Posudek dizertační práce

Předkládající: Ing. Lukáš Matějů
Název práce: Speech Activity and Speaker Change Point Detection for Online Streams

Předložená práce shrnuje postup vytvoření dvou klíčových komponent systému automatického zpracování mluvené řeči od zadání až po nasazení do skutečného komerčního produktu. Přínos práce proto spočívá zejména ve výsledcích aplikovaného výzkumu, který je podepřen množstvím experimentů a bude sloužit jako spolehlivý zdroj doporučených postupů při realizaci dalších podobných produktů.

Po formální stránce je práce naprosto v pořádku. Logické členění je přehledné, jednotlivé koncepty lze snadno dohledat a text je čtivý. Pokud mohu soudit, angličtina je rovněž velmi dobrá.

Teoretická část uvádí obšírný přehled stavu techniky. Vybrané koncepty přímo související s předloženou prací jsou rozvedeny v detailu, aby na nich bylo možné dále stavět.

Práce řeší dvě dobře definované a příbuzné úlohy, a to detekci řečové aktivity a detekci změny řečníka, oboje pro účely zpracování v reálném čase. Důraz řešení je proto kladen na minimální dobu odezvy systému a na nízké výpočetní nároky. Navržená řešení obou úloh vycházejí ze stavu techniky a rozvíjejí již známé postupy především s ohledem na uvedená omezení. Postup řešení je dobře zdokumentován a výsledky experimentů prokazují jejich správnost. Výbornou vizitkou je zejména nasazení v komerčním produktu.

Některé postupy v experimentální části nalézají optimální parametry modelů výběrem z mnoha variant s pomocí dat, například počet neuronů ve vrstvách neuronové sítě. V těchto případech bych ocenil rozšíření experimentů více k extrémům, aby bylo zřejmé, kdy věci přestanou fungovat. Při použití váhovaných překladových automatů (WFST) není zcela zřejmé algoritmické zpoždění, předpokládám ale, že byly použity mechanizmy k jeho zastropování.

Výsledky práce byly průběžně publikovány na známých konferencích a se zájmem sledovány komunitou.

Soudím, že práce Ing. Matějů naplňuje požadavky pro udělení akademického titulu a doporučuji ji k obhajobě.

V Litoměřicích, 6. února 2020

Ing. Petr Fousek, Ph.D.

FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

Posudek oponenta na disertační práci

Doktorand: Ing. Lukáš Matějů, FM TU Liberec

Název práce: Speech Activity and Speaker Change Point Detection for Online Streams

Oponent: Doc. Dr. Ing Vlasta Radová, FAV ZČU Plzeň

Vyjádření k významu práce pro obor:

Práce se zabývá detekcí řečové aktivity a detekcí změny řečníka v řečovém signálu, přičemž důraz se klade zejména na on-line použití při segmentaci řečového signálu. Oběma uvedeným tématům je odbornou veřejností pozornost věnována dlouhodobě, nicméně většinou pro off-line aplikace. Z tohoto důvodu považuji předloženou práci za aktuální a pro obor přínosnou.

Vyjádření k postupu řešení problému, použitým metodám a splnění cílů:

Doktorand si ve své práci stanovil 2 základní cíle, které jsou uvedeny v kapitole 2.3 na straně 24. Prvním cílem bylo vyvinout přístup k detekci řečové aktivity a k detekci změny řečníka, který bude využívat nejmodernější techniky, zejména DNN, bude dostatečně robustní, bude fungovat on-line s nízkou latencí a bude moci být integrován do systému pro monitorování televizního a rozhlasového vysílání vyvíjeného na pracovišti doktoranda. Druhým cílem pak bylo otestování navržených postupů a srovnání dosažených výsledků na veřejně dostupné databázi s výsledky dosaženými pomocí vybraných existujících postupů a nástrojů. Stanovené cíle považuji za disertabilní.

