

Recenzja rozprawy doktorskiej

pt. „Towards Discriminative Speaker Representations for Speaker Recognition and Diarization (W stronę różnicujących reprezentacji mówców na potrzeby ich rozpoznawania i diaryzacji)”

Pracę złożyła Magdalena Marta Rybicka z Akademii Górniczo-Hutniczej w Krakowie

W pracy badane są neuronowe reprezentacje mówców w systemach automatycznego rozpoznawania mówców (SR) i diaryzacji mówców (SD), choć dotyczy ona również separacji mówców. Tematyka ta mieści się w zakresie automatycznego przetwarzania mowy, które samo w sobie uznaje się za element sztucznej inteligencji. Badane tematy i wyciągnięte wnioski są aktualne oraz istotne dla społeczności zajmującej się badaniami i rozwojem zarówno na uczelniach, jak i w branży.

Praca składa się z 4 rozdziałów, zawiera 131 stron i opiera się na pięciu publikacjach autorki - trzech referatach z konferencji Interspeech (CORE A) i dwóch referatach opublikowanych w ważnych czasopismach dla społeczności zajmującej się przetwarzaniem mowy - IEEE Transactions on Audio, Speech and Language Processing (T-ASL) i IEEE Signal Processing Letters (SPL). Niniejsza recenzja w pierwszej kolejności skupia się na technicznej zawartości rozprawy, a następnie podsumowuje jej jakość techniczną, zawiera uwagi dotyczące kwestii formalnych, a na koniec przedstawia ogólne wnioski i zalecenia dla komisji.

Zawartość techniczna rozprawy oraz uwagi do rozdziałów i referatów

Rozdział 1 stanowi wprowadzenie do rozprawy, przedstawia ogólny przegląd problematyki i określa główne zadania. Przedstawia również główne osiągnięcia i strukturę rozprawy, a także podsumowanie prac nieobjętych pięcioma głównymi publikacjami. Rozdział ten dokumentuje szeroko zakrojoną współpracę doktorantki z Johns Hopkins University (Center of Language and Speech Processing - CLSP oraz Center of Excellence - CoE) – czołowym amerykańskim laboratorium akademickim zajmującym się przetwarzaniem mowy i języka naturalnego. Drobne uwagi do tego rozdziału: w pkt. 1.3 (Wkład), dobrze byłoby podać odniesienia do odpowiednich fragmentów rozprawy i referatów, a z akapitu dotyczącego projektów chciałbym się dowiedzieć, kto był głównym badaczem w przypadku każdego z nich.

Rozdział 2 „Tło badań” to poważny tekst przedstawiający zadania, aktualny stan wiedzy, dane i wskaźniki oceny. Rozdział ma szeroki zakres i dokumentuje doskonałą orientację doktorantki w dziedzinie. Daje to solidną podstawę do zrozumienia reszty rozprawy. Szczególnie doceniłem kompleksowy przegląd kątowych metod obliczania funkcji celu. Drobne uwagi:

1. Struktura rozdziału - moim zdaniem lepiej byłoby przedstawić przegląd **wszystkich** zadań na początku rozdziału i podać ich wskaźniki oceny, a następnie zająć się tylko aktualnym stanem wiedzy – pozwoliłoby to czytelnikowi lepiej uporządkować strukturę nabywanej wiedzy. Ponadto rozdział ten jest dość długi, radziłbym podzielić go na trzy krótsze rozdziały: 2. Zadania i wskaźniki, 3. Aktualny stan wiedzy i 4. Dane.
2. W niektórych miejscach, zwłaszcza w opisie „starszych” technik, takich jak i-wektory, wolałbym widzieć większą dyscyplinę w zastosowaniu terminów matematycznych (angielskie słowa „probability” i „likelihood” na określenie prawdopodobieństwa, macierze i wektory tłustym drukiem). W dalszych częściach dotyczących architektury neuronowej notacja jest bezbłędna.
3. Rozdział z danymi można by uzupełnić tabelami podsumowującymi, podając cel poszczególnych zbiorów danych, szczegóły techniczne i odniesienia, odpowiednio: cytowania rozdziałów/referatów, w których zostały one wykorzystane.

Rozdział 3 omawia wkład doktorantki i przedstawia pięć głównych referatów: pkt. 3.1 oraz referaty I i II dotyczą różnicujących reprezentacji mówcy, głównie struktury i szkolenia architektur neuronowych do ich ekstrakcji. Referat I skupia się na dostosowaniu parametrów obliczania kątovej funkcji straty - zadanie jest dobrze wyjaśnione, zaproponowano solidną hipotezę, eksperymenty zostały dobrze przeprowadzone, a rozwiązanie działa lepiej niż rozwiązanie podstawowe. W drugiej części referatu I badane są zmiany w strukturze sieci

neuronowych NN - łączenie bloków sieci neuronowych z opóźnieniem czasowym TDNN i ResNet. Chociaż połączona architektura działa lepiej, wolalbym zobaczyć głębszy wgląd w to, jaki jest główny problem prostych architektur i jak rozwiązują go proponowane zmiany, zamiast prezentacji tylko zmian strukturalnych i wyników. W tym względzie referat II dotyczący zastąpienia ResNet przez SpineNet i wprowadzenia przetwarzania czasowego (Squeeze and Excitation Blocks) do łańcucha przetwarzania jest znacznie lepszy, a tok rozumowania jest bardzo klarowny.

