Kraków, 11 marca 2024 r.

dr hab. inż. Konrad Kowalczyk, prof. uczelni
Wydział Informatyki, Elektroniki i Telekomunikacji
Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie
email: konrad.kowalczyk@agh.edu.pl

RECENZJA

Rozprawy doktorskiej mgr. inż. Mikołaja Pudo pt.:
„Methods of Optimizing Models and Algorithms for Automatic Speech Recognition in Mobile Applications" (tytuł w języku polskim: „Metody optymalizacji modeli i algorytmów automatycznego rozpoznawania mowy pod kątem działania na urządzeniach mobilnych")

Promotor rozprawy: dr hab. inż. Artur Janicki, prof. uczelni

Zagadnienie naukowe, główne cele i tezy pracy doktorskiej

Przedstawiona do recenzji praca doktorska mgr. inż. Mikołaja Pudo pt. „Methods of Optimizing Models and Algorithms for Automatic Speech Recognition in Mobile Applications" dotyczy zastosowania technik głębokiego uczenia maszynowego w automatycznym rozpoznawaniu mowy. W szczególności, rozprawa doktorska poświęcona jest automatycznemu wykrywaniu fraz kluczowych w sygnale akustycznym (ang. keyword spotting), powiązanym z tym tematem wyznaczaniu końca frazy (ang. end-of-speech detection) oraz adaptacji systemów rozpoznawania mowy do domeny docelowej użytkownika. W celu osiągnięcia pożądanej dokładności rozpoznawania fraz kluczowych, zastosowany został prosty model językowy o wskazanej w pracy inicjalizacji. Dodatkowo zastosowano wagowanie funkcji kosztu (straty) w istniejącej w literaturze metodzie wykrywania końca komendy. Ponadto opracowano zbiór danych przeznaczony do ewaluacji metod wykrywania fraz

kluczowych, który powstał w oparciu o popularne korpusy mowy o otwartym dostępnie. W końcowej części, rozprawa porusza zagadnienie uczenia częściowo nadzorowanego (ang. semi-supervised learning) wstępnie wytrenowanego modelu automatycznego rozpoznawania mowy. Główną wskazaną w rozprawie doktorskiej motywacją przeprowadzonych badań jest przydatność poruszanych zagadnień w kontekście stosowania inteligentnych asystentów głosowych na urządzeniach mobilnych.

Zasadniczym celem pracy doktorskiej jest automatyczne rozpoznawanie fraz kluczowych, w tym wyznaczenie końca komendy oraz dostosowanie modeli automatycznego rozpoznawania mowy do domeny docelowej użytkownika. Doktorant skupił się na opracowaniu bazy danych umożliwiającej ewaluację metod wykrywania fraz kluczowych, dodaniu niewielkiego modelu językowego z zaproponowaną inicjalizacją, zastosowaniu dodatkowego wagowania w standardowej funkcji kosztu dla istniejącego modelu detekcji końca frazy oraz stosowaniu technik uczenia częściowo nadzorowanego w dostrajaniu modelu automatycznego rozpoznawania mowy. W rozprawie doktorskiej postawione zostały trzy główne tezy badawcze:

1. Wydajność i dokładność modelu wykrywania słów kluczowych można znacznie poprawić, wykorzystując unigramowy model języka i nagrania audio generowane przez system zamieniający tekst na mowę.
2. Dokładność modelu wykrywania końca mowy można skutecznie poprawić, trenując model z proponowaną funkcją kosztu (ang. loss function).
3. Częściowo nadzorowane metody uczenia się mogą być skutecznie wykorzystywane do adaptacji modeli akustycznych nawet w przypadku małych zbiorów danych.

Zakres rozprawy

Recenzowana praca doktorska została napisana w języku angielskim, składa się z 7 rozdziałów głównych i dwóch załączników, obejmujących w sumie 82 strony. Bibliografia zawiera 118 źródeł, listę skrótów występujących w pracy jak również listę 10 tabel i 30 rysunków. Praca poprzedzona jest dwustronicowym streszczeniem w języku polskim i analogicznym streszczeniem w języku angielskim.

