Łódź, 30.06.2020

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

Recenzja rozprawy doktorskiej
Mohammad K. Nammous
„New Approaches in Speech Recognition from Isolated Words to Practical Solutions"

1. Tematyka i charakter rozprawy

Zakres tematyczny przedłożonej do recenzji rozprawy doktorskiej obejmuje analizę sygnału mowy ukierunkowaną na realizację zadania biometrycznego rozpoznawania mówcy (ang. *speaker recognition*) i dokonywaną metodami uczenia maszynowego, a także zagadnienia pokrewne, wykorzystywane pomocniczo przez Autora: rozpoznawanie treści (ang. *speech recognition*), rozpoznawanie płci mówcy (ang. *gender recognition*) i rozpoznawanie języka wypowiedzi (ang. *language recognition*). Obszar prac badawczych rozprawy zawiera się więc bez wątpienia w dyscyplinie informatyka techniczna i telekomunikacja, w której prowadzony jest przewód. Ponieważ głównym przedmiotem rozprawy są zagadnienia rozpoznawania mówcy, sformułowanie tytułu rozprawy jest wysoce niefortunne, sugerując przedstawianie w niej treści przede wszystkim związanych z rozpoznawaniem mowy. W tytule nie tylko pojawia się explicite zwrot „*speech recognition*", ale również jego konstrukcja („*from isolated words to practical solutions*") utwierdza Czytelnika w przeświadczeniu, że tematyką rozprawy będzie prezentacja nowych podejść do rozpoznawania mowy, budowanych na fundamencie podstawowych metod rozpoznawania izolowanych słów. Oczywiście, zadanie rozpoznawania mówcy można uznać za szczególny obszar dziedziny rozpoznawania mowy – w obydwu przypadkach stosowane są pokrewne mechanizmy budowania opisu sygnału mowy i metod jego klasyfikacji, ale używanie takiej interpretacji jest mało przekonujące, pozostawiając poczucie rozdźwięku między tytułem a zawartością rozprawy.

Kierunek prac podjętych w rozprawie – analiza możliwości uzyskania wysokiego stopnia poprawności dokonywanych analiz w obliczu posiadania nielicznego zbiorów przykładów, jest bezsprzecznie istotny i aktualny. Przełomowy wzrost skuteczności metod uczenia maszynowego i sztucznej inteligencji, uzyskany w efekcie pojawienia się koncepcji uczenia głębokiego i głębokich sieci neuronowych, opiera się na umiejętnym wykorzystaniu bardzo dużych zbiorów danych, niezbędnych do właściwego wytrenowania algorytmów o liczbie parametrów sięgającej dziesiątek lub setek milionów. Ponieważ istnieje wiele zadań i kontekstów analizy danych, w których zgromadzenie odpowiednio obszernych zbiorów danych nie jest możliwe, lub też ich etykietowanie jest zbyt kosztowne, oczywistym obszarem badań staje się poszukiwanie metod uczenia maszynowego zdolnych do utrzymania wysokiej poprawności analizy danych nawet w obliczu bardzo posiadania zbiorów danych treningowych o bardzo ograniczonych licznościach. Uzyskanie wkładu w tym zakresie z pewnością stanowiłoby osiągnięcie o randze spełniającej kryteria merytoryczne stawiane przed rozprawą doktorską.

Rozprawa ma charakter eksperymentalny – Autor podsumowuje w niej zbiór przeprowadzonych przez siebie eksperymentów, polegających na ocenie skuteczności rozpoznawania mówcy, mowy, płci i języka, realizowanego z wykorzystaniem różnych sposobów opisu sygnału mowy, różnych

klasyfikatorów neuronowych i dokonywanego w odniesieniu do różnych zbiorów danych, gdzie wspólnym mianownikiem jest silna redukcja liczności zbioru treningowego.

2. Teza i cele rozprawy

Obszarem zainteresowania prac Doktoranta jest analiza możliwości opracowania algorytmów rozpoznawania mówcy, bazujących na neuronowych koncepcjach uczenia maszynowego i charakteryzujących się wysokimi wskaźnikami poprawności działania, przy założeniu dysponowania zbiorami danych przykładowych o bardzo ograniczonej liczności, a więc zbiorów zawierających bardzo ograniczoną ilość informacji niezbędnej dla budowy poprawnych modeli mówców. Sytuacja rozważana przez Doktoranta jest realistycznym kontekstem wielu problemów uczenia maszynowego, rodzącym trudność prawidłowej estymacji parametrów niezwykle złożonych algorytmów rozpoznawania, jakimi są sieci neuronowe. Za główną tezę recenzowanej rozprawy można przyjąć sformułowane we wstępie rozprawy przekonanie o możliwości opracowania algorytmu rozpoznawania mówcy, bazującego na analizie swobodnych wypowiedzi i pozwalającego na poprawną realizację zadania nawet dla bardzo dużych ilości rozważanych kategorii, z wykorzystaniem prostych metod i ograniczonych zbiorów danych:

Based on the mentioned motivations, the problem statement defined by the author is that it is possible to recognize a large-scale of human identities with a simplified approach considering limited and unrestricted speech samples.

