Prof. Ing. Luděk Müller, PhD.,
Fakulta aplikovaných věd ZČU, Katedra kybernetiky
Technická 8 306 14 Plzeň
Tel.: 377 632 508, 377 632 523
Email: muller@kky.zcu.cz

Oponentský posudek disertační práce Mgr. Jana Oldřich Krůzy:

Iterativní zdokonalování přepisu zvukových nahrávek s využitím zpětné vazby posluchačů´

Posuzovaná práce pana Mgr. Jana Oldřicha Krůzy se zabývá problematikou přepisu řeči do textu, speciálně se pak věnuje možnosti postupného zdokonalování výsledného přepisu do textu s využitím systému automatického přepisu a postupného rozšiřování trénovací množiny, to vše se zaměřením na úzkou doménu, kterou je soubor zvukových nahrávek obsahujících hovory českého filozofa ing. Karla Makoně.

Předložená písemná práce o délce 109-ti stran (včetně odkazů na literaturu, seznamu tabulek a seznamu publikací autora) je psaná v češtině a má dobrou jazykovou i grafickou úroveň s minimálním počtem překlepů. Práce je rozdělena do logických částí (uvedených v obsahu práce), kdy poměrně velký prostor je věnován popisu života a nauky ing. Karla Makoně. Také v této verzi práce bych uvítal, kdyby práce čítala i samostatnou kapitolu vytyčující cíle práce, kapitolu popisující stav současného vědeckého poznání a kapitolu shrnující přínos disertační práce k posunu tohoto poznání. Práce je ale čtivá a dobře srozumitelná, její výklad je vesměs pochopitelný, i když na některých místech užívá ne zcela obvyklou terminologii.

Praktickým cílem a pravděpodobně i hlavním hybatelem práce bylo dosáhnout co nejpřesnějšího přepisu audionahrávek ing. Karla Makoně zaznamenaných na kazetách a magnetofonových kotoučích a zpřístupnit je široké komunitě zájemců využitím metod vyhledávání v (automaticky) přepsaném audiu. Pro dosažení tohoto cíle bylo třeba vyřešit řadu podcílů, a to od přípravy a analýzy dat (přípravy korpusů zvukových nahrávek), postavení systému automatického rozpoznávání řeči, jeho adaptaci na cílové prostředí a doménu, testování a vývoj metod pro postupné interaktivní rozšiřování množiny trénovacích dat s využitím kontroly (se vzdáleným přístupem) prováděné posluchači a čtenáři přepsaného textu.

K práci mám řadu připomínek, z nichž nejpodstatnější je nestandardní pojetí obsahu textu disertační práce. Očekával bych popis současného vědeckého poznání, upozornění na jeho neúplnost, na nevyřešené problémy a konečně i popis vlastního přínosu k posunu tohoto poznání a k řešení nastíněných problémů. Posuzovaná práce vyznívá spíše jako aplikační s relativně menším důrazem na výzkum nových metod z oblasti zpracování přirozeného jazyka. Dále je práce zatížena množstvím nepřesných tvrzení, která by též kromě naplnění své správné pravdivostní hodnoty měla mít (s ohledem na obor disertace) exaktnější formu než často jen prosté vágní vyjádření v přirozeném jazyce.

V následujícím seznamu uvádím jednotlivé dílčí připomínky:

1. "K nim [k MFCC koeficientům – poznámka oponenta] se najde pomocí E-M algoritmu normální distribuce, která je generuje". To je značně nešťastné tvrzení. E-M algoritmus by se pro odhad parametrů normálního rozdělení neměl používat, neboť lze odvodit rovnice v explicitním tvaru. Takže distribuce bude asi trochu složitější než autor uvádí, jaká je? (str. 30).