Jak v případě detekce řečové aktivity, tak v případě detekce změny řečníka doktorand nejprve provedl přehled používaných metod. Přehled je proveden stručně zejména odkazy na příslušnou literaturu, podrobnějšímu popisu se doktorand věnoval pouze u metod a přístupů, které v práci dále využíval. V přehledu jsou zmíněny v podstatě všechny nejdůležitější aktuální přístupy používané v úlohách detekce řečové aktivity a změny řečníka, doktorand se seznámil s poměrně velkým množstvím literatury. K této části práce mám jednu poznámku. Na str. 19 je v souvislosti příznaky, používanými při detekci změny řečníka citován článek [47], konkrétně je zde uvedeno *In the early years, more straightforward ones were successfully employed, such as zero-crossing rate or pitch [47]*. V uvedeném článku se však počet průchod nulou a základní hlasivkový tón nepoužívá k detekci změny řečníka, ale k detekci konce věty. Dále bych měla dotaz k pojmům/zkratkám SAD (Speech Activity Detection) a VAD (Voice Activity Detection). Z úvodu se zdá, že mezi těmito pojmy vidí doktorand určitý rozdíl, nicméně v kapitole 3 jsou používány spíše jako synonyma. Myslím, že by toto doktorand mohl upřesnit během obhajoby práce. A ještě bych chtěla požádat o podrobnější vysvětlení metody popsané v kapitole 3.2, konkrétně jak na obrázku 3.2 vznikla řádka označená max sequence.

V dalších částech práce se doktorand věnoval návrhu vlastního přístupu k detekci řečové aktivity a k detekci změny řečníka. U obou úloh nejprve popsal evaluační metriky, pomocí kterých vyhodnocoval provedené experimenty, po té použitá data a pak se věnoval popisu jednotlivých experimentů. Právě kapitola 5.1, kde jsou jasně a názorně vysvětleny jednotlivé evaluační metriky používané při experimentech týkajících se detekce řečové aktivity, se mi z celé práce líbí nejvíce. Přesto mám k této kapitole jeden dotaz. Na str. 53 je věta *Precision and recall are in a contradictory relationship with each other (i.e., when one improves the other one worsens)*. Čím je tento protichůdný vztah způsoben?

Co se týče použitých dat, považovala bych za vhodné, aby při obhajobě bylo podrobněji vysvětleno, jakým způsobem byla připravována. Např. v kapitole 5.2 týkající se popisu dat pro experimenty s detekcí řečové aktivity se píše, že pro trénování se používalo 30 hodin dat čisté řeči, 30 hodin hudby a 7 hodin neřečových událostí a šumu, přičemž anotace byla provedena automaticky. Znamená to, že těch 30 hodin řeči bylo souvislých, 30 hodin hudby taky a 7 hodin šumu taky, tedy v podstatě v trénovacích datech byly jen 3 segmenty? Pokud ne, tak jakým způsobem byla provedena ta automatická anotace? Podobně v kapitole 5.5 na str. 58 – Jak konkrétně byla zmixována čistá řeč s neřečovými nahrávkami při vytváření umělých trénovacích dat? Nebo v kapitole 5.6 na str. 59 se píše *two recordings were chosen randomly from the artificial training set: one speech and one non-speech.* Ale podle str. 58 ta data v artificial training set jsou již zmixována z řeči a neřeči. A také v kapitole 6.2 na straně 79 – jakým způsobem konkrétně byla provedena anotace trénovacích dat, když *The annotations of this data (for SCP detection) were generated in a fully automated way*? Pokud by automatická anotace nebyla stoprocentně úspěšná, ovlivní to výsledky jednotlivých experimentů.

Jak při experimentech s detekcí řečové aktivity, tak při experimentech s detekcí změny řečníka si doktorand nejprve stanovil základní variantu, a po té měnil jednotlivé parametry této varianty s cílem zjistit možné lepší nastavení. Výsledky těchto experimentů prezentoval v tabulkách, případně v grafech. Rozsáhlejší tabulky z těchto experimentů jsou uvedeny také v příloze práce, bohužel však na ně není v práci nikde odkaz, takže čtenář se k nim dostane až po té, co práci dočte. Výsledky provedených experimentů jsou vždy stručně okomentovány a v podstatě doktorand vždy dojde k závěru, že nejlepší možné nastavení je nastavení provedené v základní variantě. V této souvislosti mám tedy otázku, jakým způsobem bylo nastavení parametrů v základní variantě provedeno.

Závěrečné experimenty a otestování navržených postupů byly provedeny na databázi COST278 obsahující nahrávky rozhlasového vysílání v několika evropských jazycích. Dosažené výsledky experimentů prokazují, že navržené postupy lze použít k detekci řečové aktivity a k detekci změny řečníka tak, jak bylo stanoveno v cílech disertační práce. Z tohoto pohledu tedy považuji cíle práce za splněné. Jediný bod, ke kterému se nedokážu vyjádřit, je požadavek možnosti začlenit navržené postupy do systému vyvíjeného na pracovišti autora ve spolupráci s firmou NanoTrix, protože v práci není žádná zmínka o tom, jak by mělo rozhraní mezi stávající verzí systému a postupy navrženými v této práci vypadat. Předpokládám, že toto doktorand vysvětlí v odborné diskusi při obhajobě.