Kierunki badań, które Doktorant uznaje jako istotne dla osiągnięcia tego celu, zostały ujęte w postaci zbioru pytań, sformułowanych w podpunkcie 1.2 Wstępu pracy i adresujących problemy: wyboru odpowiedniej reprezentacji sygnału mowy, zasadności przeprowadzania lub możliwości uproszczenia przetwarzania wstępnego sygnału mowy, kwestii długości okna czasowego jakie ma stanowić argument analiz oraz możliwości wykorzystania dodatkowej wiedzy dla zwiększenia poprawności rozpoznawania. Szczegółowe cele prac, które odpowiadają wymienionym powyżej kierunkom badawczym i które zostały określone przez Doktoranta, są następujące (przedstawiam wolne tłumaczenie tekstu oryginalnego, który w niektórych fragmentach nie był dla mnie całkowicie jasny):

1. Określenie reprezentacji sygnału mowy zapewniającej dogodną podstawę zarówno dla uczenia klasyfikatorów zbiorem przykładów o ograniczonych rozmiarach jak i dla przeprowadzania poprawnej klasyfikacji
2. Uproszczenie lub eliminacja fazy przetwarzania wstępnego zarejestrowanego sygnału mowy
3. Opracowanie metody wykorzystującej wyniki identyfikacji płci osoby mówiącej i języka wypowiedzi jako informacji dodatkowej, wspomagającej proces rozpoznawania mówcy
4. Opracowanie metod rozpoznawania mówcy, budowanych w oparciu o ograniczone zbiory danych, skutecznych dla różnych uwarunkowań procesu rozpoznawania w tym również dla kontekstu eksperymentów wielkoskalowych

3. Znajomość stanu wiedzy i przyjęta metodyka badawcza

Punktem wyjścia dla prezentacji prac badawczych jest zamieszczona w częściach 2 i 3 charakterystyka stanu wiedzy i prezentacja metod cząstkowych, które wykorzystał Doktorant w opracowywanych przez siebie algorytmach. Z przedstawionego przeglądu widać, że Doktorant ma dobrą orientację w obszarach dotyczących tematyki rozprawy. Głównymi problemami powstającymi w sytuacji ograniczania zbiorów danych dostępnych dla treningu są zjawiska ‘zapamiętywania’ przykładów (ang. overfitting), wynikające z nadmiernej liczby parametrów algorytmów w stosunku do posiadanych informacji oraz brak lub niedostateczna reprezentacja w zbiorze przykładów wiedzy o

strukturze rozkładów próbek odpowiadających analizowanym klasom. Autor poprawnie identyfikuje te problemy i wskazuje na zaproponowane dotychczas metody redukcji ich skutków, takie jak transfer wiedzy, regularyzacja, stosowanie metod losowego przerzedzania struktury sieci podczas treningu (ang. *dropout*), przycinanie sieci, czy też wczesne przerywanie procesu uczenia (early stopping).

Cechą charakterystyczną przeglądu opracowanych dotychczas metod biometrycznej analizy sygnału mowy jest przyjęcie perspektywy wykorzystania sieci neuronowych jako głównego narzędzia realizacji rozpoznawania. W konsekwencji, Doktorant pomija szereg wcześniejszych etapów rozwoju prac w obszarze rozpoznawania mówcy i wypracowanych w nich koncepcji. Podstawowym powodem, dla którego zaniechanie obszernej prezentacji 'klasycznych' algorytmów rozważanej dziedziny może zostać zaakceptowane jest na pewno skokowy wzrost skuteczności rozpoznawania, dokonany za sprawą wprowadzenia metod uczenia głębokiego, które w niewielkim stopniu korzystają ze spuścizny wspomnianych metodyk analizy sygnału mowy. W efekcie, trudno jest się upierać, że te klasyczne metodyki stanowią niezbędną podstawę dla zrozumienia i zapewnienia możliwości prowadzenia prac nad rozważanymi aplikacjami. Mankamentem przedstawionego przez Doktoranta przeglądu jest brak jasnego uporządkowania prezentowanego materiału według czytelnych kryteriów jakościowo różnicujących różne grupy metod, chociaż należy przyznać, że klarowna systematyzacja w odniesieniu do ewoluujących błyskawicznie metod uczenia głębokiego nie jest zadaniem prostym.

Przydatnym fragmentem wprowadzenia jest podrozdział poświęcony zdefiniowaniu miar, które zostały w pracy zastosowane do oceny jakości działania proponowanych algorytmów oraz do porównań z właściwościami rozwiązań alternatywnych. Wśród przedstawionych wskaźników oceny brak jest niestety miar FAR (*False Acceptance Rate*), FRR (*False Rejection Rate*) i EER (*Equal Error Rate*) oraz krzywych funkcyjnych (DET – *Decision Error Trade-off* i ROC – *Relative Operating Characteristics*), stanowiących jedne z najczęściej stosowanych sposobów informowania o poprawności działania algorytmów weryfikacji biometrycznej.

Ostatnim elementem opisu kontekstu prac jest prezentacja cząstkowych metod, używanych jako komponenty prezentowanych dalej algorytmów. Oprócz omówienia metod wyznaczania reprezentacji sygnału mowy oraz wybranych przez Doktoranta koncepcji klasyfikatorów neuronowych, uznał On również za celowe zamieszczenie wyjaśnień absolutnie podstawowych koncepcji przetwarzania sygnałów. Części, w których Autor próbuje wyjaśniać czym jest i jakie jest znaczenie cyfrowego przetwarzania sygnałów, czym jest transformacja Fouriera, dyskretna transformacja Fouriera, funkcja okna itp., uważam za całkowicie zbędne, a ponieważ użyty w nich styl prezentacji jest bardzo powierzchowny, obniżają one w mojej opinii poziom merytoryczny pracy.

