Uniwersytet im. Adama Mickiewicza
w Poznaniu
Zakład Lingwistyki Informatycznej i Sztucznej Inteligencji
PL-61-614 Poznań, ul. Umultowska 87, tel. +48-61-8295380, fax +48-61-8295315
Kierownik Zakładu: prof. zw. dr hab. Zygmunt Vetulani
http://vetulani.home.amu.edu.pl
vetulani@amu.edu.pl

Poznań, 17.09.2019.

Recenzja rozprawy doktorskiej mgr. inż. Stanisława Kacprzaka "Spoken language recognition in i-vectorspace using cluster-based modeling" napisanej pod kierunkiem dra hab. inż. Bartosza Ziółko, promotora, oraz dra inż. Konrada Kowalczyka, promotora pomocniczego

Do oceny przedłożona została rozprawa w języku angielskim w postaci tekstu, na który składa się siedem rozdziałów merytorycznych i obszerna bibliografia zawierająca 189 pozycji).

Problematyka badawcza

Praca dotyczy badań w zakresie automatycznego rozpoznawania języka mówionego. W życiu codziennym spotykamy się od czasu do czasu z potrzebą rozpoznania, w jakim języku rozmawiają znajdujący się obok ludzie, bądź w jakim języku zwraca się do nas rozmówca (klient, interesant). Najczęściej nie mamy z tym większych problemów. Dzieje się tak dlatego, iż na ogół dysponujemy sporą wiedzą o sytuacji, w której się znajdujemy (miejsce, czas, nasza rola, otoczenie, …), a także o relacji lub jej braku z rozmówcami czy, ogólniej, osobami mówiącymi, które są przedmiotem naszego zainteresowania. Będziemy jednak bezradni, gdy wiedza ta okaże się niedostateczna, czy wręcz nie będzie jej wcale, a jedynym materiałem będą zapamiętane wypowiedzi. W takich przypadkach rozpoznanie języka mówcy na podstawie samego tylko sygnału mowy okazuje się być dużym wyzwaniem dla technologii sztucznej inteligencji. Nietrudno zidentyfikować można przynajmniej kilka ważnych obszarów, gdzie ta sytuacja ma miejsce. Dotyczyć to będzie np. monitorowania dużych zbiorowisk ludzkich o trudnym do przewidzenia, czy wręcz niemożliwym do określenia składzie etnicznym. Przykład mogą stanowić wielkie huby lotnicze, wieloetniczne zgromadzenia masowe – artystyczne, sportowe, religijne czy polityczne, a także katastrofy naturalne bądź przemysłowe w wielkich aglomeracjach. W takich sytuacjach pobrane (np. zdalnie) próby mowy będą mogły pomóc w

ustaleniu składu etnicznego czy narodowościowego danego zbiorowiska i ustaleniu jego struktury przy nieznanej *a priori* liczbie języków. W literaturze wspomina się o potencjalnych zastosowaniach w przedsięwzięciach wywiadowczych, działaniach monitorowania prewencyjnego dla potrzeb bezpieczeństwa, w projektowaniu wielojęzycznych systemów zgłoszeniowych o wielokanałowym i wielojęzycznym wejściu głosowym (typu telefon 112). Wreszcie, niezależnie od autonomicznych zastosowań, identyfikacja języka dla sekwencji sygnału głosowego może pełnić istotną rolę wspomagającą (analiza wstępna, pre-processing) w wysokopoziomowych systemach przetwarzania mowy (rozumienie, dialog człowieka z systemem informatycznym). Warto odnotować także zastosowania w ramach systemów co do zasady jednojęzycznych. Ciekawym przypadkiem jest casus języka arabskiego w wariancie algierskim, gdzie obserwuje się często spotykane zjawisko przeplatania warstwy podstawowej (arabskiej) przez niespodziewane (i niesygnalizowane prozodycznie) wstawki w postaci pełnych zdań lub ich sekwencji w języku francuskim.