2. V části 3.4.1 „Spektrální odečet šumu“ autor konstatuje, že mu „Chybovost na slovech vzrostla skoro na sto procent.“ a „Přesnou příčinu zatím neznám.“ Z jakého souboru dat se vybíral úsek ticha pro redukci šumu v konkrétní nahrávce – zahrnul či nezahrnul autor do tohoto souboru dat i jiné nahrávky než pro tuto konkrétní? (str. 32)

3. V části 3.4.2 „Neurální doménový transfer“ je síť CycleGAN použita pro odstranění přebuzení nebo nízké kvality „nízkootáčkových“ nahrávek. Výsledky uvedené v části 3.4.3 „Vyhodnocení“ nejsou dobré. Zamýšlel se autor nad použitím „augmentování“ trénovacích dat a to transferem dobrých nahrávek na přebuzené či „nízkootáčkové“ a následného natrénování nového (akustického) modelu nad těmito daty?. (str. 33)

4. "Jak by taky bylo lze nalézt funkci" – lze "lze" nahradit slovem "možno"? (str. 33)

5. Nerozumím na straně 36 tomu, proč je pro konstrukci jazykového modelu bráno jako nedostatek, že se z přepisu jednání PS ČR „vynechávají odkazy na jiné schůze“. Co se tím přesně myslí? (str. 36)

6. V části 4.2. „Číslovky a zkratky" opravdu nelze očekávat, že pouhé "zahrnutí číslic do abecedy a tedy přepis číselných výrazů přímo na číslice" bude při známé velikosti trénovacích dat (cca 500 tis. čísel) fungovat. Autor správně navrhuje nejprve před vlastním trénováním aplikovat (automatický) rozpis číslicového výrazu (asi nikoliv "číslic", jak je psáno na straně 39), ale ze všech pak vybírá jen "nejpravděpodobnější variantu". V jakém smyslu je zde chápána nejpravděpodobnější varianta (bere algoritmus např. v úvahu akustiku)? (str. 39)

7. S tvrzením, že v metodě dynamic time warping "pochází veškerá expertiza ze strany člověka" nemohu souhlasit. (str. 41)

8. Věta "Hluboké neuronové sítě nejdříve vstupují jako součást schématu s HMM, pro odhad aposteriorní pravděpodobnosti přepisu na základě akustických dat." by zasluhovala upřesnění. DNN jako součást HMM v konečném důsledku totiž poskytují odhad podmíněné hustoty pravděpodobnosti pozorování posloupnosti vektorů příznaků za podmínky, že tato posloupnot odpovídá dané akustické jednotce (např. stavu trifónu) (str. 41).

9. Práce používá dosti svérázné a nepřesné formulace, např. jen na straně 42:
   - "Zvukový signál je velmi prostý: je to spojitá funkce času do reálných čísel."
   - "transformuje proud jednoho reálného čísla šestnáctkrát za milisekundu do proudu reálně číselného vektoru stokrát za sekundu"
   - "okna jsou od sebe vzdálena setinu sekundy, takže se překrývají"
   - "Pro počítačovou implementaci jsou reálná čísla díky svému nekonečnému desetinnému rozvoji nereálná".

   (str. 42)

10. Kapitola 5.2. "Kódování signálu" je psána velmi nepřesně, především u výpočtu MFCC: Pomineme-li nepřesné formulace typu "Na každém časovém okně se provede diskrétní Fourierova transformace a rozdělí se do frekvenčních oken..." (str. 43 - co se rozdělí?); místo termínu "okno" se používá "frekvenčí pásmo"; chybí popis proměnných ve vztazích (5.1) – (5.4), "Jako třináctá hodnota se přidá buď základní frekvence," (co je základní frekvence?), "Od těchto hodnot se na základě kontextu přidá první a druhá derivace," atd., je závažným nedostatkem popisu parametrizace signálu (práce front-endu) opomenutí skutečnosti, že se (kromě preemfáze) signál musí před DFT převážit nějakým váhovým okénkem – např. Hammingovým. Ještě závažnější nedostatek popisu pak spočívá v tom, že v popisu) je zcela vynecháno aplikování logaritmu na Melovy (nekepstrální) koeficienty MFC a rovnou se na obdržené MFC aplikuje zpětná diskrétní Fourierova tranformace IDFT (ve skutečnosti se navíc aplikuje DCT). Také delta a delta delta koeficienty MFCC se počítají trochu jinak, než autor uvádí. Navíc zkratka DFT, kterou autor používá, není v textu práce nikde vysvětlena. (str. 41-43)