Metodyka prac Autora to wspólny dla wszystkich podejmowanych przez Niego wątków schemat postępowania, w którym dla realizacji określonego zadania stosuje typowe dla dziedziny rozpoznawania mówcy podejście: wyznacza (na różne sposoby) reprezentację sygnału mowy, która jest następnie wykorzystana w celu uczenia klasyfikatorów neuronowych (czyli budowania modeli mówców). W odniesieniu do realizacji pierwszego zadania, Doktorant wykorzystuje kilka koncepcji, zarówno znanych (reprezentacja sygnału mowy za pomocą modelu autoregresyjnego Burga, klasycznych współczynników MFCC i deskryptora nazwanego przez Niego 'metodą projekcji'), jak i jednej oryginalnej, autorskiej meotdy (zbioru wartości własnych macierzy zbudowanych na podstawie odpowiednio przetworzonego widma sygnału w oknie analizy). W odniesieniu do zadania klasyfikacji, korzysta zarówno z prostych architektur neuronowych, jak i z rekurencyjnych architektur głębokich.

Wspólnym mianownikiem wszystkich prac jest dzielenie posiadanego zbioru przykładów na części testową i treningową w proporcjach wybitnie preferujących liczebność części testowej. Zabieg ten ma stanowić odzwierciedlenie rzeczywistych uwarunkowań rozpoznawania, gdzie liczba posiadanych przykładów wypowiedzi danej osoby, będzie stanowić niewielką cząstkę wszystkich wypowiadanych przez tę osobę słów. Realizując taki zamysł, Autor przyjmuje arbitralne proporcje (np. 15% przykładów stanowi zbiór treningowy, zaś 85% jest umieszczanych w zbiorach walidacyjnym, jeśli występuje, i testowym) i próbuje budować na ich podstawie modele mówcy i algorytm jego rozpoznawania. Wydaje się, że celem przedstawionego podejścia jest zastosowanie innej niż

powszechnie stosowana perspektywy korzystania z danych przykładowych: Doktoranta interesuje bardziej symulacja oceny działania algorytmów w warunkach docelowych (stąd powiększanie rozmiaru zbioru testowego) niż dążenie do maksymalizacji ilości informacji używanej do budowy modeli mówców (realizowane przez powiększanie udziału części treningowej). Na uznanie zasługuje obszerna weryfikacja eksperymentalna algorytmów, oceniająca ich działanie dla różnych zbiorów parametrów używanych dla procedur cząstkowych (zarówno dla tworzenia reprezentacji sygnału mowy, jak i budowy klasyfikatora).

4. Merytoryczna ocena prac badawczych Doktoranta

Recenzowana rozprawa, w części opisującej dorobek naukowy Doktoranta. podzielona jest na części odpowiadające różnym uwarunkowaniom procesu rozpoznawania mówcy, w których Autor stosuje lub rozwija różne metody analizy. Przyjęty układ odpowiada kolejnym etapom prac wykonywanych przez Autora, w ramach których stopniowo zwiększano skalę złożoności podejmowanego problemu: od identyfikacji osób należących do niewielkiego zbioru klas, dokonywanej na podstawie ustalonych wypowiedzi, poprzez rozpoznawanie realizowane dla ograniczonego zbioru mówców, ale tym razem, na podstawie dowolnych wypowiedzi (mowy swobodnej), rozszerzone następnie o wykorzystanie dodatkowej wiedzy o płci i języku wypowiedzi, kończąc na rozpoznawaniu osób na podstawie mowy swobodnej w warunkach zbioru mówców zawierającego bardzo dużą liczbą kategorii. Przyjęcie takiego scenariusza prac badawczych jest naturalne, chociaż nieco brakuje mi głębszej analizy ograniczeń stosowanych podejść i wynikających z tego pomysłów ich modyfikacji, zwiększających potencjał w konfrontacji z trudniejszymi problemami. Mimo tego zastrzeżenia, przyjęta koncepcja stanowi właściwą konstrukcję dla możliwości oceny przydatności stosowanych przez Doktoranta różnych szczegółowych koncepcji rozwiązań.