Przy całej doniosłości problematyki identyfikacji języka mówcy zainteresowanie tą tematyką na szerszą skalę ma stosunkowo krótką historię, choć pierwsze ślady tych badań sięgają już lat 1970-tych. Publikacji z tego okresu jest jednak niewiele i są niekiedy trudno osiągalne, jak np. (cytowany w Frozprawie jako pozycja [47]) raport techniczny z Air Force Rome Air Development Center, 1974. Istotna intensyfikacja prac w kierunku identyfikacji języka przypada na koniec pierwszej dekady XXI wieku (za wyjątkiem pionierskiej pracy Reynolds'a i innych (1998), która jednak – jak się zdaje – nie miała bezpośrednich kontynuatorów), w związku z wypracowaniem metod, które okazały się skuteczne we wcześniejszych badaniach w zakresie identyfikacji mówcy i w sposób naturalny znalazły zastosowanie w rozpoznawaniu języka dla fragmentu mowy ciągłej. Metody te to:
1) „Joint Factor Analysis" /metoda będąca wariantem analizy czynnikowej (factor) polegającej na redukcji opisu zależności wyrażonej przez zaobserwowane powiązane zmienne losowe do opisu przy pomocy mniejszej liczby nieobserwowanych niezależnych zmiennych losowych/,
2) oraz uproszczenie powyższej metody zwane metodą „i-vector" / od *i(ntermediate)-vector*/.
Podobieństwo problemu identyfikacji mówcy do problemu identyfikacji języka określonego fragmentu tekstu sprawiło, że stosowanie i-wektorów w rozpoznawaniu języka stało się obok metod głębokiego uczenia (ang. *Deep Learning*) techniką odpowiadającą obecnemu stanowi wiedzy (por. str. 9 i 16 rozprawy).

W Polsce, wg mojej wiedzy, problematyka automatycznej identyfikacji języka segmentów mowy praktycznie nie jest podejmowana poza zespołem prof. Bartosza Ziółko z AGH (w ciągu ostatnich 5 lat).

Sformułowanie zagadnienia badawczego rozprawy doktorskiej

Doktorant sytuuje główne zagadnienie badawcze rozprawy w obszarze „rozpoznawania języka mówionego" (ang. Spoken Language Recognition /SLR/), przez co rozumie on „proces automatycznego rozpoznawania bądź weryfikowania języka w nagraniu mowy" („process of automatically identifying or verifying the language spoken in a recording of the speech", str. 5). Doskonale orientując się w prowadzonych na świecie przed rokiem 2015 badaniach (czemu daje wyraz w obszernym i kompetentnym przeglądzie stanu badań zawartym w Rozdziałach 2 i 3 /ponad 35% tekstu głównego/) Doktorant odnotowuje daleko idące analogie między metodami wyodrębniania grup podobieństwa pomiędzy użytkownikami języka a metodami wyróżniania grup podobieństwa próbek językowych („Although speaker and language clustering differ only in the selection of attribute classes that are a target of a clustering algorithm, the latter has received much less attention in the literature", str. 28). W konsekwencji, określa on **zakres przedmiotowy rozprawy** jako badania nad wyodrębnianiem grup podobieństwa próbek językowych reprezentowanych w pewnej przestrzeni wektorowej (ang. i-vector space) *(„This thesis investigates the problem of language clustering in the i-vector space"*, str. 1). W świetle tego co stwierdziliśmy powyżej, stosowanie metod analizy skupień (ang. *clustering analysis*) do nagrań próbek mowy reprezentowanych w postaci przestrzeni i-wektorowej należy uznać za podejście prawidłowe.

Cel i tezy

Celem rozprawy jest analiza możliwości użycia algorytmów klasteryzacji w zagadnieniach rozpoznawania języka mówionego. Tezy badawcze przytaczamy dosłownie za Autorem:

1. „Parametryzacja i-wektorowa umożliwia otrzymanie klasteryzacji zbioru nagrań, gdzie klastry odpowiadają językom."
2. „Klasteryzacja nagrań wewnątrz danej klasy językowej pozwala na lepsze modelowanie danych i prowadzi do polepszenia działania systemu rozpoznawania języka mówionego wzgl. pozwala na użycie prostszych klasyfikatorów liniowych przy zachowaniu skuteczności porównywalnej do skuteczności dla klasyfikatorów nieliniowych."