Rozdział 1 (ang. Introduction) wprowadza czytelnika w tematykę przetwarzania mowy i jej zastosowania w asystentach głosowych. Następnie po krótce omówione zostają główne składowe przedstawionego modułu automatycznego rozpoznawania mowy tj. rozpoznawanie komend, detekcja granic frazy, poprawa jakości sygnału, modele akustyczne i językowe, a także zagadnienia związane z przetwarzaniem mowy na urządzeniu mobilnym lub w chmurze.

Rozdział 2 (ang. Literature Review) poświęcony jest przeglądowi stanu wiedzy w zakresie trzech głównych tematów poruszanych w pracy doktorskiej, czyli technik rozpoznawania fraz kluczowych (komend) wraz z przeglądem ogólnodostępnych baz danych do ich ewaluacji, detekcji końca frazy (ang. end-of-speech detection) oraz częściowo nadzorowanego uczenia maszynowego na potrzeby rozpoznawania mowy.

Rozdział 3 (ang. Contribution of this Thesis) przedstawia trzy główne hipotezy badawcze, krótko opisuje główne osiągnięcia pracy doktorskiej oraz wskazuje publikacje konferencyjne tematycznie powiązane z każdą z hipotez.

Rozdział 4 (ang. Custom Keyword Spotting) opisuje dwa główne osiągnięcia Doktoranta w zakresie wykrywania fraz kluczowych. W pierwszej części tego rozdziału przedstawiona została stworzona w ramach pracy doktorskiej baza danych do ewaluacji metod rozpoznawania fraz kluczowych. Po omówieniu głównych wymagań do takiej bazy, którymi Doktorant kierował się przy jej tworzeniu, omówione zostały dane i metadane bazy MOCKS oraz sposób jej utworzenia z danych pochodzących z dwóch popularnych korpusów mowy LibriSpeech i Mozilla Common Voice. Ponadto porównano błędy detekcji dla opracowanej bazy MOCKS dla języka angielskiego z błędami otrzymanymi dla popularnej bazy Google Speech Commands. W drugiej części rozdziału 4, omówiono pochodzący z literatury model akustyczny do wykrywania fraz kluczowych, proponując zwiększenie prawdopodobieństwa rozpoznania fraz kluczowych poprzez zastosowanie dodatkowego prostego unigramowego modelu językowego oraz jego odpowiednią, zaproponowaną przez Doktoranta inicjalizację. Inicjalizacja polega na nadaniu wartości hiperparametrowi, który powinien przyjąć wartość większą niż jeden. Optymalna wartość hiperparametru wyznaczona została na podstawie przeprowadzonych eksperymentów, które opisane zostały w dalszej części rozdziału.

Rozdział 5 (ang. End-of-Speech Detection) poświęcony jest opisowi znanego z literatury modelu wykrywania końca frazy, który Doktorant stosuje w swojej pracy doktorskiej. Modyfikacją zaproponowaną przez Doktoranta jest dodanie do klasycznej w tym rozwiązaniu funkcji kosztu (ang. loss function) opartej o binarną entropię krzyżową (ang. binary cross entropy) wagi wpływającej na agresywność detekcji końca frazy, w zależności odległości danej ramki sygnału dźwiękowego od faktycznego końca frazy. Przedstawiono też wyniki przeprowadzonych eksperymentów z różną długością sekwencji wejściowej oraz z różnymi wartościami wagi.

Rozdział 6 (ang. Semi-supervised Learning for Speech Recognition) poświęcony jest w całości tematyce adaptacji (inaczej dostrajania) pre-trenowanego modelu rozpoznawania mowy do nowej domeny używając metod uczenia częściowo nadzorowanego (ang. semi-supervised

learning). Przeprowadzone eksperymenty dotyczyły m.in. użycia stałych lub iteracyjnie wyznaczanych pseudo-etykiet dla nieoznaczonego zbioru, w trakcie dostrajania wstępnie wytrenowanego modelu automatycznego rozpoznawania mowy.