Elementem spajającym wszystkie opracowane przez Doktoranta algorytmy jest wykorzystanie klasyfikatorów neuronowych, słusznie wskazanych jako najbardziej obecnie skuteczna metodyka rozwiązywania problemów rozpoznawania. Podstawową cechą modeli neuronowych, przede wszystkim zaś, dominujących pod względem poprawności przetwarzania, modeli głębokich, jest ich ogromna złożoność, implikująca konieczność stosowania wielkich zbiorów treningowych, co rodzi konieczność wprowadzania odpowiednich, wspomnianych wcześniej mechanizmów redukcji negatywnych skutków braku możliwości zapewnienia odpowiednio licznego zbioru danych. Pierwszym sposobem na redukcję rozmiaru zestawu przykładów uczących było wykorzystanie przez Doktoranta możliwie najprostszych koncepcji klasyfikatorów neuronowych: Probabilistycznej Sieci Neuronowej (ang. *Probabilistic Neural Network*, PNN) oraz dwuwarstwowej, jednokierunkowej sieci o radialnych funkcjach bazowych (ang. *Radial-Basis Neural Networks*, RBNN), a następnie sprawdzenie stopnia złożoności analiz, które można realizować z satysfakcjonującą poprawnością. Dla realizacji zadań rozpoznawania w kontekście problemów trudnych, Doktorant zastosował jedyne rozsądne podejście – sieć głęboką, ale wzbogacił ją o dwa istotne mechanizmy zmniejszania wrażliwości na ograniczone zbiory danych: augmentację oraz uzupełnianie procesu uczenia o dodatkową wiedzę, przy czym zabieg ten został przeprowadzony na dwa różne sposoby. W pierwszym przypadku, Autor postanowił wzbogacić proces uczenia o wiedzę na temat płci i języka wypowiedzi, stanowiących przesłanki pozwalające na zawężenie puli hipotez stawianych co do tożsamości osoby, a tym samym, zwiększenie poprawności rozpoznawania. Drugi zabieg, znacznie ciekawszy, polegał na zastosowaniu koncepcji dwukierunkowej propagacji informacji, oferowanej przez sieci BiLSTM (ang. *Bi-directional* LSTM). Wykorzystanie tej koncepcji pozwala na uzyskanie nieprzyczynowej, kontekstowej informacji dotyczącej sposobu wypowiadania słów, wzbogacając w porównaniu do zwykłych sieci rekurencyjnych sposób biometrycznej reprezentacji sygnału, a tym samym dając podstawy do oczekiwania podobnych rezultatów treningu obydwu koncepcji przy znacznie uboższych zasobach danych uczących.

Ponieważ prace badawcze Doktoranta przedstawiane w kolejnych częściach rozprawy są powiązane jedynie używaniem podobnych metod obliczeniowych, analizę wkładu Doktoranta do dyscypliny wygodnie jest przedstawić osobno dla materiału przedstawionego w każdym z rozważanych modułów

rozprawy. W odniesieniu do pierwszego poruszonego w rozprawie zadania – rozpoznawania mówców należących do niewielu klas, realizowanego na podstawie izolowanych słów, jedynym wartym odnotowania wynikiem prac jest wskazanie przez Doktoranta możliwości skutecznego rozwiązania postawionego problemu przy użyciu bardzo prostych metod, zarówno w odniesieniu do przyjętego opisu sygnału mowy, jak i w odniesieniu do zastosowanych metod klasyfikacji. Okazuje się, że wykorzystanie deskryptora zawierającego jedynie komponenty widma mocy, wyznaczonego dla całego wyrazu, jak również deskryptorów zbudowanych na jego podstawie (wartości własne macierzy Toeplitza) oraz prostych klasyfikatorów (najbliższej średniej, PNN i RBN) pozwala na uzyskanie efektów, które dla wielu zastosowań mogłyby okazać się satysfakcjonujące. Niestety, przyjęta w pracach metodyka rozpoznawania: używanie całych izolowanych słów jako podstawy budowy modelu mówcy i konkretny zbiór danych o niewielkiej liczbie klas, wykluczają możliwość uogólnienia uzyskanych wniosków. Dodatkowo, zastosowanie 'mocniejszych' algorytmów dla rozważanego zadania mogłoby skutkować uzyskaniem 100% skuteczności (na przykład, ciekawe jaki byłby wynik użycia metody SVM). Biorąc pod uwagę te dwie konkluzje, uważam, że z punktu widzenia naukowego, przedstawiony rozdział nie wnosi zauważalnego wkładu do dziedziny.