Treść rozprawy

Wśród siedmiu merytorycznych rozdziałów rozprawy wyróżniam dwie części: wstępną (rozdziały 1,2 i 3) i główną (pozostałe rozdziały). Obydwie są dla rozprawy ważne.

Część wstępna

Rozdział 1, nazwany „Introduction" (4 strony) zawiera zwięzłe przedstawienie sformułowania zagadnienia badawczego (1.1 „Problem Statement"), wskazanie celów i uzyskanych wyników, które Doktorant uważa za główne (1.2 „Objectives of the Research" oraz 1.3 „Research Contribution"), motywacje, a w tym okoliczności zainteresowania się problematyką, do której odnosi się recenzowana rozprawa (1.4 „Motivation"). Wstęp ten jest wyczerpujący i klarowny. Z kolei Rozdział 2 („Spoken Language Recognition") stanowi właściwe wprowadzenie do tematyki *rozpoznawania języka mówionego,* przez co rozumie się rozpoznanie (lub potwierdzenie przypuszczenia co do tego), w jakim języku wypowiedziany został określony fragment zarejestrowanej mowy. Na 20 stronach części głównej rozprawy (liczącej ogółem 90 stron) Autor dokonuje przeglądu pojęć i metod stosowanych w pracach dotyczących identyfikacji języka tekstu. W rozdziale tym Doktorant odwołuje się do 104 opublikowanych doniesień naukowych (plac lub raportów), w tym: 3-ciu prac z lat 1936 - 1963, 74 prac z lat 1974 – 2015 oraz 26 publikacji z okresu prac Doktoranta nad rozprawą doktorską, tj. 2016-2019 (a także do 4 niedatowanych tekstów internetowych). 15 wśród wymienionych publikacji odnosi się bezpośrednio do zagadnienia identyfikacji języka w okresie poprzedzającym prace Doktoranta (1974 – 2002). Obserwowany w tym przeglądzie rozkład zainteresowania się tematyką identyfikacji tekstu przez badaczy tłumaczy się początkową koncentracją prac na zagadnieniach analizy mowy. Prace te zaowocowały powstaniem warsztatu badawczego, który w dalszym ciągu pozwolił na szybki rozwój prac nad identyfikacją języka.

Rozdział 2 („Spoken Language Recognition") składa się z 6 sekcji, z których najważniejsze to sekcje 2.3 – 2.6.

Dwie pierwsze sekcje (2.3 „Approaches to Language Recognition" i 2.4 „Review of Language Recognition Research") poświęcone są odnotowaniu głównych podejść i prac z zakresu rozpoznawania języka. Na 2 stronach tekstu znalazły się odniesienia do 37 tekstów źródłowych, siłą rzeczy omówionych bardzo pobieżnie. Tym niemniej uważam sporządzone zestawienie za przydatne źródło informacji bibliograficznych do wykorzystania przez osoby zainteresowane podejmowaną w rozprawie tematyką.

Dwie kolejne **sekcje** (2.5 „I-vector Based Language Recognition System)" i 2.6 „Language Recognition in the I-vector Space") mają inny charakter. Stanowią one właściwe wprowadzenie do wykorzystywanego przez Doktoranta warsztatu badawczego, który polega na zakodowaniu informacji o obiektach, którymi są nagrane fragmenty mowy w formie tzw. reprezentacji i-wektorowej (ang. i-vector). Metody oparte na reprezentacji i-wektorowej były z powodzeniem

wykorzystane w pracach nad identyfikacją mówcy i w tym zakresie są uważane nadal za metody odpowiadające aktualnemu stanowi (*state-of-the-art*) technologii mowy (obok metod bezpośrednio wykorzystujących technologie uczenia się maszynowego, np. neuronalne). Dzięki zastosowanej (i-wektorowej) metodzie reprezentacji prób tekstu można sprowadzić zagadnienie identyfikowania klas podobieństwa tekstów (odpowiadających określonym językom naturalnym) do metod matematycznej teorii skupień (*clustering analysis*).

