POSUDEK NA DISERTAČNÍ PRÁCI

Téma práce: Speaker Diarization

Doktorand: Ing. Marie KUNEŠOVÁ

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

Předložená práce ing. Kunešové se věnuje problematice diarizace řečníků. Tato úloha nalézá uplatnění v úlohách přepisu souvislé řeči, kdy rozpoznávané promluvy neobsahují řeč pouze jednoho mluvčího, např. automatizovaný přepis rozhlasového či televizního zpravodajství, přepisy konferencí, telefonních konverzací, apod. Aktuálnost řešení dané problematiky je dána mnoha problémy, které přinášejí uvedené aplikace v reálných podmínkách či za zhoršené kvality nahrávek a které stále nejsou zcela dostatečně dořešeny. To přináší nepochybně prostor pro disertabilní výzkum.

Obecné cíle disertační práce jsou uvedeny hned v úvodní kapitole a později upřesněny v kapitole 6 po popisu současného stavu řešené problematiky. Lze je stručně shrnout následovně:

1. popsat a srovnat známé a používané metody offline a online diarizace,
2. vybrané metody implementovat v rámci tvorby nového diarizačního systému, metody experimentálně ověřit a navrhnout jejich modifikace a vylepšení,
3. navrhnout vhodný detektor překrývající se řeči a zvýšit tak přesnost výsledné diarizace mluvčích.

Takto stanovené cíle jsou disertabilní, autorka tyto cíle v předložené práci naplnila a již z prvních odstavců a formulací v textu je zřetelný její nadhled nad danou problematikou. V průběhu svého doktorského studia musela řešit bouřlivý posun výzkumu ve studované oblasti, a to zejména v oblasti reprezentace mluvčích od technik na bázi GMM aktuálních v době zahájení svého doktorského studia, přes pozdější standardy na bázi i-vektorů směrem k technikám využívajícím hluboké neuronové sítě (DNN, x-vektory). Stejně jako v jiných oblastech zpracování řečového signálu se přístupy na bázi DNN postupně staly častěji používaným standardem pro reprezentaci mluvčích či promluv. S tímto problémem se autorka vyrovnala a nejvýznamnější přínosy předložené práce bych shrnul v následujících bodech.

- Práce předkládá rozsáhlý přehled používaných metod diarizace mluvčího. V kapitolách 2-5 autorka nejprve definuje teoretické základy, následně podrobněji popisuje používané off-line a on-line systémy se všemi dílčími kroky diarizace jako extrakce příznaků, VAD, segmentace, shlukování a resegmentace. Všechny dílčí kroky autorka s nadhledem popisuje a nakonec uvádí potenciální problémy, na základě kterých zpřesňuje konkrétní cíle svého disertačního výzkumu. V kapitole 7 pak autorka navazuje na obecný popis rozsáhlým srovnáním výsledků dosahovaných jinými autory v úlohách off-line i on-line diarizace aplikovaných na různé typy vstupních dat. Oceňuji rozsah tohoto srovnání, ale i vždy stručný a velmi výstižný popis každé jednotlivé techniky, který dává přehledně rámcovou představu o principu jednotlivých srovnávaných technik.
- V experimentální části autorka popisuje realizované experimenty s implementovanými systémy pro off-line resp. on-line diarizaci. Trochu nadbytečně rozsáhlá se mi zdá první

část věnovaná technikám na bázi GMM, byť se jedná o první významné výsledky autorčina výzkumu. Je sice prezentováno zajímavé srovnání diarizace pro různé typy dat, ale zestručnění této části by práci jistě neuškodilo.

Autorčino následné zacílení experimentální části práce na techniky na bázi i-vektorů považuji za pochopitelné a správné. Použití těchto technik v diarizačních systémech může být stále výhodné díky jejich jednoduchosti a menším nárokům na množství trénovacích dat ve srovnání se systémy na bázi DNN.

K této části bych měl dotaz týkající se LFCC příznaků, které v navrhovaných systémech používáte namísto častěji používaných MFCC. Zmiňujete prezentované lepší výsledky pro rozpoznávání ženských hlasů, ale také skutečnost, že všeobecně tento přístup příliš rozšířený není. Jestli jsem tomu dobře rozumněl, tak LFCC jsou používány ve všech popisovaných systémech, kde se pracuje na úrovni příznaků s kepstrem. Nicméně volba LFCC místo MFFC není podložena konkrétními výsledky v experimentální části. Je zlepšení opravdu tak významné? Není případně nevýhodné pracovat s odlišnými příznaky pro diarizaci a případné následné rozpoznávání obsahu, kde je asi vhodnější spíše použití MFCC?

- Významným přínosem potvrzující autorčiny schopnosti a přínosy jejího výzkumu je participace v týmu, který se úspěšně účastnil dvou evaluačních kampaní DIHARD Speaker Diarization Challenge, viz kap. 9.4. V roce 2018 v rámci DIHARD I se tým s autorkou umístil na výborné 5. pozici ze 14, což potvrzuje kvalitu popisovaného systému off-line diarizace na bázi i-vektorů. V rámci DIHARD II v roce 2019 se tým s autorkou umístil 11. místě z 20 účastníků, původní i-vektorový systém byl použitý jako základ modifikace pro práci s x-vektory resp. pro kombinovaný xi-vektorový systém. Kromě srovnání v rámci DIHARD challenge, je uvedeno i srovnání s baseline systémem v široce používaném KALDI toolkitu, kde bylo dosaženo lepších výsledků ve srovnání s referenčním systémem.
- Přínosem disertační práce je i navrhované řešení problému detekce překrývající se řeči, což si autorka zvolila jako jeden z konkrétních dílčích cílů svého výzkumu. Nejprve autorka uvádí v kap. 8 shrnutí dříve publikovaných řešení a v kap. 9.5 pak popisuje své řešení na bázi CNN. Klíčovým krokem je kompletace dat pro trénování dané sítě, kde autorka volí vytvoření souboru dat s uměle generovaným překryvem ze signálů z databází TIMIT resp. VoxLibri s doplněním o augmentovaná data (data s přidaným šumem resp. dozvukem). Kompletace trénovacích resp. testovacích dat pro účely detekce překrývající se řeči je vzhledem k menší dostupnosti dat tohoto typu jistě významným vedlejším produktem prezentované práce, jak sama autorka zmiňuje.

