POSUDEK NA DISERTAČNÍ PRÁCI

Téma práce: Iterativní zdokonalování přepisu zvukových nahrávek s využitím zpětné vazby posluchačů

Doktorand: Mgr. Jan Oldřich Krůza

Posudek vypracoval: Doc. Ing. Petr POLLÁK, CSc.
ČVUT FEL K13131, Technická 2, 166 27 Praha 6

Předložená práce Mgr. Krůzy zpracovává interdisciplinární problematiku přepisu existujících zvukových nahrávek hovorů českého filozofa, ing. Karla Makoně. Ač se jedná o v současné době často řešenou problematiku s výsledky dostupnými pro světové jazyky i pro češtinu, zpracování off-line audioarchivů představuje vždy výzkum vázaný na konkrétní nahrávky, kdy je nutné řešit mnoho konkrétních problémů souvisejících zejména s různou kvalitou i konkrétním obsahem zpracovávaných záznamů, což přináší prostor pro disertabilní výzkum.

Cíle práce stanovené v úvodní kapitole považuji disertabilní, autor tyto cíle v předložené práci s určitými výhradami naplnil a nejvýznamnější přínosy s připomínkami z mé strany lze shrnout v následujících bodech.

- V první části práce autor shrnuje život Karla Makoně a obsah jeho díla, jímž je evidentně významně osloven. Zpřístupnění tohoto díla formou přepisů vnímám jako zajímavé z hlediska obecné filosofie, teologie resp. religionistiky, což lze považovat za jeden z autorem zmíněných interdisciplinárních přínosů dané práce. Celkový význam Makoňova díla však neposuzuji, to je na odbornících ze zmíněných filosofických oborů.

  Dále autor realizuje rozbor resp. analýzu obsahu Makoňova díla, včetně automatické identifikace témat textovým vyhledáváním. Realizovaný popis a přehled jednotlivých témat je logickým prvním krokem, viz kapitola 2.2, zdá se mi však až příliš podrobný. V kontextu této práce vidím význam analýzy témat zejména z hlediska využití při následné anotaci nahrávek v korpusu resp. možnosti přizpůsobení automatického přepisu danému tématu. Pro tyto účely resp. pro účely korpusové lingvistiky by byl podle mého názoru vhodnější stručnější seznam témat.

- Poměrně složitým problémem, se kterým se autor musel vypořádat, jsou akustické vlastnosti resp. nižší kvalita zpracovávaných audio nahrávek různého stáří. Autor analyzuje kvalitu různě degradovaných záznamů, v první řadě na časových průbězích a spektrogramech; 9 velmi podobných obrázků (3.1-3.9) je ovšem nadbytečné množství a dává jen velice orientační informaci, zejména pak v případě relativně dlouhých záznamů. Doporučil bych ilustrativní zobrazení menšího počtu podobných obrázků a hlavně pak použití výrazně kratších signálů, optimálně okolo 10 sekund či méně, kdy lze lépe srovnat spektrální obsah nejen v pauzách, ale i v řečových úsecích.

  Vhodnějším způsobem analýzy je jistě použití objektivní akustické metriky s následným hierarchickým shlukováním podobných signálů. Bylo by ale třeba použitou metriku podrobněji popsat, autor však jen odkazuje na prameny [17] a [18] a pouze definuje pojem "akustická vzdálenost". Použitý algoritmus zdaleka není založen jen na zmíněných MFCC, jak autor stručně uvádí. Nicméně funkčnost použité metriky potvrzují dosažené výsledky, tj. zejména srovnání vzdáleností mezi různými nahrávkami z jednání Poslanecké sněmovny Parlamentu ČR vs. srovnání promluv v Makoňově korpusu

(výrazně větší hodnoty). Výsledek provedeného shlukování však není zmíněn. Vedlo shlukování k rozlišení různých typů zkreslení zmíněných výše?

- Kompenzaci akustických nedostatků se autor pokusil řešit dvěma metodami. Nejprve na bázi spektrálního odečítání pomocí obecně používaného nástroje "sox", kdy z vybraných úseků získá vzorek pozadí, které se následně ve spektrální oblasti odečítá. Dochází k závěru popsanému v mnoha předchozích pracích, že uvedenou techniku lze použít pouze na odstranění šumů stacionárních a podobných vybranému referenčnímu úseku. Zhoršení výsledků přepisu díky dalšímu vzniklému zkreslení signálu je pochopitelné. Z tabulky 3.1 navíc není zcela zřejmé, zda byl uvedený postup použitý jednotlivě pro různá zkreslení (napočítané shluky). Mohl byste toto upřesnit? V každém případě konstatování *"Přesnou příčinu neznám, ovšem vzhledem k nízkým očekáváním, které jsem od metody měl, nepovažuji za účelné se po ní pídit."* nepatří do disertační práce.

  Druhou zvolenou metodou je transformace signálů pomocí neuronové sítě typu CycleGAN, což autor vyzkoušel na dvou typech dat, bohužel opět bez výsledků umožňujících praktické použití. Problémem zde je použití hotového nástroje s malou možností případné modifikace pro dané účely.