Sekcjom 2.5 („I-vector Based Language Recognition System") oraz 2.6 („Language Recognition in the I-vector Space") Autor poświęcił 15 stron rozprawy (str. 9 – 25). Sekcja 2.5 zawiera dokładny i starannie osadzony w literaturze przegląd technik sprowadzania zapisu audio do reprezentacji w postaci przestrzeni i-wektorowej. Z kolei sekcja 2.6 stanowi przegląd klasyfikatorów statystycznych wykorzystujących odległość kosinusową (Cosine Distance Scoring /kosinusowa ocena podobieństwa/, Generative Gaussian Classifier /klasyfikator gausowski/, Mixtures of von Mises-Fisher Distributions /mikstury (mieszniny) rozkładów von Mises-Fishera, Support Vector Machines /Maszyna Wektorów Nośnych/, Logistic Regression /Regresja logistyczna/, Probabilistic Linear Discriminant Analysis /Probabilistyczna liniowa analiza dyskryminacyjna/) oraz technik kompensacji (Length Normalisation /Normalizacja Długości/, Whithening /Wybielanie/, Within-Class Covariance Normalisation /Wewnątrz-klasowa normalizacja kowariancji/, Linear Discriminant Analysis /Liniowa Analiza Dyskryminacyjna/) wykorzystywanych w badaniach prezentowanych w rozprawie.

Niewątpliwie jest to materiał niezbędny do gruntownego zrozumienia zarówno warsztatu pracy jak i uzyskanych wyników, co jednak wymaga gruntownej znajomości wykorzystywanego aparatu matematycznego z zakresu statystyki matematycznej (włączając w to zaawansowany aparat analizy matematycznej). Autor najwyraźniej adresuje ten tekst do ekspertów z dziedziny przetwarzania mowy (speech processing), o których można założyć, że mają odpowiednie, specjalistyczne przygotowanie matematyczne, na takiej samej zasadzie jak adresowanych jest większość, jeżeli nie wszystkie, z cytowanych publikacji. Ten rozdział, będzie bardzo przydatny dla początkujących specjalistów, a w szczególności potencjalnych kontynuatorów prac z przygotowaniem jak wyżej.

Rozdział 3 („Spoken Language Clustering") ma podobny charakter jak rozdział poprzedni, wprowadza mianowicie najważniejsze pojęcia i algorytmy klasteryzacji języków. Rozdział ten składa się 5 sekcji. Trzy pierwsze poświęcone są zdefiniowaniu problemu klasteryzacji języków (ang. *language clustering*) jako analogonu zagadnienia klasteryzacji mówców (ang. *speaker clustering*) i wskazanie powodów, dla których klasteryzacja języków jest trudniejsza niż

analogiczny problem identyfikacji mówców. Doktorant zwraca uwagę (str. 28), że klasteryzacja języków okazuje się być trudniejsza niż rozpoznawanie mówcy przy zastosowaniu analogicznej metody wykorzystującej reprezentację i-wektorową segmentów mowy (uzyskiwaną metodą uczenia nienadzorowanego, por. Rozdział 2) (metoda wypracowana na potrzeby rozpoznawania mówców w serii challengów amerkańskiego National Institute for Standards and Technology). Bierze się to z konieczności abstrahowania od specyficznych cech mówcy.

Główną częścią tego rozdziału są sekcja 3.4 („Clustering Algorithms") poświęcona przeglądowi wybranych algorytmów oraz sekcja 3.5 (Evaluation of Clustering) stanowiąca przegląd wybranych metod oceny klasteryzacji.

Pięć omówionych w sekcji 3.4 algorytmów klasteryzacji to algorytmy wykorzystujące pojęcia odległości kosinusowej i centroidu (sferyczny algorytm centridów /spherical k-means/, mean shift, klasteryzacja hierarchiczna /aglomerative hierarchical clustering/ oraz HDBSCAN (Hierarchical Density-based Spatial Clustering of Applications with Noise). Będą one wykorzystywane w części głównej, eksperymentalnej. Z kolei w sekcji 3.5 omówione są miary jakości klasteryzcji („clustering quality measures"), także wykorzystane w części eksperymentalnej (Rozdział 5 „Language Clustering Experiments").

