POSUDEK NA DISERTAČNÍ PRÁCI

Téma práce: Automatická strukturalizace počítačem přepsaných mluvených dokumentů z multimediálních archivů

Doktorand: Ing. Marek BOHÁČ

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

Předložená práce ing. Boháče se věnuje problematice strukturalizace automaticky přepsaných mluvených dokumentů, která nachází v současné době četné aplikace při off-line zpracování rozsáhlých audioarchivů. Autorův výzkum je konkrétně vázán na zpracování rozsáhlého audioarchivu Českého resp. Československého rozhlasu, uvedená problematika je však aktuální a je intenzivně řešena i v širším mezinárodním měřítku, kde lze nalézt řešení podobných úkolů, ze kterých autor vycházel a na které ve své práci navazoval.

Cíle předložené disertační práce jsou přehledně uvedeny po úvodním popisu a přehledu stavu řešené problematiky ve 4. kapitole práce a lze je stručně shrnout následovně:

1. provést podrobnou rešerši existujících metod v oblasti zpracování rozsáhlých multimediálních archivů,
2. navrhnout a realizovat schémata strukturalizace zpracovávaných nahrávek,
3. na základě navržených vyhodnocovacích kritérií navržená schémata otestovat na vhodně vybraných testovacích datech, včetně analýzy dílčích bloků systému a vazeb mezi nimi,
4. a nakonec připravit reálné nasazení a vyhodnotit funkčnost v reálném provozu navrženého systému.

Výše stanovené cíle jsou jistě disertabilní a autor tyto cíle v předložené práci naplnil. Předložená práce představuje jednoznačný přínos v dané problematice na naší národní i mezinárodní úrovni a nejvýznamnější dosažené výsledky lze shrnout v následujících bodech.

- V první řadě autor naplňuje první cíl, kdy zpracoval stručnou a přehlednou rešerši aktuálního stavu problematiky (kapitola 3), která zahrnuje shrnutí přístupů k automatickému rozpoznávání spojité řeči, k detekci řečové aktivity či změn v nahrávce, ke klasifikaci charakteru nahrávky a především pak obsahuje popis existujících systémů pro zpřístupnění multimediálních archivů (MALACH, SpeechFind, SPRACH, RadioOranje, InForMedia, TaiwanNews).
- Přehled systémů pro zpřístupnění multimediálních archivů je potom doplněn popisem modulů a nástrojů vyvinutých pro strukturalizaci dokumentů na pracovišti doktoranda. Soubor všech dílčích modulů (od detekce řeči přes identifikaci řečníka, diarizaci záznamů, LVCSR a mnoho dalších až po doplnění interpunkce a výslednou strukturalizaci) je samozřejmě výsledkem mnohaletého vývoje na školícím pracovišti a pochopitelně přesahuje rozsah této práce. V této souvislosti zmiňuje autor velmi obecně svůj podíl na vývoji některých popisovaných modulů. Pro přesnější přehled o konkrétních autorových aktivitách by určitě bylo vhodné zmínit trochu přesněji, na vývoji jakých částí konkrétních modulů se přesně podílel.