  V obou zmíněných případech se navíc jedná o techniky, generující opět akustický signál. Metody na bázi spektrálního odečítání lze však použít v rámci extrakce příznaků a pro standardní aditivní šum zlepšení při rozpoznávání jistě přinášejí. Podobně při použití neuronových sítí se domnívám, že správnější cestou je použití případně modifikace sítě generující přímo přepis resp. sítě v DNN-HMM systému než snaha o získání nezkresleného akustického signálu, který je následně rozpoznáván.

- Jádrem práce je realizace automatického přepisu daného korpusu. Autor vycházel z mnoha prací řešených v širším mezinárodním měřítku, nejvýznamnější práce jsou citovány, systematická podrobnější rešerže však chybí. Podrobnější popis existujících systémů přepisu audio resp. multimediálních archivů, a to včetně diskuse nad dosahovanými výsledky a rozdíly v přepisovaných datech, bych viděl jako klíčovou část práce, ze které pak může vycházet zdůvodnění pro použití použitého systému přepisu a jeho potřebné přizpůsobení realizované úloze.

  Autor nakonec použil dva různé rozpoznávače řeči: první na bázi GMM-HMM ve starší implementaci pomocí HTK Toolkitu v poměrně standardní konfiguraci a se standardními trénovacími postupy a jako druhý pak volně dostupný systém DeepSpeech založený na hlubokých neuronových sítích (DNN).

  Hlavní přínos práce v této oblasti vidím osobně ve skutečnosti, že autor se k automatickému přepisu nahrávek Makoňova korpusu v akceptovatelné kvalitě dopracoval. Klíčovým krokem k relativně úspěšnému přepisu je iterativní vylepšování použitých systémů na základě korigovaných přepisů. Dosažená přesnost je relativně dobrá, autor zmiňuje WER 22,2 %. Osobně jsem pak při poslechu několika úseků ve zpřístupněném korpusu byl spíše překvapen, chyby v přepisech byly menší a rozhodně jsem v poslechnutých částech zaregistroval spíše jen menší změnu smyslu sdělení. To považuji při přepisu záznamů zmíněné kvality za dobrý výsledek. Přesto bych měl k této části pár následujících připomínek a otázek.

- V první řadě implementaci na bázi HTK Toolkitu považuji za nešťastnou. Volba GMM-HMM architektury použitého rozpoznávače může mít logické důvody, autor zmiňuje snažší implementaci, možnost natrénování s menším množstvím dat. GMM-HMM systém lze však realizovat i pomocí nástrojů KALDI, a to jistě v pokročilejší konfiguraci.

Autor by si ušetřil řadu komplikovaných situací a snadněji by mohl využívat průběžné upgrady KALDI či přejít relativně snadno k vlastní implementaci DNN-HMM systému.

Autor se pak snaží o vylepšení v principu velmi standardního systému na různých úrovních, zejména pomocí kepstrální normalizace a aktivního učení s výběrem dat na bázi míry spolehlivosti. Obojí však ke zvýšení přesnosti přepisu nevedlo. Předpokládám, že rozumím dobře textu, že tyto metody nebyly nakonec použité?

U kepstrální normalizace lze obecně předpokládat horší výsledky v případě příliš velké variability akustických podmínek. Pokusy s vynecháním neřečových úseků toto spíše nemohou překonat. Pro dlouhé záznamy by měl být fonetický obsah dostatečně bohatý, aby průměrné kepstrum obsahovalo informaci o případně přítomné konvoluční složce. Zásadní je ovšem předpoklad existence konvolučního zkreslení. Z textu není zcela jasné, zda byly při výpočtu průměrného kepstra rozlišovány různé typy zkreslení. Obecně by na základě provedených analýz různých variant degradace měla být přizpůsobena kompenzační technika dosaženému výsledku: např. echo se bude kompenzovat jinak než aditivní šum či silné nízkofrekvenční zkreslení u nahrávek malou rychlostí. Jedná se o podobný problém jako u použití spektrálního odečítání. Prosím o upřesnění!

- Segmentace přepisovaných dat na kratší úseky je významným krokem a realizace na základě aktuálně existujících přepisů dává smysl. Možná šikovnější přístup v případě první segmentace by bylo použití jednoduchého detektoru řečové aktivity místo segmentace na 15s úseky. Navíc je toto zmíněno v kapitole 6.3 při popisu segmentace pro zpřístupnění ve WEBovém rozhraní.
- Systém přepisu založený na DNN je realizován pomocí DeepSpeech a autorův přínos je v použití manuálně přepsaných dat Makoňova korpusu resp. doplněných o další trénovací data použitá i pro trénování GMM-HMM systému. Při analýze dosahované úspěšnosti v tabulce 5.3 je zřejmé, že systém na bázi DNN je přesnější. Ač se jedná o převzatý nástroj, popis by měl být podrobnější, ne omezený jen na odkazy.
- Analýza přesnosti přepisů je shrnuta v tabulce 5.3. Dosahované výsledky jsou očekávatelné pro spontánní (mnohdy neformální) promluvy s obecně horší akustickou kvalitou. Proč však nejsou uvedeny i výsledky na dalších testovacích sadách i pro GMM-HMM systém (v práci navíc podrobněji diskutovaný)?