Część główna

Rozdziały 4, 5 i 6 składające się na część główną rozprawy dotyczą przeprowadzonych przez Doktoranta eksperymentów w zakresie identyfikacji języka tekstu (nagrania) z wykorzystaniem narzędzi przedstawionych w rozdziałach 1, 2 oraz 3.

Rozdział 4 („Language Data Description and Analysis") jest krótki (3.5 strony), lecz bardzo ważny bo zawiera ogólne przedstawienie danych empirycznych wykorzystanych w pracy. Dane te mają postać zbioru i-wektorów pochodzących z bazy „NIST 2015 Language Recognition I-vector Machine Learning Challenge Database" i ich znajomość jest istotna do zrozumienia wyników pracy. Udostępniony przez NIST zbiór i-wektorów odpowiadający próbkom audio zawiera ok 28000 elementów dla 65 języków, z czego w eksperymentach opisanych w Rozdziale 5 zostało użytych 15000 i-wektorów stanowiących dane treningowe dla testowanych w Challenge'u algorytmów klastrujących dla łącznie 50 języków (300 i-wektorów na język). Doktorant zwraca uwagę na fakt, że wybór danych do prowadzenia eksperymentów opisanych w dalszej części pracy był podyktowany umożliwieniem porównywalności otrzymanych wyników z wynikami innymi opisanymi w literaturze (pozycje [8] – [11] w załączonej bibliografii, patrz też str. 87). W części poświęconej analizie danych w eksperymentach

challenge'u NIST Doktorant skupia się na analizie obrysu (silhouette) (jako miary spójności klastrów wizualizacji) oraz wizualizacji.

Niestety opis danych zawiera utrudnienia polegające na stosowaniu pojęć niezdefiniowanych (jak np. *development set*) należących do „folkloru terminologicznego dyscypliny", lecz których znaczenie może nie być oczywiste dla sporej części czytelników..

Innym powodem, dla którego rozdział czyta się trudno są niekonsekwencje terminologiczne w opisie danych. Powodem trudności jest np. wieloznaczne użycie określenia „dane ewaluacyjne", które to określenie jest w jednym miejscu równoznaczne z określeniem „dane testowe", a w innym oznacza podzbiór danych testowych (por. *„The data was split into training, development and evaluation subsets"* (str. 37, 4 wiersz od dołu) oraz *„The test set contains (...). This set was split into progress set and evaluation set.",* str. 38, wiersze 4 – 6 od góry).

W Rozdziale 5 („Language Clustering Experiments") przedstawione zostały eksperymenty przeprowadzone przez Doktoranta w zakresie klasteryzacji danych reprezentowanych w przestrzeni i-wektorów, o których jest mowa w rozdziale poprzednim oraz ich ocena. Używane miary oceny wyników eksperymentów klasteryzacji to: miara nieczystości klastrów /cluster impurity measure, miara dyspersji języka (rozrzut języka na różne klastry) /language impurity measure (str 34.)/, zmodyfikowany Indeks Randa /Adjusted Rand Index/, skuteczność klasteryzacji /clustering efficiency (BBN)/, estymator czystości otoczenia klastra /Nearest Neighbor Purity Estimator/, współczynnik obrysu /Silhouette coefficient/. Celem przeprowadzonych eksperymentów było „uzyskanie klastrów możliwie najlepiej odpowiadających językom reprezentowanym w zbiorze danych („Our goal is to obtain clusters that would correspond to languages as close as possible.", str. 45).