W pkt. 3.2 oraz referatach III i IV proponowane jest wzmocnienie reprezentacji mówców w neuronowej diaryzacji mówców w systemach *end-to-end* (EEND), a autorka odnosi się głównie do etapu grupowania osadzania. Wprowadzenie atraktorów nieautoregresywnych (NAA) zdefiniowanych w referacie III stanowi bardzo udany element badań - zaproponowano kilka systemów, przeprowadzono ewaluację z pozytywnymi wynikami, a pracą zainteresowała się społeczność badaczy (referat III ma 19 cytowań w Google Scholar). Publikowany w czasopiśmie artykuł – referat IV stanowi rozbudowę referatu z konferencji i wprowadza kilka innowacji oraz szerszą ocenę eksperymentalną (w tym analizę wpływu komponentów na rezultat, tzw. *ablation study*). Szczególnie doceniam wprowadzenie bloku wykrywania aktywności pojedynczego mówcy – *Single Speaker Activity Detection* (SSAD), znacząco poprawiającego wyniki wszystkich zawierających go schematów. Jednak chętnie bym zobaczył bardziej szczegółowe i intuicyjne wyjaśnienie tego, co doprowadziło do wprowadzenia kilku wariantów (od EEND-NAA-Overest do EEND-NAA-1step), tj. co było nie tak z prostymi podejściami i jak naprawiły to kolejne warianty. Ponadto w przypadku iteracyjnego wyszukiwania atraktorów (referat III), dobrze byłoby skomentować wymagania obliczeniowe tego podejścia, ponieważ wydaje się, że iteracje odbywają się w czasie wnioskowania, oraz określić, jak ustalono optymalną liczbę iteracji.

Wreszcie, w pkt. 3.3 i w referacie V jest mowa o wspólnym zadaniu z wykorzystaniem diaryzacji i separacji mowy, co stanowi naturalne rozwinięcie wcześniejszych prac z dziedziny przetwarzania sygnałów w oparciu o dobre reprezentacje mówców. Ta generalizacja zadziałała bardzo dobrze, a eksperymenty udokumentowały wyższość proponowanych atraktorów również na potrzeby separacji mowy. Niewielką niedogodnością jest konieczność wcześniejszego ustalenia, czy w sygnale źródłowym nie mamy do czynienia z nakładającymi się głosami wielu mówców (atraktory klasteryzacji działają lepiej), czy mamy z tym do czynienia (atraktory diaryzacji wykazują wyższość); na potrzeby dalszych badań ciekawe by było, gdyby został zaproponowany jeden schemat działający w różnych scenariuszach.

Rozdział 4 to standardowe podsumowanie i informacje o pracy planowanej w przyszłości - mogę zgodzić się ze wszystkimi wymienionymi punktami (dobre plany inżynierskie), ale brakuje mi trochę szerszego obrazu, z bardziej „odległymi” celami, a także uwag na temat obecnych trendów w przetwarzaniu mowy (wstępnie wytrenowane modele SSL i modele mowy z użyciem dużych modeli językowych LLM).

Podsumowanie w odniesieniu do treści technicznej rozprawy

Rozprawa wyraźnie pokazuje cechy doktorantki - zdolność do studiowania ważnej literatury z kilku dziedzin, proponowania własnych nowatorskich rozwiązań, ich wdrażania, starannego testowania i omawiania wyników eksperymentów. Pokazuje szeroki przegląd zarówno mowy, jak i przetwarzania sygnałów, uczenia maszynowego i nowoczesnej architektury neuronowej. Eksperymenty są dobrze przeprowadzone, dobrze udokumentowane, dobrze przeanalizowane i komentarze do nich są odpowiednie. Niektóre ustalenia (np. atraktory nieautoregresywne NAA i kombinacje na potrzeby diaryzacji i separacji mowy) już wywarły wpływ na szeroką społeczność badawczą lub mają duży potencjał, aby mieć wpływ w najbliższej przyszłości.