11. "Pravděpodobnost vstupních dat"? Ta se blíží nule! (str.44)

12. "Markovovský model realizuje akustický model jako sumu odhadnutých pravděpodobností přechodů stavy automatu při pozorování dané vstupní posloupnosti:" - opět nešťasně formulováno. (str. 44)
13. Co je M v rovnici (5.7)? (str. 44)
14. Na obrázku 5.5. je uvedeno nepřesné označení "akustická data" - co je přesně O, které mimochodem v textu není nikde vysvětleno? (str. 45)
15. Nepřesné a nepravdivé: "Vymezení [v prostoru vstupních dat? - poznámka oponenta] se provádí pomocí pravděpodobnostní distribuce určené středem a variancí v každé dimenzi". (str. 45)
16. "ticho modeluje i neřečové události a často odpovídá dlouhým úsekům, povoluje se u něho, jakož i u krátké pauzy, přechod mezi druhým a čtvrtým stavem v obou směrech". Nikoliv ticho, ale model ticha. (str. 45)
17. Popis k rovnici (5.8) "Jestliže pak rozeznáváme u češtiny čtyřicet jednu hlásku plus ticho plus krátkou pauzu, bude tento jednoduchý akustický model mít $41 \times (6 + 2 \times 39) + 2 \times 39 + 8 + 9 = 3539$ volných parametrů" je mi nejasný. Co jednotlivá čísla znamenají? (str. 45)
18. Neobvyklou psanou formu má též výraz "gaußovskými směsmi" (str. 41). I další termíny jako např. "markovovských modelů" (např. str. 31), „automated speech recognition" (str. iii a str. 40) jsou neobvyklé a v literatuře nezavedené. (např. str. iii, 31, 40, 41, 55, ...)
19. Tvrzení "Všechny hlásky se inicializují jako shodné." na str. 54 u popisu trénování akustického modelu HMM v systému HTK je velmi vágní. (str. 54)
20. Podobně vágní jsou výrazy typu: "Střed a variance [složek GMM – poznámka oponenta] jsou určeny identicky podle globálních hodnot" (str. 54).
21. "Vybere se kritérium, které log likelihood zvýší nejvíce" – jaké kritérium? Jedná se nejspíše o výběr otázky v rozhodovacím fonetickém stromu HTK (str. 55).
22. "Každá [gassovská distribuce – poznámka oponenta] má svůj střed, svoji varianci a svoji váhu, jejichž celkový součet musí být roven jedné" – tvrzení je nepřesné. Jedná se o vícerozměrné gaussovské rozdělení! (str. 55).
23. Co je malé m v rovnici (5.24)? (str. 57)
24. Obrat „Z každé dvacáté jsem snížil na polovic nejen abych neplýtval trénovacími daty, nýbrž také protože vyhodnocování směsí zabírá při trénování zdaleka nejvíce času, a ten je přímo úměrný velikosti sady heldout." je nejasný (str. 59).
25. Kepstrální normalizace MFCC se standardně již dlouhá léta provádí na částech neobsahujících neřečový úsek (ticho). Metoda uváděná v části 5.8 „Experiment s kepstrální normalizací" na straně 60 tedy není nijak nová, navíc často obsahuje nejen CMN ale i CVN. (str. 60)
26. V části 5.13 na str. 63 autor uvádí, cituji: „*Out of vocabulary*, tedy mimo slovník. Tak se zove jev, kdy výstupem rozpoznávání řeči je slovo, které jazykový model nezná. Může k tomu dojít ve dvou případech: Buď když je neznámé slovo součástí jazykového modelu nebo když je rozpoznávání řeči schopno vydávat i slova mimo slovník jazykového modelu." OOV jsou slova neobsažená ve slovníku (používaného jazykovým modelem) rozpoznávacího systému, která jsou však obsažena v rozpoznávaném souboru akustických nahrávek, a to bez ohledu na to, zda tato OOV slova jsou či nejsou výstupem (end-to-end systému) rozpoznávání. (str. 63)
27. Obrat "ručně přepsaná slova a automaticky přepsaná slova mají tendenci se shlukovat" také není zcela přesným popisem daného jevu. Co přesně se tímto shlukováním myslí? (str. 74)
28. Nerozumím tomu, proč by úloha nalezení bodu předělu (tj. místa, kde lze zvukovou nahrávku rozdělit pro další zpracování, aniž bychom ji dělili uprostřed slova) neměla mít řešení pro případy, že délka ticha trvá déle než 120 sec. Větší délka ticha na začátku či na konci nahrávky nevadí a