Doktorant przeprowadził eksperymenty na materiale empirycznym wykorzystywanym w projekcie „NIST 2015 Language Recognition i-Vector Machine Learning Challenge" i udostępnionym przez National Institute of Standards and Technology (NIST). Na ten materiał składają się próby nagrań audio z 50 języków (przy czym na jeden język przypada około 300 i-wektorów) należące do zbioru treningowego („training set") (str. 37), charakteryzowanego tym, że i-wektory są etykietowane językiem i długością nagrania. Spośród 15000 i-wektorów tego zbioru do eksperymentów klastrowania Doktorant wybrał 5811 odpowiadających segmentom więcej niż 30 sekundowym. Powodem tej decyzji była obserwacja, że wielkość próbki audio ma wpływ na jakość rozpoznania („As i-vectors are calculated from utterance statistics, the longer the utterance, the more reliable the i-vector is", str. 39).

Analiza eksperymentów opisanych w rozdziale 5 wykazała, że spośród rozważanych algorytmów Mean Shift Clustering zwraca najlepszą klastryzację charakteryzującą się najmniejszym zanieczyszczeniem klastrów (niska wartość parametru *cluster impurity*) przy zbliżonej liczbie klastrów i rozróżnionych języków, oraz stosunkowo niskiej wartości parametru dyspersji językowej *(language impurity).* **Oznacza to uzyskanie stosunkowo prawidłowego (*relatively pure*) określenia języka próbki mowy na podstawie jej przynależności do klastra, a tym samym osiągnięcie jednego z założonych celów rozprawy.**

Ważną obserwacją jest zwrócenie uwagi na to, że dla najlepszej (przy zastosowaniu przyjętych metod ewaluacji) uzyskanej klastryzacji, dla niektórych z rozważanych języków, próby znalazły się w dwóch lub trzech klastrach. Obserwacja ta jest punktem wyjścia do zaproponowania dalszych modyfikacji systemów rozpoznawania języka tekstu.

Rozdział 6 („Language Recognition with Clustering-based Modeling Experiments") składa się z 3 sekcji.

Sekcja 6.1 („Cluster-based Modeling") zawiera opis nowatorskiej koncepcji systemów rozpoznawania języka biorącej za punkt wyjścia wyniki eksperymentów klastryzacji opisane w rozdziale poprzednim, a mianowicie zaobserwowanie dodatniej dyspersji dla niektórych języków przy niskim zanieczyszczeniu klastrów. Doktorant proponuje przeprowadzenie eksperymentów klastryzacji dla poszczególnych języków, w wyniku czego uzyskuje się modele tych języków. Przy ich pomocy można ustalić język nieznanego nagrania audio przez przyrównanie odpowiedniego i-wektora do modeli poszczególnych języków (por. rys. 6.1 na str. 71). Doktorant zwraca uwagę na zalety tej metody polegającej na możliwości stosowania liniowych klasyfikatorów do poszczególnych języków. Metoda prowadzi do wykrycia podklas i-wektorów odpowiadających różnym podmodelom dla danego języka.

W sekcji 6.2 (Language Recognition Experiments and Results) Doktoranat przeprowadza porównania działania różnych systemów opisanych w literaturze. W szczególności pokazuje on, że jego prosty klasyfikator wykorzystujący kosinusową miarę podobieństwa (CDS) jest konkurencyjny w stosunku do nieliniowych klasyfikatorów opracowanych w ramach NIST Challenge 2015 (tab. 6.12, str. 87). Porównanie to było możliwe dzięki stosowaniu w eksperymentach tego samego zestawu danych (i-wektorów) testowych. Do porównania zastosowana została funkcja kosztu (NIST Cost Function) mierząca poziom błędów klastryzacji (opisana w sekcji 6.2.1.1, str. 72). W przedstawionych porównaniach najlepiej wypada system Fusion („fusion of logistic regressions and CDS with cluster-based modeling") przedstawiony w rozprawie (por. tab. 6.12, str. 87).

W sekcji 6.3 („Results Analysis and Discussion), obok interesującej analizy dotyczącej rozpatrywanych w rozprawie języków i refleksji statystycznej, na uwagę zasługuje analiza złożoności obliczeniowej systemów działających w oparciu o metodę Cluster-based Modeling. Autor wykazuje, że algorytmy zaproponowane w tej rozprawie mają liniową złożoność czasową, a dodatkowo architektura systemu umożliwia paralelizację istotnej części obliczeń.