Rozdział 7 (ang. Conclusions) zawiera krótkie podsumowanie wskazujące na najważniejsze osiągnięcia rozprawy w odniesieniu do trzech hipotez badawczych. Ponadto przedstawia w jednym paragrafie praktyczne zastosowania osiągnięć Doktoranta w projektach Samsung R&D Institute Poland oraz formułuje pytania do dalszych prac badawczych.

Załącznik A zawiera listę publikacji konferencyjnych Doktoranta na konferencjach międzynarodowych i krajowych oraz jednego patentu. Natomiast Załącznik B przedstawia dodatkowe wykresy uzupełniające, będących uzupełnieniem wyników przedstawionych w rozdziale 3.

OCENA MERYTORYCZNA PRACY

Analiza źródeł

Bibliografia zawiera 118 źródeł i ze względu na niewielką liczbę przedstawia jedynie wycinek stanu wiedzy w zakresie automatycznego rozpoznawania mowy czy wykrywania fraz kluczowych. W większości są to publikacje zaprezentowane na międzynarodowych konferencjach naukowych. Bibliografia zawiera małą liczbę artykułów opublikowanych w czasopismach naukowych oraz pozycji książkowych. Wiele publikacji konferencyjnych pojawia się na liście referencji jako publikacje CoRR (arXiv) podczas gdy zostały one opublikowane w ramach konkretnej konferencji, , na przykład główna dla rozdziału 4 referencja [97] opublikowana na konferencji ICLR w 2018 roku.

Mała liczba źródeł wpłynęła na znaczne ograniczenie omawianych w pracy metod stosowanych w automatycznym rozpoznawaniu mowy i wykrywaniu fraz kluczowych, a w konsekwencji dość ograniczony przegląd stanu wiedzy głównych zagadnień badawczych pracy.

Strona edycyjna pracy

Praca doktorska jest poprawna od strony redakcyjnej, edycja rozprawy jest staranna, a rysunki i tabele są czytelne. Bardzo pomocne są też liczne i starannie wykonane schematy blokowe. Rozprawa napisana jest poprawnym językiem angielskim, a znaczna część tekstu rozdziałów 4, 5, 6 została wcześniej starannie przygotowana przy opracowaniu wieloautorskich publikacji konferencyjnych.

Metodyka badawcza

Metodyka badawcza stosowana przez Doktoranta w rozdziałach 4, 5 i 6 dla poszczególnych zagadnień badawczych jest właściwa i wskazuje na znajomość problemów oraz metodyki prowadzenia badań naukowych w obszarze automatycznego rozpoznawania mowy. Eksperymenty zostały zaplanowane i przeprowadzone zgodnie z dobrymi praktykami i standardami prowadzenia badań. W przeprowadzonych eksperymentach, dane treningowe i testowe zostały poprawnie dobrane, właściwie dobrane zostały miary ewaluacyjne, a prezentacja wyników umożliwia krytyczną dyskusję nad wpływem poszczególnych założeń czy elementów rozwiązań.

Oryginalność rozwiązania problemu naukowego i uzyskanych wyników badań

Oryginalne rozwiązania problemów badawczych przedstawione zostały w rozdziałach 4, 5 i 6, a wcześniej zostały krótko podsumowane w ramach dwustronicowego rozdziału 3, który wskazuje odniesienia do trzech głównych hipotez badawczych oraz publikacje naukowe powiązane z głównymi osiągnięciami przedstawionymi w pracy. Wszystkie oryginalne rozwiązania przedstawione w pracy doktorskiej zostały opisane w 4 publikacjach będących materiałami konferencyjnymi.