- Hlavní přínos práce je jednoznačně v návrhu schémat strukturalizace dokumentu, což je obsahem kapitoly 6. Autor volí dva přístupy: strukturalizaci s izolovaným rozhodováním, kdy konkrétní informace je používána pouze v okamžiku jejího získání, a strukturalizaci s kumulovaným rozhodováním, kdy jsou nejprve shromážděny všechny potřebné informace a dílčí rozhodování bere informace z více zdrojů. Výhodou kumulovaného rozhodování je menší citlivost na chybu v jednotlivém dílčím kroku. Kromě rozdílného přístupu k rozhodování na základě výstupů dílčích modulů se oba navrhované systémy liší i v použití různých rozpoznávačů, tj. GMM-LVCSR vs. DNN-LVCSR. V této souvislosti bych se zeptal, jaké byly důvody použití dvou různých konfigurací rozpoznávače v navrhovaných strukturalizačních schematech? Bylo to dáno především historickým vývojem práce na celém projektu či použití DNN-LVCSR v systému s kumulovaným rozhodováním bylo motivováno ještě dalšími aspekty?
  Uvedené dva základní přístupy jsou ještě doplněny o postup strukturalizace dokumentu s dostupným přepisem, který nabízí úsporu strojového času při zpracování správně přepsaných úseků určených na základě ohodnocení zarovnání dostupného přepisu s odpovídající nahrávkou.
- Pro experimentální část práce i pro hodnocení systémů strukturalizace obecně má velký význam vytvoření evaluačních množin záznamů rozhlasových pořadů rozdělených podle doby pořízení, což v sobě zahrnuje rozdělení podle kvality záznamu, míry výskytu slovenského jazyka, míry spontaneity promluvy, apod.
- Co se týče dosahovaných výsledků experimentů, autor prezentuje velmi vysokou přesnost detekce bodů změny v nahrávce i pro data s nízkou akustickou kvalitou (cca 97-99%), v případě klastrování nahrávky je nutné zmínit vysokou přesnost VAD, vysokou přesnost identifikace mluvčího pro data standardní kvality (horší výsledky pro data nižší kvality jsou pochopitelná).
  Dosahovaná přesnost automatického přepisu je v navrhovaných systémech relativně vysoká, zejména u kvalitnějších záznamů (Acc cca 92-96%), dosahované nižší úspěšnosti (cca 65%) pro záznamy nižší kvality odpovídají aktuálně dosahovaným standardům. Při srovnávání LVCSR systémů na bázi GMM vs. DNN je prezentována nižší přesnost detekce neřečových událostí. V této souvislosti bych měl otázku, jaké jsou rozdíly v modelování neřečových událostí v GMM resp. DNN systému a co má největší vliv na horší výsledky detekce neřečových událostí pro DNN systémy?
  Pro úlohu doplnění interpunkce, kde autor analyzuje přesnost 2 navržených interpunkčních schémat využívajících informaci o prozodii promluvy neřečových událostech v promluvě resp. informaci o celkové strukturalizaci nahrávky, je dosahovaná přenost v obou případech opět relativně vysoká a blíží se k 90%. Autor zmiňuje i kladnou zpětnou vazbu od uživatelů týkající se nepříliš negativního vnímání lehce nadbytečné interpunkce.
- Pro reálné nasazení je zásadní i informace o výpočetní náročnosti strukturalizace, autor prezentuje RT faktor 3,85 pro strukturalizaci s izolovaným rozhodováním resp. RT faktor 3,2 pro variantu s kumulovaným rozhodováním.
- Z aplikačního hlediska je nutné vyzdvihnout skutečnost, že navrhované metody nejsou analyzovány pouze na menším množství vybraných evaluačních dat, ale byly prakticky nasazeny ke zpracování části archivu Českého rozhlasu a dosažené výsledky jsou zpřístupněny a již používány pro vyhledávání v rozhlasovém archivu na

adrese http://naki.ite.tul.cz. Autor ve své práci prezentuje i výsledky zpětné vazby od uživatelů, specifikující jaké chyby jsou pro uživatele více či méně zásadní.

Po formální stránce a z hlediska sazby a grafické úpravy je předložená práce na velmi dobré úrovni. Několik málo typografických chyb nesnižuje zásadně celkový dojem z práce. Členění je logické a text je psaný velmi pěknou češtinou (malou méně podstatnou výhradu bych měl snad jen k používání termínu frame, zejména pak ve skloňovaných česko-anglických tvarech framy, framu, apod.). Častější formátovací chyby jsou v přehledu literatury (zejména z hlediska používání velkých a malých písmen).

Originální přínos autora v dané oblasti potvrzují autorovy publikace, mezi kterými jsou nejvýznamnější publikace v časopisech EURASIP Journal on Audio, Speech, and Music Processing (první autor) a Journal of Multimedia. Přínosné jsou samozřejmě i publikace na mezinárodních konferencích, workshopech a seminářích, zejména pak na Text, Speech, and Dialogue, Telecommunications and Signal Processing, a dalších. Celkem autor publikoval 17 prací, kde v 11 případech je prvním autorem.