Materiał przedstawiony w Rozdziale 5 dotyczy problemu rozpoznawania mówców na podstawie swobodnych wypowiedzi, przy czym Autor rozważa tu zarówno zadanie identyfikacji mówcy w obrębie zadanego zbioru klas oraz zadanie weryfikacji tożsamości mówcy przeprowadzane dla otwartego zbioru mówców. Dla realizacji założonych celów Autor używa metod testowanych we wcześniejszej części, uzupełniając opis sygnału mowy dodatkowo o deskryptory zawierające współczynniki MFCC. Dokonanie oceny możliwości oferowanych przez proste koncepcje klasyfikacji: PNN i RBN w zadaniu rozpoznawania mówcy dokonywanym w odniesieniu do trudniejszych danych, pochodzących z rzeczywistego, a więc praktycznego kontekstu, może być uznana za wkład Autora do dziedziny. Uzyskane przez Autora wskaźniki ilościowej poprawności wykonywania rozważanych analiz mogą stanowić interesujące przesłanki do ewentualnego wykorzystania rozważanych koncepcji w analizach dokonywanych przez coraz powszechniejsze autonomiczne systemy o ograniczonych mocach obliczeniowych i zasobach (tzw. edge computing). Efekty prac Autora potwierdzają duży potencjał stosowania prostych metod analizy, mogących dawać wystarczające z punktu widzenia wielu rodzajów zastosowań wyniki, sygnalizowanych wielokrotnie wcześniej w literaturze w odniesieniu na przykład do koncepcji ELM (ang. extreme learning machines), będącej rozwiązaniem pośrednim między rozważanymi przez Autora sieciami PNN i RBNN. Wątkiem prac Doktoranta zasługującym na większą uwagę, stanowiącym kolejny element autorskiego wkładu do dziedziny, jest przyjęty przez Niego sposób korzystania z opisu sygnału mowy, wykorzystującego współczynniki MFCC. Otóż, klasyfikatory są przez Doktoranta trenowane na podstawie uśrednianych wektorów współczynników MFCC, wydzielanych z kolejnych okien w obrębie dwusekundowych fragmentów analizowanych wypowiedzi. Takie podejście jest rzadko spotykane, ale nie oznacza to, że jest błędne. Przedstawiony zabieg nie miałby żadnego sensu w odniesieniu do rozpoznawania mowy, ale w odniesieniu do rozpoznawania mówcy ma uzasadnienie, bowiem może być interpretowany jako mechanizm redukcji wewnątrzklasowej zmienności parametrów opisujących pracę narządów aparatu fonacji i artykulacji, przy czym wydaje się, że o ile w odniesieniu do procesu fonacji zabieg ten jest pożądany, o tyle w odniesieniu do procesu artykulacji, podejście to może być dyskusyjne. Szkoda, że Autor nie poświęcił więcej uwagi i miejsca na dyskusję i zbadanie omawianej kwestii, zbywając rozważania stwierdzeniem, że uzyskał większą dokładność rozpoznawania niż w przypadku wykorzystywania w rozpoznawaniu wszystkich wektorów wydzielanych z analizowanych segmentów mowy (chociaż podejmował takie próby, na przykład, analizując w algorytmach informacje zawarte w macierzach kowariancji zbiorów wektorów MFCC).

Zagadnienia opisane w Rozdziale 6 rozprawy dotyczą prac nad rozpoznawaniem płci i języka wypowiedzi, zmierzających do uzyskania dodatkowej informacji, przydatnej w procedurze rozpoznawania mówcy. Kierunek prac jest jak najbardziej uzasadniony, a uzyskany wzrost poprawności rozpoznawania, potwierdza zasadność podjęcia tego wątku. Do realizacji postawionego zadania Autor używa sieci LSTM, analizującej kolejne fragmenty sygnału mowy, reprezentowane za pomocą wektorów MFCC. Jako metodę referencyjną Autor stosuje metodę łączącą opis sygnału za pomocą współczynników widma obliczonych metodą Burga i klasyfikator PNN, co daje możliwość

konfrontacji tej prostej metodyki z kolejną kategorią problemów. Na uznanie zasługuje bardzo obszerny zbiór eksperymentów mających na celu poszukiwanie optymalnej architektury sieci LSTM (kompromis między poprawnością analizy a liczbą elementów sieci i wielkością wektora wejściowego). Bazując na uzyskanych wynikach, Doktorant formułuje wyrażenie mające pełnić rolę wskazówki dla doboru złożoności sieci, uzależniającej liczbę jednostek w warstwach od liczby wejść i liczby klas. Zaproponowany przez Doktoranta współczynnik jest jednak w moim przekonaniu adekwatny tylko w odniesieniu do konkretnego, dysponowanego przez Niego zbioru danych i nie ma szans stanowić wskazówki o charakterze ogólnym, która mogłaby być przydatna dla szacowania wymaganej złożoności architektur sieci. Również niezwykle obszerne są prace Autora nad zbadaniem zależności poprawności rozpoznawania jako funkcji długości wektora współczynników MFCC. Pewien niedosyt budzi zaniechanie próby sprawdzenia efektu pomijania początkowych współczynników MFCC (typowego w rozpoznawaniu mówcy) na dokładność analizy i skupienie się wyłącznie na różnicowaniu liczby uwzględnianych współczynników (zakładam, że wektory są zawsze tworzone z początkowych współczynników, ale być może tak nie jest, o czym Autor jednak nie informuje). Wartościowym elementem prac Doktoranta jest przedstawienie (w odróżnieniu od opisu podanego Rozdziale 5) osobno wyników eksperymentów dla przypadków treningu klasyfikatora na bazie uśrednianych wektorów MFCC i treningu wykorzystującego wektory MFCC obliczane dla wszystkich segmentów rozważanych wypowiedzi.

Zwieńczeniem prac przedstawionych w rozdziale 6 są wyniki eksperymentów rozpoznawania płci i języka wypowiedzi oraz rozpoznawania mówcy, wykorzystującego wcześniej zidentyfikowany język wypowiedzi, dokonywanych za pomocą obydwu rozważanych przez Doktoranta strategii klasyfikacji. Wnioski płynące z analizy uzyskanych wyników są ciekawe i w większości, uwypuklone przez Autora. Pierwszy z nich to uzyskanie większej poprawności rozpoznawania w identyfikacji języka wypowiedzi i płci, a także rozpoznawania mówcy wspieranego informacją o zidentyfikowanym języku wypowiedzi, w odniesieniu do kombinacji MFCC-LSTM. Wniosek ten wcale nie jest oczywisty w kontekście stosunkowo ograniczonej liczby przykładów użytych w procesie treningu. Drugim ciekawym spostrzeżeniem jest całkiem niezła poprawność rozpoznawania mówcy, języka i płci dla kombinacji MFCC-PNN. Szkoda, że Autor nie przedstawił wyniku eksperymentów rozpoznawania mówcy dla tej kombinacji, uwzględniającej wykorzystanie wiedzy o zidentyfikowanej płci i języku wypowiedzi – wyniki takie byłyby interesujące z perspektywy wspominanej już wcześniej dziedziny ‘edge computing’.