navíc lze přeci i libovolně dlouhý úsek ticha vystřihnout. (část 6.3.3 „Výběr bodů předělu“ (str. 79)

29. Algoritmus výběru bodů předělu na straně 80 popsaný slovy: “Začneme s množinou všech tich a iterujeme přes ně od nejkratšího po nejdelší. Ticho z množiny odebereme, pokud sloučením sousedních segmentů nevznikne segment delší než 60 sekund. Přes vybraná ticha znova iterujeme a ticho odebereme, jestliže jeden z jeho sousedů má méně než 30 sekund.” by měl být formulován exaktněji.(str. 80).
30. Na straně 71 pak věta “Čtvrtá kombinace fonetického zápisu a špatné výslovnosti se pochopitelně nevyskytuje.” není asi přesná a pravděpodobně znamená “Čtvrtá kombinace správného fonetického zápisu a špatné výslovnosti se pochopitelně nevyskytuje.”. Anebo je tomu jinak? (str. 87).
31. Tvrzení “recall je není vyloženě žalostný” je podivné. Na základě dat tabulky odhaduji, že u tématu Lazar by odhadnutý recall na základě „interpolace“ (extrapolace?) mohl být pod 50 %. (str. 92)

Přes uvedené připomínky musím kladně hodnotit opravdovou snahu autora o dosažení co nejlepšího výsledku v úloze automatického přepisu korpusu audionahrávek ing. Karla Makoně zaznamenaných na kazetách a magnetofonových kotoučích. Práce je co do šířky tématu značně obsáhlá a její zvládnutí zabralo autorovi mnoho úsilí a času.
Přínosem práce je především návrh, realizace a vyhodnocení metody pořízení kvalitního zarovnaného přepisu velkého množství dat se zapojením pouze malého počtu laických přispěvatelů, metoda získávání specifických trénovacích dat pro aktivní učení od dobrovolných anotátorů, nový téměř dvoutisícihodinový korpus pro úlohu S2T, který autor dává svobodně k dispozici, i dostupnost všeho autorem použitého kódu pro řešení úlohy. Domnívám se, že z tohoto důvodu lze dovozovat, že výsledky provedených experimentů jsou reprodukovatelné. Bohužel, srovnání dosažených výsledků s výsledky dosaženými jinými autory není možné z důvodu specifičnosti zkoumané úlohy, která není řešena nikým jiným, než autorem. Kladně hodnotím úsilí autora o zvládnutí celé šíře zkoumané problematiky i jeho snahu o aplikování aktuálních state-of-the-art metod, které se v průběhu řešení disertační práce poměrně výrazně vyvíjely. Práci proto doporučuji k obhajobě s tím, že při ní od autora očekávám zdůraznění hlavních přínosů práce k současnému stavu vědeckého poznání a jejího významu pro rozvoj studovaného vědního oboru.

V Plzni 7. 9. 2021

.....
Luděk Müller