Dodajmy, że klastery wyodrębnione dla określonego języka mogą mieć w niektórych przypadkach interpretację lingwistyczną (np. może okazać się, że klastry odpowiadają różnicom dialektalnym lub socjo-lingwistycznym, jak np. rejestr).

Obserwacje redakcyjne

Tekst jest napisany jest poprawną i zrozumiałą angielszczyzną, a usterki jak w zdaniu *„To be able to indicate the best clustering algorithm for the language clustering task it is necessary to assess their performance”* (str. 33) są wyjątkiem. Nie czując się kompetentnym do oceniania poprawności językowej, wstrzymuję się od sygnalizowania przypadków, które wydają mi się wątpliwe.

Układ treściowy rozprawy także nie budzi moich zastrzeżeń, może poza podrozdziałem 2.4 („Review of Language Recognition Research”), który szokuje zagęszczeniem przytoczonej literatury w stosunku do tekstu omawiającego poszczególne pozycje (50 pozycji na 1 stronie tekstu).

Silne strony rozprawy

1.

Niewątpliwie najważniejszą (i najciekawszą) częścią rozprawy jest Rozdział 6, w którym Autor wyciąga wnioski z badań eksperymentalnych omówionych w rozdziałach poprzedzających. Wnioski te zamykają się w propozycji nowego, w stosunku do dotychczasowych, podejścia do zagadnienia rozpoznawania języka określonego zapisu audio, interesującego zarówno pod względem jakości rozpoznawania jak też czasowej złożoności obliczeniowej.

Jednocześnie uważam, że Doktorant nie do końca wyeksploatował obserwację, iż najlepsze z dotychczasowych algorytmów rozpoznawania języka zwracają klastry o stosunkowo wysokiej czystości przy niezłej przeciętnej dyspersji językowej, z tym że dyspersja dla pewnych języków może być podwyższona, a dla innych bliska idealnej (czyli zerowej). Oznacza to, że gdyby ograniczyć zastosowanie algorytmu klastrującego do jednego z języków, dla których dyspersja jest znaczna, to naturalną interpretacją musiałoby być stwierdzenie, że

grupowanie się i-wektorów zawiera informację o języku (o ile o zbiorze analizowanych nagrań można założyć że jest reprezentatywny dla języka). Prowadzi to do postawienia problemu badawczego (wykraczającego poza wąsko rozumianą informatykę), którym jest pytanie o interpretację lingwistyczną zaobserwowanych skupień i-wektorów (klastrów). Nie jest wykluczone, że prowadzone w tym kierunku prace doprowadziłyby do wykrycia niezidentyfikowanych do tej pory zjawisk i prawidłowości wewnątrzjęzykowych czy socjolingwistycznych. Dodać należy, że w postulowanych badaniach można z powodzeniem wykorzystać dane z Challenge'u NIST 2015 (jako, że można je traktować za reprezentatywną, i tym samym doskonale nadającą się do porównawczych badań lingwistycznych, próbę dla 50 języków).

Reasumując, wyniki uzyskane przez Doktoranta zdają się mieć potencjał istotnie wykraczający poza problematykę informatyczną.

2.

Poza walorami o charakterze naukowym, rozprawa stanowi ważne źródło wiedzy, które może być cenną pomocą dla osób zainteresowanych tematyką rozprawy i zamierzających ją zgłębić. Sam tylko Rozdział 2 mógłby stanowić podstawę opisu rocznego wykładu monograficznego pt. „Metody statystyczne rozpoznawania języka", a tym samym narzędziem dający pogląd na ten zasób wiedzy statystycznej, którą adept tej dyscypliny winien dysponować.

3.

Cennym źródłem informacji jest wyjątkowo obszerna, licząca 189 pozycji bibliografia zamieszczona w rozprawie.

Słabe strony rozprawy

1.