Ostatnia część rozprawy dotyczy rozpoznawania mówcy w warunkach bardzo obszernego zbioru klas, dokonywanego na podstawie mowy swobodnej. Punktem wyjścia dla badań Doktoranta było wykorzystanie jednego ze znanych podejść do realizacji tego zadania – użycie kombinacji, zawierającej współczynniki MFCC jako formę reprezentacji informacji o sygnale mowy i głębokiej sieci LSTM (ang. *Long Short-Term Memory*), jako klasyfikatora. Doktorant przeprowadził obszerne eksperymenty, których celem było stwierdzenie wpływu redukcji ilości informacji dostarczanej w procesie treningu sieci przez zbiór przykładów na dokładność rozpoznawania. Ponieważ konkluzja, jaka wynikała z prac nie była obiecująca, Doktorant postanowił zaproponować inną metodę realizacji zadania – wykorzystać w charakterze klasyfikatora sieć BiLSTM. Pomysł wykorzystania ‘dwukierunkowego’ kontekstu analizy jest interesujący: osobnicze zmiany współczynników MFCC tradycyjnie były modelowane poprzez dodawanie do wektora cech różnic pierwszego i drugiego rzędu (delta-MFCC i delta-delta-MFCC). Zastosowane przez Doktoranta rozwiązanie zachowuje pełną informację o sygnale i pozwala na analizę kontekstu, nie wprowadzając redundancji reprezentacji, skutkującej niepotrzebnym zwiększaniem liczby wymaganych parametrów algorytmu analizy. Przedstawione rozwiązanie ma dobre rokowania – Doktorant potwierdził wysoką poprawność rozpoznawania (znaczącą lepszą niż w przypadku użycia sieci LSTM) z użyciem wielkoskalowej bazy nagrań (o złożoności rzadko stosowanej w dotychczas prowadzonych eksperymentach) i z wykorzystaniem ograniczonych rozmiarów zbioru treningowego.

Podsumowując ocenę zawartości merytorycznej pracy, Doktorant zawarł w niej zbiór rozwiązań, pomysłów i wyników badań, stanowiących ciekawy i pożyteczny wkład do dziedziny rozpoznawania mówcy. Za najistotniejsze osiągnięcie Autora, dające podstawy do uznania jego prac za spełniające

wymagania sformułowane dla uzyskania stopnia doktora, uważam przede wszystkim koncepcję wykorzystania w rozpoznawaniu sieci BiLSTM. Również wartościowym wynikiem prac Doktoranta jest opracowanie metody wzbogacania procesu biometrycznego rozpoznawania mówcy poprzez odpowiednie włączanie do algorytmu analizy dodatkowej wiedzy – wyniku identyfikacji języka wypowiedzi. Kolejnym ciekawym elementem prac, wartym dalszych pogłębionych studiów, jest wykazanie skuteczności redukcji zmienności opisu mówcy, wykorzystującego współczynniki MFCC poprzez proste uśrednianie. Mimo braku osiągnięcia satysfakcjonujących rezultatów, na uznanie zasługuje również podejmowanie przez Autora prób poszukiwania nowych, dyskryminatywnych sposobów reprezentacji sygnału mowy (do tego wątku odniosę się w dalszej części rozdziału). Wreszcie, cennym wynikiem prac jest samo przedstawienie uzyskiwanych parametrów procesu rozpoznawania w różnych uwarunkowaniach, dla algorytmów wykorzystujących bardzo proste struktury neuronowe. Na uznanie zasługuje też zgromadzenie i opracowanie bardzo obszernego materiału eksperymentalnego, który może stanowić cenną podstawę dla prowadzenia dalszych prac badawczych zarówno przez Doktoranta i członków jego zespołu, jak i innych środowisk naukowych.