Uwagi dotyczące kwestii formalnych

Chociaż ogólnie preferuję rozprawy „linearne” od takich, które składają się z kilku publikacji, to ta rozprawa mi się podobała. Autorka starannie wybrała ograniczoną liczbę najistotniejszych publikacji, rozdział 3 wprowadzający treść tych publikacji był wystarczająco „lekki”, a treść referatów się nie pokrywała w znacznym stopniu. Struktura pracy (z wyjątkiem długiego rozdziału 2) jest dobra, a autorka dołożyła wszelkich starań, aby przedstawić wykonane prace w logicznej kolejności. Rozprawa jest napisana bezbłędnie w języku angielskim, występuje w niej bardzo niewiele błędów stylistycznych. Z przyjemnością przekażę autorce wersję dokumentu z komentarzami.

Prezentacja wyników jest doskonałej jakości, tabele i rysunki są czytelne, dobrze opisane i dostarczają prostych

informacji na temat tego, co osiągnięto. Zapis matematyczny (poza drobnymi wyjątkami) jest przejrzysty i spójny. Wspaniale została wykorzystana literatura, wykazy piśmiennictwa w rozprawie i referatach są obszerne i dokumentują szeroki zakres wiedzy doktorantki. Sugerowałbym tylko alfanumeryczny styl odniesień w pracy - [Watanabe2020] jest czytelniejsze niż [193].

Podsumowanie i zalecenia

Z uwagą zapoznałem się z rozprawą doktorską Pani Magdaleny Marty Rybickiej. Pomimo drobnych uwag krytycznych podniesionych powyżej, moim zdaniem jest to solidna praca, która wnosi wkład w postępy w badaniach nad neuronową reprezentacją mówcy. Przeanalizowałem również dorobek publikacyjny doktorantki i stwierdzam, że spełnia on standardy do uzyskania stopnia doktora na uznanej uczelni.

Stwierdzam, że w przedłożonej do recenzji rozprawie doktorskiej Pani Magdaleny Marty Rybickiej pt. „Towards Discriminative Speaker Representations for Speaker Recognition and Diarization (W stronę różnicujących reprezentacji mówców na potrzeby ich rozpoznawania i diaryzacji)” doktorantka wykazała się wiedzą z zakresu przetwarzania mowy, architektur neuronowych oraz uczenia maszynowego. Uważam, że oryginalny wkład w dziedzinę nauki opisaną w rozprawie doktorskiej, a także dorobek publikacyjny są znaczące. W mojej ocenie przedstawiona praca spełnia wymogi stawiane rozprawom doktorskim w obowiązującej ustawie o stopniach naukowych i tytule naukowym. W związku z powyższym wnoszę o dopuszczenie przez Radę Dyscyplinarną Pani Magdaleny Marty Rybickiej do kolejnych etapów postępowania o nadanie stopnia doktora.

Biorąc pod uwagę jakość techniczną rozprawy, jakość eksperymentów przeprowadzonych na otwartych zbiorach danych, ważność wyników i ich znaczenie dla międzynarodowej społeczności naukowej, a także międzynarodową współpracę z prestiżowym amerykańskim laboratorium mowy/NLP, sugeruję, aby nagrodzić tę rozprawę wyróżnieniem.

Proponuję zadać następujące pytania podczas obrony pracy doktorskiej:

1. Proszę wyjaśnić rozumowanie i intuicyjne myślenie stojące za połączeniem bloków schematu TDNN i ResNet w pkt. 3.1 i referacie I.
2. Proszę podać intuicyjne wyjaśnienie, co doprowadziło do wprowadzenia kilku wariantów (od EEND-NAA-Overest do EEND-NAA-1step) w pkt. 3.2 i referacie IV - tj. co było nie tak z prostymi podejściami i jak te warianty to naprawiały.
3. W kilku miejscach w rozprawie i w referacie IV stwierdza Pani, że kodowanie pozycyjne nie było używane - proszę podać przyczynę.
4. Schemat separacji/diaryzacji w pkt. 3.3 i referacie V jest szkolony z użyciem funkcji celu łączącej kilka kryteriów. Czy uważa Pan, że byłoby wykonalne trenowanie tylko z zastosowaniem kryterium separacji mowy, które powinno zawierać wszystkie informacje na temat diaryzacji (tj. gdy mówca nie jest aktywny, jego sygnał powinien wynosić zero)?

Brno, 14 stycznia 2026 r.

[podpis]

Prof. dr Jan "Honza" Černocký

Kierownik Katedry Grafiki Komputerowej i Multimediów

Odpowiedzialny za grupę BUT Speech@FIT

Wydział Informatyki, Politechnika w Brnie

Bozetěchova 2, 612 66 Brno, Republika Czeska

Tel: +420 5 41141284 Tel. kom.: +420 604738324,

<mailto:cernockv@fit.vutbr.cz> <http://www.fit.vutbr.cz/~cernocky>

<http://www.fit.vutbr.cz/> <https://cs-cz.facebook.com/FIT.VUT>

<http://speech.fit.vutbr.cz> <https://www.facebook.com/BUT-Speech/>