Główne osiągnięcie przestawione w rozprawie zostało opisane w rozdziale 4 i dotyczy wykrywania fraz kluczowych przy pomocy modelu, którego głównym elementem jest model akustyczny. Autorskie rozwiązanie polega na dodaniu do zaczerpniętego z literatury naukowej rozwiązania do automatycznego rozpoznawania mowy [97], prostego unigramowego modelu językowego (ang. unigram language model) oraz zaproponowanie jego inicjalizacji. Zaproponowana metoda inicjalizacji polega na takim ustaleniu wag modelu językowego tak aby wartości dla reprezentacji składającej się na frazę kluczową były na stałym poziomie większym niż jeden oraz wartości jeden dla pozostałych wag modelu językowego. Na podstawie licznych przeprowadzonych eksperymentów wskazano na wartość 32 jako właściwą dla tego pojedynczego hiperparametru. Ponadto w kontekście rozpoznawania fraz kluczowych nieobserwowanych w trakcie treningu modelu akustycznego, wskazano na możliwość wykorzystania wygenerowanej przez syntezator mowy. Eksperymenty przeprowadzono głównie przy użyciu przygotowanej przed Doktoranta bazy danych MOCKS zawierającej frazy kluczowe pochodzące z popularnych otwartych korpusów mowy LibriSpeech i Mozilla Common Voice, która w mojej ocenie stanowi drugie najważniejsze osiągnięcie Doktoranta

zaprezentowane w ramach pracy doktorskiej. Omówione rozwiązanie zostało przestawione w referacie opublikowanym w ramach konferencji FedCSIS 2023 (20 pkt. MEiN), a w eksperymentach użyto stworzonej bazy danych MOCKS opisanej w referacie opublikowanym na konferencji Interspeech 2023 (140 pkt. MEiN).

Autorskim osiągnięciem przestawionym w rozdziale 5 jest niewielka modyfikacja standardowej funkcji kosztu (ang. loss function) opartej o binarną entropię krzyżową (ang. binary cross entropy) przy treningu zaczerpniętego z literatury (publikacja [64]) modelu do wykrywania końca frazy (ang. end-of-speech detection). Konkretnie, do standardowej funkcji kosztu dodana została waga pozwalająca na kontrolę intensywności detekcji końca frazy, która zmienia wartość w zależności od odległości przetwarzanej ramki sygnału mowy od faktycznego końca frazy. Omówione rozwiązanie zostało przedstawione w referacie opublikowanym w ramach konferencji MoMM 2020 (70 pkt. MEiN).

W rozdziale 6 przedstawiono metody dostrajania (ang. fine-tuning) wstępnie wytrenowanego (ang. pre-trained) modelu automatycznego rozpoznawania mowy do domeny użytkownika przy użyciu uczenia częściowo nadzorowanego (ang. semi-supervised training). Jako główne osiągnięcie Doktorant wskazuje na eksperymentalną ewaluację użycia niewielkiej ilości danych, podczas gdy to znane w automatycznym rozpoznawaniu mowy podejście stosowane jest zwyczajowo przy większej ilości danych. Wyniki eksperymentów zostały przedstawione w referacie opublikowanym w ramach konferencji RTSI 2022 (20 pkt. MEiN).

W mojej opinii autorski wkład naukowy jest dość niewielki jak na pracę doktorską.

Główne uwagi i pytania dotyczące recenzowanej rozprawy

Po przeczytaniu rozprawy doktorskiej, nasuwa się kilka uwag i pytań. Główne uwagi dotyczące pracy:

1) Opis stan wiedzy dotyczący tematyki wykrywania fraz kluczowych jak i rozpoznawania mowy przedstawiony w pracy doktorskiej jest dość ograniczony i zawiera jedynie część istniejących w literaturze rozwiązań. Na przykład, prezentując stan wiedzy, nie zostały omówione dominujące w ostatnich pięciu latach podejścia do rozpoznawania mowy bazujące na neuronowych transducerach (ang. neural transducers) czy bardziej aktualne podejścia bazujące na dużych modelach językowych (ang. large language models). Oba podejścia cechuje wysoka dokładność i są istotne dla poruszanej tematyki. Ponadto takie modele nie wymagałyby dodatkowego zewnętrznego modelu językowego ani modułu realizującego detekcję końca frazy, zwracając odpowiedni token.

2) Temat pracy doktorskiej jest w mojej opinii zbyt ogólny w stosunku do autorskich rozwiązań zaprezentowanych w rozdziałach 4, 5, 6 oraz wskazanych przez Doktoranta w rozdziale 3. Główne osiągnięcia dotyczą wykrywania fraz kluczowych w sygnale akustycznym oraz powiązanym z tym tematem wyznaczaniu końca frazy, a ogólny model automatycznego rozpoznawania mowy jest poddany eksperymentom jedynie w rozdziale 6.