Na základě výše uvedených skutečností lze tedy jistě konstatovat, že předložená práce popisuje originální výsledky vědecké práce, a proto ji **doporučuji** k obhajobě za účelem získání vědecké hodnosti doktora na Technické univerzitě v Liberci.

V Praze dne 13. května 2016

Oponentský posudek disertační práce Ing. Marka Boháče

„Automatická strukturalizace počítačem přepsaných mluvených dokumentů z multimediálních archivů"

Předložená disertační práce řeší velmi komplexní problematiku z oblasti automatického zpracování řeči a to konkrétně segmentaci audio nahrávek. Segmentace je zpracovávána na několika úrovních - řeč/hudba/neřeč, řečník, jazyk, šířka přenosového pásma, formátování textu a interpunkce. Tyto činnosti jsou provedeny pro potřeby vyhledávání v archivních nahrávkách Českého rozhlasu v rámci projektu NAKI[1].

Hlavním cílem práce je návrh a implementace modulů pro segmentaci řečových přepisů, což zahrnuje adaptaci a integraci řady existujících modulů do funkčního celku. Autor zde srovnává dva základní přístupy: 1) využití výsledků samostatných modulů; 2) rozhodování na základě spojení výsledků dostupných modulů. Zde je evidentní, že první přístup (významně jednodušší) musí poskytovat horší výsledky, než přístup druhý, který využívá komplexnější informaci. Proto bych se zde chtěl zeptat autora, proč se rozhodl oba přístupy srovnávat, když je výsledek takového srovnání celkem předvídatelný?

1. **Význam disertační práce pro obor**

   Jak sám autor uvádí, již existuje řada systémů, které se zabývají podobnou problematikou. Nicméně jako jeden přínos autora považuji „dotažení" zpřístupnění audio nahrávek v podobě přehledné vizualizace výsledného dokumentu (dokument + metadata) s možností vyhledávání.

   Dále autor navrhuje a srovnává (jak již bylo uvedeno výše) dvě schémata, izolované a komplexní (kumulované) schéma pro řešení segmentace nahrávek.

   Za největší přínos považuji nové metody pro doplňování větné interpunkce. Experimenty ukázaly, že navrhované přístupy úspěšně doplňují interpunkční znaménka do vět a zpravidla překonaly testované referenční metody.

   Autor také navrhuje nové evaluační metriky. Jedna z navrhovaných je $accuracy_{M2N}$ (viz eq. 7.9). Tato metrika považuje za správně rozpoznaná slova i v případě přidání/smazání bílých znaků. Dále považuje za správně rozpoznaná i slova s chybnou koncovkou. Další metrika je vážená F-míra ($F-measure_w$, je tedy vážena i přesnost a úplnost). Tato míra započítává různé chyby detekce změn v nahrávce různou váhou. Potřebnost těchto metrik není bohužel dostatečně zdůvodněna. Vysvětlete prosím potřebnost Vámi navržených metrik (oproti klasickým ACC a $F-m_1$).

2. **Postup řešení, použité metody a splnění stanovených cílů**

   Jak jsem již uvedl výše, autor v předkládané práci řeší velmi komplexní problematiku a je nucen využívat/spojovat řadu dalších modulů (LVCSR[2], segmentace řeč-neřeč, diarizace nahrávky apod.). Je proto velmi složité vše popsat vyčerpávajícím způsobem. Autor se rozhodl stručně popsat všechny kroky nutné pro realizaci kompletního systému. To je vhodné z úhlu pohledu čtení práce nepříliš erudovaným

[1] Projekt Ministerstva kultury ČR: DF11P01OVV013

[2] Rozpoznávání spojité řeči

čtenářem (i ten pochopí celkovou problematiku). Nicméně díky této volbě nejsou související studie popsány dostatečně (pouze 14 stran textu). V této práci je popisována řada úloh, které pro ní nejsou příliš důležité, jako např. ASR[3] (4 str. textu). Autor pak popisuje klíčové související práce na pouhých čtyřech stranách disertace (str. 28-30), což se mi zdá celkem povrchní. Proto je pak velmi složité hledat, co autor navrhl nového oproti state-of-the-art.