Oprócz przedstawionych powyżej elementów, świadczących o wkładzie Autora do rozważanej dyscypliny, w pracy znajdują się fragmenty i tezy, które uważam za nieuzasadnione, dyskusyjne lub nie do końca zrozumiałe. Wątkiem prac Autora pozostawiającym niedosyt poznawczy jest przedstawiony w rozprawie nowy sposobu opisu sygnału mowy, zaproponowany przy współudziale Autora i opublikowany wcześniej w renomowanym czasopiśmie (IEEE Transactions on Industrial Electronics, pozycja nr 110 spisu literatury). Sformułowana koncepcja reprezentacji opisu okna czasowego sygnału mowy – całkowicie odmienna od stosowanych, jest poddana analizy możliwości zapewnienia dobrej zdolności dyskryminacyjnej w eksperymentach rozpoznawania mówcy, opisanych w Rozdziałach 4 i 5. Oczekiwane korzyści zastosowania metody – uzyskanie potencjalnie znaczącej redukcji złożoności obliczeniowej procesu wyznaczania reprezentacji, są uzasadnione, stanowiąc ciekawe i potencjalnie atrakcyjne uzasadnienie dla jej wprowadzenia. Niestety, przedstawiony w pracy opis koncepcji jest całkowicie pozbawiony prób wyjaśnienia idei, przyświecającej tworzeniu tej szczególnej formy opisu. Zaprezentowany przez Autora pomysł, to przekształcenie dyskretnego widma mocy sygnału w oknie (uzyskanego metodą autoregresyjną Burga), najpierw - do postaci ilorazu wielomianów, zbudowanych na podstawie współrzędnych próbek widma, a następnie, do postaci macierzowych deskryptorów zbudowanych ze współczynników rozwinięcia rozważanego ułamka w szereg Taylora. Deskryptory mają strukturę macierzy Toeplitza, dla których wyznaczane są następnie minimalne wartości własne, stanowiące elementy deskryptora rozważanego fragmentu sygnału mowy. Droga przebyta między sygnałem oryginalnym i deskryptorem jest zbyt skomplikowana, by zrozumieć (przynajmniej w moim przypadku) istotę przedstawionego pomysłu. Autor w przedstawionej pracy nie tłumaczy, dlaczego zdecydował się na taką formę deskryptora, nie wiadomo więc, dlaczego właściwości osobnicze mogą być przez ten deskryptor poprawnie uchwycone (być może, istotne jest raczej tłumienie zmienności wewnątrzklasowych?). Również w pracy oryginalnej [110] nie znalazłem odpowiedzi na to pytanie, więc mimo niewątpliwie fundamentalnie nowego podejścia do opisu sygnału mowy, trudno mi wnioskować o trafności sformułowanej propozycji. Dodatkowo, nie rozumiem przedstawionych przez Autora rozważań dotyczących decymacji widma mocy i wyboru próbek reprezentujących sygnał mowy, a także rozważań dotyczących reprezentacji widma w postaci obrazu (nie chodzi tu bynajmniej o spektrogramy). Być może istotne znaczenie w trudności tłumaczenia myśli ma problem bariery językowej, bo przedstawione rozważania są sformułowane w sposób daleki od jednoznaczności i jasności. Wątek konstrukcji deskryptora uważam za niewykorzystaną okazję: być może kryje się w nim spory potencjał, który nie został w moim odczuciu odpowiednio zbadany (należy zaznaczyć, że analiza spektralna macierzy kołowych prowadzi do transformacji DFT – gdyby klarownie zidentyfikować istotę i cel dokonywanych przez Autora przekształceń, być może wyłoniłaby się interesująca i pożyteczna interpretacja przedstawionej, skomplikowanej procedury).

Motywem przewijającym się przez całą pracę jest uwypuklanie znaczenia proporcji między liczebnością części testowej i treningowej zbioru przykładów posiadanych dla uczenia sieci. Nie rozumiem znaczenia tego, forsowanego przez Autora rozprawy, parametru i nie wiem jakie są podstawy dla przywiązywania temu parametrowi istotnego znaczenia. Zapewnienie wysokiej skuteczności działania algorytmu uczenia maszynowego wymaga zbudowania właściwego modelu

problemu, stanowiącego kompromis między wspominanymi przez Autora biegunami: nadmiernego dopasowania do próbek (przy zbyt małej liczbie przykładów w stosunku do liczby estymowanych parametrów) i braku dostatecznej 'pojemności' algorytmu dla reprezentacji rzeczywistej struktury danych odpowiadających problemowi (zbyt mała liczba parametrów algorytmu w stosunku do wymaganej złożoności modelu). Oznacza to, że w przypadku oceny możliwości rozwiązania problemu z wykorzystaniem danej architektury obliczeniowej, istotna jest relacja między trzema czynnikami: liczbą parametrów algorytmu, złożonością rzeczywistego rozwiązania problemu i ilością informacji dostarczanej o problemie w danych uczących. Wskazywana przez Autora proporcja jest w takim przypadku zupełnie nieistotna i nie ma związku z procesem budowania poprawnie działającej sieci neuronowej. Co więcej, jeżeli dla danej sieci Autor stwierdzi poprawne działanie dla zadanej proporcji liczności próbek zbiorów treningowego i testowego, jakakolwiek zmiana struktury tej sieci (dodawanie lub usuwanie jednostek) może doprowadzić do zupełnie innej konkluzji. Również, z faktu uzyskania określonej poprawności działania dla danej sieci i danej proporcji, nie wynika możliwość uzyskania podobnego wyniku dla sieci o innej architekturze. Wreszcie, ta sama proporcja może dla danej sieci zapewnić odpowiednią ilość informacji dla treningu lub nieodpowiednią ilość informacji, zależnie od zbioru używanych danych i sposobu losowania próbek między część testową i treningową. W konsekwencji, uważam podniesienie przedstawionego przez Autora wskaźnika do rangi istotnego parametru procedury uczenia sieci neuronowej za nieporozumienie, a fragmenty tekstu skupiające się na tym zagadnieniu, za niepotrzebne (chyba, że istnieje jakieś inne wyjaśnienie dla jego istotności, na które nie wpadłem).