3) Dla większości rozwiązań przedstawionych w rozdziałach 4, 5, 6, brak jest porównania wyników rozwiązań Doktoranta z wynikami osiąganymi przez istniejące w literaturze rozwiązania dla tych samych danych testowych (i treningowych). Eksperymentalna ewaluacja ogranicza się przede wszystkim do znalezienia optymalnych dla rozważanych modeli i zbiorów danych hiperparametrów, w szczególności wag modelu językowego w rozdziale 4 i wag w funkcji kosztu (straty) modelu detekcji końca frazy. Porównania dokonane są jedynie z tą samą metodą przy ustaleniu wag na wartość jeden, co traktowane jest jako jedyny punkt odniesienia. Wyjątkiem jest utworzona baza danych, która jest porównywana z istniejącą i popularną bazą komend Google Speech Commands.

4) W pracy nie wykazano jednoznacznie wdrożenia rozwiązań opisanych w rozdziałach 4, 5, 6. Jedyny opis poruszający kwestię wdrożenia to drugi paragraf na stronie 75, w którym wskazano na dalsze prace prowadzone nad rozwiązaniem z rozdziału 4, nad użyciem rozwiązania 5 przy przygotowywaniu danych treningowych lub testowych oraz bardzo generalnie o wykorzystaniu metod dostosowywania się do domeny użytkownika w produktach firmy przy pomocy danych bez etykiet, co każda firma oferująca analogiczne produkty faktycznie robi.

Szczegółowe pytania i uwagi:

5) Na szczególne uznanie zasługuje przygotowanie dużego zbioru danych MOCKS do wykrywania fraz kluczowych w pięciu językach, który opisano w rozdziale 4.2. W zbiorze tym uwzględniono fonetyczne podobieństwa i różnice fraz do frazy kluczowej.

6) Jaki wpływ na zaproponowany w rozdziale 4.4.1 system wykrywania fraz kluczowych miałaby zamiana modelu rozpoznawania mowy z bazującego na modelu akustycznym [97] na bardziej współczesny model bazujący na neuronowych transducerach lub dużych modelach językowych? Czy potrzebny byłby wtedy dodatkowy mały model językowy opisany w rozdziale 4.4.2 z zaproponowaną inicjalizacją oraz detekcja końca frazy z rozdziału 5?

7) Rozbudowana eksperymentalna ewaluacja rozwiązania polegającego na dodaniu modelu językowego opisana w rozdziale 4.5 dotyczy głównie znalezieniu stałej wartości jednego hiperparametru, z lub bez dodania fraz wygenerowanych przy pomocy syntezatora mowy. Czy

potwierdzono eksperymentalnie wpływ otrzymanej w jej wyniku wartość 32 dla innych modeli rozpoznawania mowy?

8) W tabeli 5 przedstawiono miary EER, podając w opisie poniżej, że wartości te nie wynikały z przecięcia krzywych FPR i FNR. Moim zdaniem brakuje choćby przykładu takiego wykresu aby można było unaocznić czytelnikowi jak ta wartość została wyznaczona.

9) W eksperymentalnej ewaluacji przedstawionej w rozdziale 4 brakuje rezultatów dla metod referencyjnych, innych niż ustalenie hiperparametru na wartość jeden.

10) W eksperymentalnej ewaluacji przeprowadzonej w rozdziale 5, brakuje porównania do innych metod umożliwiających detekcję końca frazy lub choćby do systemu wykrywania aktywności głosowej (ang. voice activity detection), jak to miało miejsce w przypadku ewaluacji w publikacji [64] prezentującej używany w pracy model.

11) Jaki wpływ na otrzymane w rozdziale 5 wyniki miało sztuczne wydłużenie długości nagrań dźwiękowych zawierających mowę o 2 sekundy ciszy? Czy wpłynęło ono w równym stopniu na wyniki uzyskane dla stosowania zaproponowanych wag w stosunku do standardowej funkcji kosztu (straty) bazującej na binarnej entropii krzyżowej?