Vyjádření k výsledkům disertační práce a k přínosu doktoranda:

Jak už jsem uvedla v předchozí části posudku, výsledky experimentů jsou shrnuty v tabulkách a grafech a jsou stručně okomentovány. Podle mého názoru jsou dosažené výsledky správné, nicméně bych předpokládala, že doktorand ve zhodnocení více rozebere případné příčiny toho, proč jsou výsledky právě takové, jaké jsou. Například u kapitol 5.7.1, 5.7.2 a 5.7.3 by mě zajímalo, jestli se doktorand zamýšlel nad důvody, proč neuronová síť se širšími skrytými vrstvami dosahuje horších výsledků než ta, která má skryté vrstvy užší. Podobně síť s větším počtem vrstev. Nebo vyhodnocení detekce řečové aktivity v různých podmínkách šumu (obr. 5.20, 5.21 a 5.22 na str. 72 až 74) – z grafů vyplývá, že výsledky v různých prostředí jsou vždy relativně stejné: nejlepší car, pak home, street, cafe a nejhorší reverb. Lze toto nějak vysvětlit? A také výsledky detekce změny řečníka pro on-line i off-line režim v grafech 6.9 a 6.10 v kapitole 6.11 ukazují, že nově navržený přístup dosahoval lepších výsledků ve srovnání s nástrojem LIUM toolkit pro baskičtinu, španělštinu, galštinu, chorvatštinu a slovinštinu, zatímco pro češtinu, na kterou byl v předchozích experimentech laděn, dosahoval výsledků horších. Je pro to nějaké vysvětlení?

Přesto jednoznačně konstatuji, že předložená práce je v odborné komunitě přínosná.

Vyjádření k systematičnosti, přehlednosti, formální úpravě a jazykové úrovni práce:

Práce je zpracována systematicky a přehledně. Je psána v angličtině, k formální a jazykové úpravě nemám v podstatě výhrady. V práci se také vyskytuje minimum obsahových chyb, nicméně zcela se jim doktorand nevyhnul (např. ve vztahu (4.12) na str. 44 u 3. členu na pravé straně výrazu chybí log).

Vyjádření k publikacím doktoranda:

V seznamu svých publikací doktorand uvádí 12 prací za období let 2015 až 2019, všechny publikace souvisejí s tématem disertační práce. Ve většině případů se jedná o publikace na významných mezinárodních konferencích v daném oboru, v jednom případě jde o časopisecký článek v časopise zařazeném do Q1. Všechny publikace jsou publikovány v autorském kolektivu, v 5 případech byl spoluautorem školitel doktoranda. U časopisecké publikace jde o publikaci v mezinárodním autorském kolektivu. Dle mého názoru je publikační aktivita doktoranda velmi dobrá a pro doktorské studium naprosto dostačující.

Závěr:

I přes výše uvedené připomínky předložená disertační práce dle mého názoru splňuje požadavky stanovené zákonem č. 111/1998 Sb. o vysokých školách. Práce obsahuje původní a uveřejněné výsledky a ing. Matějů v ní prokázal schopnost a připravenost k samostatné činnosti v oblasti výzkumu nebo vývoje. **Předloženou disertační práci jednoznačně doporučuji k obhajobě.**

Otázky do diskuse:

- Jak je navržený postup pro detekci změny řečníka úspěšný v případě rychlých změn nebo úseků, kde si řečníci skáčou do řeči apod.?
- Je databáze COST278 veřejně dostupná, tj. je k dispozici i pracovištím, které se na jejím vývoji nepodílely?
- Jak při detekci řečové aktivity, tak při detekci změny řečníka jste využíval neuronovou síť se dvěma výstupy znamenajícími rozhodnutí řeč/neřeč resp. změna nastala/nenastala. Neuvažoval jste v případě tohoto binárního rozhodování o síti s jedním výstupem? Respektive jak jste se v případě sítě se dvěma výstupy vypořádával se situací, že rozhodování této sítě není binární?

V Plzni, 10.5.2020

Doc. Dr. Ing. Vlasta Radová