Autor použil pro řešení problematiky odpovídající metody.

Na základě předložené práce se zdá, že autor splnil všechny vytyčené cíle. Nicméně nikde jsem nenašel, jaké cíle byly stanoveny v rigorózní práci. Žádám proto studenta o objasnění této skutečnosti.

3. **Využití výsledků práce v praxi**

   Za **největší přínos** této práce považuji její praktické využití. Řada disertací sice navrhuje nové velmi sofistikované metody/modely, které však bohužel nikdy nejsou využity v praxi. Předkládaná práce sice (podle mého názoru) nenavrhuje příliš nové přístupy, spíše využívá/adaptuje/spojuje existující osvědčené metody pro danou úlohu, nicméně vzniklý systém je plně funkční (viz `https://naki.ite.tul.cz`). Z praktického hlediska je proto přínos disertační práce velký.

4. **Formální stránka práce**

   Disertační práce je napsána v českém jazyce a skládá se z devíti kapitol a příloh. Je psána srozumitelnou formou a je dobře strukturována. Po jazykové stránce je práce na slušné úrovni, obsahuje jen několik překlepů a chybějících slov ve větách. Obsahuje několik drobných formálních prohřešků: číslování práce (první číslovaná strana by měla být úvod, zde je to str. 15); některé zkratky nejsou definovány při prvním použití (např. str. 35 - CI-RNN-HMM); některá tvrzení by bylo vhodné doplnit referencemi (např. str. 28, sec. 3.1.3).

5. **Publikační aktivita studenta**

   Autor publikoval výsledky své práce ve dvou časopisech a v patnácti článcích na mezinárodních konferencích. Tento počet převyšuje požadavky kladené na studenta doktorského studia. Důležitost publikací je potvrzena **dobrým citačním ohlasem** (7 x DB Scopus). Na druhou stranu v uvedeném seznamu postrádám prestižní konference typu *A a A** z oblasti automatického zpracování řeči (např. Interspeech apod.).

6. **Připomínky / Dotazy**

   - V práci jsou na některých místech (např. sec. 7.2 - vyhodnocovací metriky) popsány některé skutečnosti, které by měl zasvěcený čtenář již znát. V těchto místech by byl odkaz dostatečný.
   - V práci srovnáváte dva akustické dekodéry, kde první je založen na směsi Gaussových funkcí (LVCSR-GMM) a druhý pak na hlubokých neuronových sítích (LVCSR-DNN). Při tomto srovnání je u LVCSR-GMM používána adaptace na mluvčího, ale v případě LVCSR-DNN nikoliv. Jak můžete srovnávat oba modely v rozdílných konfiguracích? Vysvětlete prosím.

[3] Automatické rozpoznávání řeči

- V sekci 5.3.2 popisujete konfiguraci a topologii Vaší DNN[4]. Volba parametrů této sítě bohužel nikde není zdůvodněna. Vysvětlete prosím volbu parametrů (počet vrstev, počet neuronů ve skrytých vrstvách, velikost okolí zkoumaného framu apod.) této sítě.
- V sekci 7.3 (str. 83 dole) máte nastaveny váhy w_H a w_S na 1 a 0,7. Zdůvodněte prosím tuto volbu.

7. **Shrnutí**

Na základě prostudování předložené práce konstatuji, že disertační práce pana Ing. Marka Boháče splňuje podmínky samostatné vědecké práce a obsahuje řadu původních autorem publikovaných výsledků. Disertační práci proto **DOPORUČUJI** k obhajobě.

V Plzni dne 22. května 2016

doc. Ing. Pavel Král, Ph.D.
Katedra informatiky a výpočetní techniky
Západočeská univerzita v Plzni

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky
②

[4] Hluboká neuronová síť