Przyjęta dla rozprawy metoda prezentacji typu konferencyjnego stwarza, że praca nie jest informacyjnie samowystarczalna, a w szczególności wymaga znacznie większej wiedzy matematycznej niż można oczekiwać od wielu zainteresowanych dziedziną. Jest to poważny defekt z punktu widzenia roli rozpraw doktorskich: doktorant winien nie tylko zdać sprawę ze swoich kompetencji badawczych poprzez uzyskane wyniki, lecz także przedstawić wyniki swoich prac w sposób dostępny i przekonywujący specjalistów z dziedzin pokrewnych, a także przydatny dla adeptów dyscypliny.

Przykładem jest brak w rozprawie jasnych i wyraźnych (w terminologii Kartezjusza) definicji kluczowych dla pracy pojęć, nie mówiąc już o ilustracji trudnych zagadnień przykładami. W szczególności Autor nie zdobył się na podanie definicji tytułowego pojęcia *i-vector space* mimo, że termin i-vector pojawia się w pracy ponad 211 razy (nie licząc bibliografii), brak jest też jest jakichkolwiek przykładów ilustrujących to pojęcie w poświęconym mu rozdziale 2.5 (I-Vector Based Language Recognition System). Trudno w tym przypadku dopatrywać się waloru publikacji samodzielnej (w rozumieniu ang. *self-contained*).

2.

Istotnym brakiem przedstawionej rozprawy, nb. pokutującym w niektórych środowiskach naukowych, jest brak indeksu pojęć, który to brak jest dotkliwy w przypadku, kiedy czytelnik (np. recenzent) nie ma do dyspozycji przeszukiwalnej wersji elektronicznej, podczas gdy praca jest obszerna i nasycona terminologią. Świadczyć to może o tym, iż autor nie zakłada, że dzieło może być czytane w wersji innej niż elektroniczna. Brak ten dziwi o tyle, że rozprawa zawiera „pomoce" o znikomej przydatności, jak np. sporządzone automatycznie lista ilustracji czy tabel.

3.

Rozprawa zredagowana zastała w języku angielskim, co jest dopuszczalne przez obowiązujące przepisy. Tym niemniej, fakt redagowania rozprawy wyłącznie w języku angielskim uważam za jej słabą stronę, zdając sobie sprawę, że w niektórych środowiskach naukowych ten pogląd będzie odosobniony. Jednak uważam, że wartościowa naukowo praca doktorska winna pełnić także funkcję popularyzatorską w środowisku naukowym, w pierwszym rzędzie rodzimym, i zachęcać (w pierwszym rzędzie) początkujących naukowców do zainteresowania się uzyskanymi wynikami. Istotną barierą jest stosowanie specjalistycznej terminologii niemającej odpowiedników w języku, w którym adept zapoznawał się z podstawami dziedziny na poziomie akademickim. W kraju, gdzie studia w języku angielskim nie stanowią standardu, a takim jest Polska, należy dbać o udostępnianie wiedzy wysokospecjalistycznej także w języku, w którym prowadzone są studia wyższe. Warunkiem minimum, równie ważnym jak definiowanie pojęć technicznych (co do których zachodzi przypuszczenie, że mogą być nieznane osobom zainteresowanym) będzie wprowadzenie polskich odpowiedników dla terminologii specjalistycznej tam, gdzie istnieją. W przypadku wprowadzania nowych pojęć, obok zaproponowania nowego terminu w języku rozprawy, winien zostać zaproponowany odpowiednik polski. Rozwiązaniem optymalnym, z punktu widzenia zasygnalizowanych

przeze mnie potrzeb, byłoby redakcja rozprawy jednocześnie w dwóch wersjach językowych (ojczystej /tu polskiej/ i angielskiej).

Pragnę przy tym zaznaczyć, że wskazane powyżej „słabe strony" nie umniejszają wysokiej oceny wyników naukowych opisanych w rozprawie i nie rzutują na zamieszczoną poniżej konkluzję.

Konkluzja

Biorąc pod uwagę sumę poczynionych obserwacji stwierdzam, że przedłożona rozprawa spełnia warunki stawiane przez Ustawę o Stopniach i Tytułach Naukowych przed rozprawami doktorskimi i wnioskuję o jej przyjęcie oraz o dopuszczenie jej autora do dalszych etapów przewodu doktorskiego.

Prof. zw. dr hab. Zygmunt Vetulani