe, liczbę kanałów oraz liczbę rodzaju cech) opisujących poszczególne segmenty sygnału mowy. Tak wygenerowane dane stanowiły wejście konwolucyjnej sieci neuronowej mającej za zadanie rozpoznawanie fonemów i wykrywanie ich błędnych realizacji. Sieć ta zawiera typowe warstwy niezbędne m.in. do przeprowadzenia filtracji splotowej oraz klasyfikacji danych, posiada też elastyczną strukturę umożliwiającą wybór i łączenie różnych bloków przetwarzania danych oraz redukcję liczby występujących warstw. Zdefiniowano również rodzaje klas fonemów rozpoznawanych przez sieć oraz określono liczebności zbiorów uczących, testowych i treningowych. Sieci konwolucyjne wymagają dużej ilości danych uczących, niezbędnych dla zapewnienia zdolności sieci do generalizacji wyników uczenia. Z tego powodu zastosowano metody augmentacji zbioru danych (liczącego pierwotnie od 80 do 200 próbek, w zależności od klasy) stosując preemfazę oraz podział analizowanego segmentu na fazy artykulacyjne. Ta ostatnia metoda umożliwiła trzykrotne zwiększenie zbioru uczącego. Podsumowując, w rozdziale tym opisano metody umożliwiające przetworzenie analizowanych sygnałów akustycznych do postaci danych, które mogą stanowić wejście dla głębokiej sieci neuronowej przygotowanej do wykrycia cech tych sygnałów w dziedzinie czasu i częstotliwości, umożliwiających detekcję fonemów i określenie poprawności ich realizacji.

Wyniki badań przedstawione w rozdziale 5 były częściowym potwierdzeniem przyjętych wcześniej założeń badawczych. Wykazano skuteczność zastosowania wielokanałowego systemu akwizycji sygnału mowy, który umożliwił (dzięki analizie wartości skutecznej zarejestrowanego sygnału) rozróżnienie prawidłowej i nieprawidłowej realizacji badanych fonemów. Uzyskano tu pełną zgodność z opisem badanych fragmentów sygnału mowy wykonanych przez logopedów. Analiza jakościowa właściwości czasowo-częstotliwościowych takich fragmentów doprowadziła do wniosku, że wartości energii banku filtrów (oraz ich pochodnych) stanowią właściwy zestaw cech umożliwiający detekcję poprawności realizacji głosek /s/ oraz /ʃ/, zatem zastosowanie cech FBE jako danych wejściowych dla sieci konwolucyjnej jest wyborem poprawnym.

Rozdział szósty zawiera wyniki testów opracowanej metody analizy sygnału mowy, szczególności wyniki walidacji zastosowanej sieci neuronowej. Eksperymenty przeprowadzono w różnych wariantach. Badano zależność skuteczności rozpoznawania poszczególnych realizacji głosek od liczby kanałów (stosowane warianty obejmowały wersje wykorzystania 1, 5 oraz 15 kanałów). Badano także wpływ zakresu częstotliwości zespołu filtrów pasmowoprzepustowych oraz wpływ struktury sieci (liczby warstw, zakres filtracji splotowej danych wejściowych). Niezależnie, dla każdego z powyższych wariantów przeprowadzono pięć eksperymentów dotyczących klasyfikacji różnego zakresu danych (poprawne realizacje głosek /s/ oraz /ʃ/, poprawne realizacje głoski /s/ oraz dwóch jej realizacji nienormatywnych, analogiczny eksperyment dla głoski /ʃ/, połączenie dwóch ostatnich eksperymentów generujące w wyniku 6 klas oraz normatywnych i wszystkich wariantów nienormatywnych realizacji badanych głosek). Wyniki każdego z eksperymentów poddano krytycznej dyskusji, dokonując przekonujących prób wyjaśnienia uzyskanych wyników, w szczególności przyczyn błędów klasyfikacji. Na szczególne uznanie zasługuje druga część rozdziału, gdzie przeprowadzono analizę wrażliwości wyników klasyfikacji dla wybranych parametrów sieci. Takie podejście nieczęsto spotyka się badaniach dotyczących klasyfikacji danych, mimo że analiza wrażliwościowa umożliwia optymalizację klasyfikatora. W tych badaniach uwzględniono m.in. właściwości struktury sieci (rozmiar i liczba filtrów konwolucyjnych), rodzaje danych wejściowych, parametry sterujące procesem uczenia. Do zbadania statystycznej istotności różnic wartości uzyskanych parametrów mierzących jakość klasyfikacji zastosowano odpowiednio dobrane testy statystyczne. Dzięki przeprowadzonym eksperymentom udało się zidentyfikować optymalną strukturę sieci oraz sposób przetwarzania danych wejściowych maksymalizujący poprawne rozpoznawanie realizacji (poprawnej i niepoprawnej) badanych głosek.

Liczący 194 pozycji wykaz literatury obejmuje wszystkie najważniejsze pozycje literatury światowej dotyczące tematyki związanej z rozprawą. Wykaz ten zawiera również 3 publikacje współautorskie mgr. Kręcichwosta zamieszczone w materiałach konferencyjnych i w czasopiśmie Biocybernetics and Biomedical Engineering (IF = 2,159).