Oprócz dwóch wskazanych, najistotniejszych dyskusyjnych elementów rozprawy, znajduje się w niej również kilka fragmentów wątpliwych merytorycznie lub trudnych do zrozumienia, takich jak:

- W podrozdziale 3.6.1 (Classical Approach) Autor opisuje klasyfikację metodą najbliższej średniej (NM – nearest mean), ale czyni to w sposób wyjątkowo zagmatwany, ilustrując dodatkowo proste koncepcje niezrozumiałym wzorem (3.9), który prawdopodobnie miał pokazać, że wynikiem jest etykieta klasy, dla której odległość wektora średniego klasy (wzorca) do nieznanego wektora jest najmniejsza. Autor modeluje klasę za pomocą wektora średniego, a jako miarę podobieństwa stosuje miarę L1. Nie rozumiem dlaczego Autor przyjmuje taki wariant metody NM jako metodę bazową. Jest to w oczywisty sposób wybór zły: po pierwsze, zakłada że klasy mają jednomodowy rozkład Gaussa, co w przypadku analizy mowy jest błędne. Po drugie, założony dla każdej klasy rozkład Gaussa ma diagonalną macierz kowariancji, co jest uproszczeniem również kompletnie nie pasującym do rozważanej dziedziny aplikacji. Nie wiem dlaczego jako bazowej metody Autor nie użył, równie prostej koncepcyjnie, metody k-NN?
- Opisy referencyjnych metod klasyfikacji zawierają błędy. Na przykład, w odniesieniu do metody SVM podane wyjaśnienie tzw. 'kernel trick': „*… the kernel trick means transforming data into another dimension that has a clear dividing margin between classes of data*." jest nieprawdziwe. Podobnie, prezentując problemy treningu sieci neuronowych pisze, że: „*Using incorrect parameters for training the NNs may lead to underfitting and using insufficient information may lead to overfitting*.", Pierwsza część myśli jest błędna (w zjawisku nie chodzi o 'złe parametry')), zresztą parę zdań dalej Autor poprawnie identyfikuje przyczyny i sposoby przeciwdziałania zjawisku 'niedopasowania'.
- W rozprawie przewija się koncepcja traktowania sygnału mowy lub jego widma w kategoriach obrazu, ale sposób rozumienia tej odpowiedniości nie jest dla mnie jasny. W przypadku estymacji widma metodą autoregresyjną, jako obraz Autor traktuje dyskretną funkcję – dla prążków widma, amplitudy traktuje jako rzędne punktów, zaś częstotliwości, jako odcięte . Stwierdzenia typu : „*Next the complete speech signal is processed as an image the by power spectrum estimation (PSE)*." powiększają skalę niejasności, bo nie do końca wiadomo jak rozumieć takie stwierdzenia, a przede wszystkim, jakie korzyści dla analizy sygnału wynikają z forsowanego 'obrazowego' opisu funkcji jednowymiarowych.

5. Ocena formalnej strony rozprawy

Ostatni wątek przedstawionej recenzji dotyczy oceny formalnej strony przedstawionego tekstu. Język angielski nie jest językiem ojczystym ani Doktoranta, ani moim, więc przedstawiony w języku angielskim tekst, stanowiący formalny interfejs przekazu myśli, stanowi potencjalne źródło problemów w poprawnym rozumieniu przedstawianych wywodów. Niestety, z mojej perspektywy, język rozprawy nie daje wystarczająco precyzyjnej podstawy dla śledzenia zawartych w niej treści, co mogło działać na niekorzyść Doktoranta. Jako przykład, braku precyzji formułowania myśli, który uważam za reprezentatywny dla całej pracy, chcę przytoczyć drugie zdanie wstępu rozprawy, które brzmi:

„*Speaker recognition (SR) in particular is a performance trait, i.e., the user performs a task to be recognized*."

Przy wykazaniu pewnego wysiłku można dostrzec, co Autor chciał w tym zdaniu powiedzieć: ma to być informacja, że mowa jest cechą behawioralną, ale identyfikacja tej treści wymaga całkowitego przeformułowania przedstawionego tesktu, bo w początkowej wersji jest on pełen błędów. Po pierwsze, rozpoznawanie mówcy (*speaker recognition*) nie jest cechą (*trait*). Również, to nie zadanie (*task*) a użytkownik (*user*) ma być celem rozpoznawania, jak wynikałoby z gramatycznej konstrukcji zdania. Niestety, trudność precyzyjnego wysławiania się w języku dziedziny prowadzi do często frustrujących problemów ze zrozumieniem innych wywodów Autora, których sens nie zawsze udawało mi się przeniknąć, choć być może były poprawne. Problem komunikacji pojawia się również w zakresie doboru słownictwa – prawdopodobnie sposób upatrywania znaczenia pewnych terminów jest funkcją naleciałości języka ojczystego, gdy perspektywa staje się różna, pojawia się obszerny obszar niezrozumienia. Przykładowym słowem, często stosowanym przez Autora w sposób odmienny od znaczenia, jakie ja temu słowu przypisuję jest wyraz „*meanwhile*'.

6. Wniosek końcowy

Przedstawiona do recenzji rozprawa doktorska Pana magistra inżyniera Mohammada Nammmous pt. „New Approaches in Speech Recognition from Isolated Words to Practical Solutions" **spełnia** według mnie wymagania określone w odnośnej ustawie o stopniach i tytule naukowym, czyli zawiera elementy, stanowiące wkład do dyscypliny naukowej informatyka techniczna i telekomunikacja i tym samym **wnioskuję o dopuszczenie jej Autora do publicznej obrony**.