12) Proszę o wyjaśnienie dlaczego w rozdziale 5.4, model trenowany przy pomocy funkcji kosztu z zaproponowanym wagowaniem uzyskiwał gorsze lub podobne rezultaty dla nagrań w warunkach szumowych?

13) Przedstawiając wykresy 18, 19, 20, Doktorant zrezygnował z miar ewaluacyjnych latencji EP50 i EP90 stosowanych jako główne miary w [64] na rzecz wybranych przez siebie miar Early/Fine/Fail EOS. Zaproponowane wagowanie w funkcji kosztu sprzyja poprawie miar Early EOS i Fine EOS, ale powoduje zwiększenie liczby przypadków, w których w ogóle nie wykryto końca frazy, a więc Fail EOS, jak również zwiększa wartości miar latencji. Dlaczego dokonano takiego wyboru miar zamiast porównania latencji (miary EP50 i EP90)? Która z miar jest najbardziej istotna ze względu na doświadczenie użytkownika przy wykrywaniu fraz kluczowych?

14) Czy wyniki przedstawione na wykresach 21, 22 i 23 przedstawiają miarę błędnie rozpoznanych słów (ang. Word Error Rate- WER) jedynie dla zbioru, do którego model jest dostrajany? Jeśli tak, to jaki wpływ na miarę WER dla zbiorów danych z innych domen ma takie dostrajanie?

15) Wykres 21 przedstawia wyniki uczenia częściowo nadzorowanego dla 4 zbiorów danych opisanych w tabeli 10, w różnych wariantach takiego uczenia sprawdzanych przez Doktoranta. Po przeczytaniu rozdziału 6 mam wrażenie, że eksperymenty przedstawione w rozdziale 6 nie pozwalają na wyciągnięcie jednoznacznych wniosków co do skuteczności sprawdzanych rozwiązań dla małego zbioru danych bez etykiet z domeny docelowej, co było celem wskazanym przez Doktoranta do przeprowadzenia tych badań. W wielu przypadkach proces uczenia częściowo nadzorowanego powoduje wzrost błędnie rozpoznanych słów podczas procesu dostrajania, a bez etykiet nie byłoby wiadomo kiedy dokładnie należałoby przerwać ten proces. Poprawę miary błędnie rozpoznanych słów można uzyskać jeżeli wykorzysta się też uczenie nadzorowane danymi z etykietami pochodzącymi z domeny docelowej. Jednak nawet w tym przypadku użyteczność analizowanych zabiegów treningowych jednoznacznie poprawia dokładność rozpoznawania mowy jedynie dla 2 z 4 testowych zbiorów danych. Które ze zbadanych zabiegów treningowych powinien zastosować czytelnik w celu dostrojenia modelu rozpoznawania mowy, posiadając dostęp jedynie do małego zbioru danych bez etykiet z domeny docelowej?

Wnioski końcowe

Podsumowując stwierdzam, że w przedłożonej do recenzji rozprawie doktorskiej mgr. inż. Mikołaja Pudo pt. „Methods of Optimizing Models and Algorithms for Automatic Speech Recognition in Mobile Applications", Doktorant wykazał się znajomością zagadnienia wykrywania fraz kluczowych jako specjalnego przypadku automatycznego rozpoznawania mowy oraz świadomością ograniczeń występujących przy wykonaniu takiego zadania na urządzeniach o ograniczonych zasobach obliczeniowych. Opisany w pracy doktorskiej autorski wkład w dziedzinę naukową oceniam jako niewielki, choć wydaje się on być wartościowy. Również dorobek publikacyjny Doktoranta jest niewielki. W mojej opinii, przedstawiona praca spełnia w stopniu minimalnym warunki stawiane rozprawom doktorskim w aktualnie obowiązującej ustawie o stopniach i tytule naukowym. W związku z tym wnioskuję do Rady Naukowej Dyscypliny Informatyka Techniczna i Telekomunikacja Politechniki Warszawskiej o dopuszczenie Pana mgr. inż. Mikołaja Pudo do dalszych etapów przewodu doktorskiego.