  Důležitou součástí ovlivňující přesnost přepisu je také použitý jazykový model (LM) popisovaný v 5. kapitole, jehož významnou složkou jsou Makoňovy spisy a manuální přepisy nahrávek. Zkoušel jste při tvorbě jazykových modelů pro tyto účely také pracovat s identifikovanými tématy definovanými v 2. kapitole? Dle textu jsou prezentované výsledky v tab. 5.3 získány s LM bez rozlišení témat.
- Jak bylo výše řečeno, zásadním přínosem této práce je metodika postupného zlepšování přepisů na bázi iterativních upgradů s postupně rostoucím množstvím manuálně přepsaných dat získaných korekcí automatických přepisů. Toto však v práci není podrobně dokumentováno. Autor uvádí přírůstky ručních přepisů na obr. 2.4, ale nikoliv jaký byl vývoj přesnosti přepisu s nárůstem trénovacích dat z Makoňova korpusu. Mohl by toto autor při obhajobě doplnit? Též by mě zajímalo, od kterého okamžiku bylo pracováno s DNN systémem? Předpokládám, že první přepisy byly získány bázi GMM-HMM.
- Jednoznačným přínosem předložené práce je také vytvořené Webové rozhraní s dostupným přepisem Makoňova korpusu, možnostmi fulltextového vyhledávání a editace přepisů. Autor se inspiruje jinými zpřístupněnými korpusy a programy pro anotaci signálů. S dostupnými WEBovými technologiemi pak vytváří finální aplikaci se zaměřením

na efektivní ruční korekci aktuálně dostupných přepisů či efektivní dostupnost odpovídajících audiosouborů. Detaily implementace neposuzuji, z uživatelského pohledu je na základě osobního vyzkoušení aplikace funkční a použití velmi srozumitelné.

- Nakonec bych ocenil i mnoho formální a organizační práce typu zaškolení anotátorů a organizace potřebných ručních korekcí přepisů Makoňova korpusu, kterou musel autor odvést. V této souvislosti považuji za významný přínos a velmi dobrý výsledek dosažení zmíněného rozsahu asi 110 hodin ručně přepsaných dat.

Publikační výstupy autora jsou slabší, publikoval pouze 4 práce související s tématem disertace. Publikace v prestižním časopise či na některé z nejprestižních mezinárodních konferencí (Interspeech, ICASSP) chybí. Významnější je publikace na konferenci Text, Speech, and Dialogue (zaindexované ve Web of Science), avšak ta je již z roku 2012 a jsou zde popisovány první výsledky realizovaného projektu. Nicméně vzhledem ke skutečnosti, že i další publikace byly prezentovány na mezinárodních fórech (byť spíše lokálního významu či bez standardního sborníku), lze tyto výstupy akceptovat jako potvrzení originálního přínosu práce autora, byť na hraniční resp. minimalistické úrovni.

Po formální stránce je text práce psaný obecně dobrou češtinou, místy je použitá terminologie resp. slovní zásoba trochu neobvyklá, např. termín "trifoném" místo obvyklejšího "trifón", výrazy "umenšení chybovosti", "svobodný program", apod. Členění kapitol by mohlo být vhodnější, související témata jsou občas popisovaná v různých kapitolách. Ve finální sazbě lze také nalézt několik formálních prohřešků, např.: časté přetečení tiskového zrcadla, což je spíše menší typografický prohřešek, avšak zvýrazněný tučnou vertikální čarou generovanou použitou šablonou; nejednotné či neúplné citace, časté chyby v malých a velkých písmenech v citacích, neúplné citace; umístění seznamu obrázků a tabulek na konec práce; ne acela optimální členění odstavců, které jsou často velmi krátké; apod. Uvedené prohřešky principiálně neznehodnocují hlavní sdělení práce, avšak jistě jej nevylepšují, působí rušivě a ukazují na kompletaci práce v časovém presu.

Na základě všech výše uvedených skutečností a i přes zmíněné výhrady lze konstatovat, že předložená práce přináší originální výsledky vědecké práce s konkrétními praktickými výstupy a považuji ji za disertabilní zejména díky její šíři a interdisciplinaritě. Obecným problémem, který se táhne napříč prací, je autorovo používání různých dostupných nástrojů pro realizaci dílčích kroků, a to často bez detailnějšího zaměření na principy jejich funčnosti. Nicméně chápu, že množství problémů, které musel autor řešit, bylo veliké, zasahovaly do různých oblastí, a mnohé by mohly dát na samostatnou disertační práci. Z tohoto hlediska je potom naopak jednoznačně pozitivním výsledkem, že autor dokázal za pomoci zmíněných nástrojů nakonec realizovat přepisy s dobrou přesností a dosáhnout jednoznačně konkrétního a hmatatelného výsledku, a to efektivního zpřístupnění Makoňova korpusu ve zmíněné WEBové aplikaci. Práci proto **doporučuji** k obhajobě za účelem získání vědecké hodnosti doktora na Matematicko-fyzikální fakultě Karlovy Univerzity.

V Praze dne 31. srpna 2020