Podsumowując merytoryczną ocenę rozprawy stwierdzam, że Doktorant osiągnął zdefiniowany w pracy cel badawczy oraz udowodnił postawione tezy. Opracował i zweryfikował model akustyczny wybranych głosek dentalizowanych oraz wykazał jego skuteczność w klasyfikacji normatywnych i nienormatywnych realizacji fonemów dentalizowanych. Należy podkreślić, że osiągnięcie celu pracy wymagało wykazania się przez mgr. Michała Kręcichwosta szeroką wiedzą nie tylko z dyscypliny inżynieria biomedyczna i biocybernetyka (obecnie: inżynieria biomedyczna) w zakresie opracowania metody detekcji realizacji, wraz z oceną jej poprawności dla wybranych głosek, ale również z dyscypliny elektronika (obecnie: automatyka, elektronika i elektrotechnika), w zakresie znajomości metod projektowania i budowy układów elektronicznych do akwizycji, przetwarzania i analizy sygnału mowy. Do najważniejszych osiągnięć Autora pracy zaliczam:

1. wykonanie oraz wszechstronne przetestowanie urządzenia (maski akustycznej) umożliwiającego wielokanałową akwizycję sygnału mowy, które zostało zaprojektowane w taki sposób, aby mogło być wykorzystane do pracy z dziećmi w wieku 6-7 lat;
2. opracowanie, implementację oraz dokładną walidację metody rozpoznawania i oceny poprawności realizacji wybranych sybilantów, która może być wykorzystana we wspomaganiu diagnostyki i terapii sygmatyzmu;

Praca jest napisana poprawnym językiem, nie budzi również większych zastrzeżeń od strony redakcyjnej. Zauważyłem tylko pojedyncze usterki, np. „korekty Welch'a" -> „korekty Welcha" (str. 98); symbol * czasem oznacza mnożenie (str. 39), a czasem splot (str. 82); na rys. 4.6 w częściach a i b podano różne wartości liczb mikrofonów N (str. 60). Z kolei na rys. 6.8, 6.13, 6.18, 6.22, 6.31 i 6.32 wartości dokładności sieci podane w kolorowych prostokątach są mało czytelne.

Lektura pracy nasunęła mi również kilka przedstawionych poniżej uwag polemicznych i dyskusyjnych.

1. Uzyskane wartości dokładności klasyfikacji realizacji głosek /s/ oraz /ʃ/ wynoszą średnio 0.97, co jest wynikiem bardzo dobrym. Jednak przy klasyfikacji patologicznych realizacji tych głosek oraz eksperymentów mieszanych wartości są dużo mniejsze i zawierają się w przedziale (0.75-0.79). Czy taka dokładność metody jest wystarczająca z punktu widzenia zastosowania jej jako narzędzia do wspomagania diagnostyki sygmatyzmu? Jaka jest opinia logopedów dotycząca zastosowania opracowanej metody w praktyce terapeutycznej?

2. W rozdziale 6 wspomniano o wykonaniu 10 niezależnych prób eksperymentów klasyfikacyjnych. Czy należy przez to rozumieć, że przeprowadzono 10-krotną walidację krzyżową sieci konwolucyjnej? Czy przedstawione wartości parametrów oceniających działanie sieci są średnią ze zrealizowanych 10 eksperymentów?

3. Jako parametry analizowanych fragmentów mowy wykorzystano energie banków filtrów. Czym był spowodowany taki akurat wybór i czy nie warto byłoby przetestować innych cech szeroko wykorzystywanych w analizie widmowej sygnału mowy, np. parametrów mel-cepstralnych lub momentów widmowych?

4. Czy Doktorant rozważał wykorzystanie innych, poza siecią neuronową, klasyfikatorów? Naturalnie, w takim przypadku należałoby dokonać redukcji liczby cech opisujących analizowane fragmenty mowy. Potencjalną korzyścią z takiego podejścia byłoby określenie, które z nich są rzeczywiście istotne z punktu widzenia rozróżniania poszczególnych sybilantów (np. czy istnieją określone zakresy częstotliwości mające szczególne znaczenie przy odwzorowywaniu patologicznych realizacji fonemów).

5. W pracy zabrakło informacji dotyczących oprogramowania wykorzystywanego do implementacji modeli akustycznych głosek oraz sieci konwolucyjnej. Nie podano również czasu niezbędnego do nauczenia sieci oraz szybkości działania opracowanej metody przetwarzania i analizy zarejestrowanego sygnału mowy.

Wszystkie moje uwagi krytyczne i dyskusyjne w żadnym stopniu nie wpływają na jednoznacznie pozytywną ocenę recenzowanej pracy, która ma charakter interdyscyplinarny, co zasługuje na szczególne uznanie (Autor wykazał osiągnięcia w dyscyplinach inżynieria biomedyczna oraz automatyka, elektronika i elektrotechnika). Stwierdzam, że praca „Analiza przestrzennych modeli akustycznych głosek dentalizowanych w diagnostyce sygmatyzmu" z nadmiarem spełnia wymagania stawiane rozprawom doktorskim zgodnie z Ustawą o stopniach naukowych i tytule naukowym z dnia 14 marca 2003 r. z późniejszymi zmianami, oraz wnioskuję o przyjęcie tej rozprawy i dopuszczenie mgr. inż. Michała Kręcichwosta do publicznej obrony.