  Natrénovaná síť je testovaná opět na uměle vytovřených datech, ale také na datech s překryvem z jiných korpusů (AMI, SSPNet). Autorka orientačně srovnává své dosažené výsledky s výsledky jiných autorů (přesné srovnání není možné z důvodů velké různosti testovacích dat) a lze konstatovat, že dosahuje lepších nebo srovnatelných výsledků. Měl bych jen doplňující otázku, zda máte alespoň první konkrétní výsledky, jak použití Vámi navrhovaného detektoru snižuje celkovou chybu diarizace?
- Nakonec bych zmínil neobyčejně rozsáhlý seznam pramenů, které reprezentují ve veliké šíři aktuální stav problematiky diazrizace řečníků v celosvětovém měřítku a které autorka cituje v řešeršní části práce i při srovnání vlastních výsledků s výsledky dosahovanými jinými autory. Uvedený seznam i konkrétní citované výsledky dokazují,

že práce autorky není odtržená od aktuálního celosvětového směřování výzkumu ve studované oblasti.

Po formální stránce má předložená práce logické a přehledné členění. Výklad je velmi srozumitelný a i grafické zpracování je precizní. Je psaná v anglickém jazyce na velmi dobré úrovni, nenašel jsem zásadní gramatické či stylistické chyby.

Celkový přínos práce je neodiskutovatelný a originální přínos autorky v dané oblasti potvrzuje i její publikační činnost. V práci je zmíněno 11 publikací (v 5 případech jako první autor), kde nejvýznamnější jsou publikace na prestižních konferencích Interspeech, nicméně i většina dalších prací byla publikována na kvalitních mezinárodních konferencích se sborníky zaindexovanými ve WoS.

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

Kromě výše zmíněných otázek v textu bych měl do diskuse ještě následující 2 otázky:

- Ve své práci jste zpracovala rozsáhlý přehled technik diarizace, řadu technik jste experimentálně testovala, participovala jste v mezinárodních srovnávacích kampaních. Který z přístupů s hlubokými neuronovými sítěmi byste považovala v současné době a na základě Vašich zkušeností za nejperspektivnější resp. kde byste viděla největší prostor pro případná vylepšení aktuálních systémů.
- Z aplikačního hlediska je vždy významným přínosem nasazení popisovaných systémů v reálném provozu. Zmiňujete různé úlohy, některé experimenty byly realizovány se záznamy jednání poslanecké sněmovny parlament ČR. V jakých systémech realizace přepisů případně on-line titulkování byly případně Vámi navržené systémy použité?

V Praze dne 11. října 2021

Ing. Petr Červa, Ph.D.
Technická univerzita v Liberci
Ústav informačních technologií a elektroniky
Studentská 2, 461 17 Liberec

Oponentský posudek disertační práce Ing. Marie Kunešové:

„Diarizace řečníků"

Disertační práce Ing. Marie Kunešové se týká problematiky automatické diarizace mluvčích v off-line i on-line režimu. Je členěna do celkem deseti kapitol. První čtyři obsahují podrobný popis teoretického pozadí úlohy a aktuální přehled existujících metod. V následující páté kapitole jsou detailně shrnuty hlavní úskalí diarizace mluvčích. V kapitole šest je pak stanoveno pět hlavních cílů práce. První dva z nich zahrnují vytvoření přehledu existujících metod diarizace a identifikaci úskalí v rámci této úlohy. Třetí cíl disertační práce, spočívající v porovnání publikované úspěšnosti existujících metod na standardních datových sadách, je obsahem kapitoly sedm. Ta dále popisuje metriky pro vyhodnocení úspěšnosti diarizace a úlohu detekce překrývající se řeči. Čtvrtý cíl práce pak představuje implementace některé z existujících diarizačních metod do nového diarizačního systému a návrh nové nebo vylepšení stávající diarizační metody. Posledním pátým cílem práce je zlepšit výsledky diarizace řešením některého z popsaných úskalí. Konkrétně je vybrána problematika detekce překrývající se řeči. Postup řešení uvedených posledních dvou cílů práce je popsán v kapitole devět, kde jsou také prezentovány dosažené experimentální výsledky.

Stanovené cíle práce jako takové jsou podle mého názoru dizertabilní. Uvedené členění práce, systematičnost postupu, přehlednost i formální úpravu a jazykovou úroveň předložené práce hodnotím kladně.

Z hlediska stanovených cílů práce platí, že první tři cíle mají spíše rešeršní charakter a jsou splněny v příslušných kapitolách jedna až pět a dále pak v kapitole sedm. Z hlediska hodnocení významu pro obor, vlastního přínosu předkladatelky i její publikační činnosti považuji za nejdůležitější čtvrtý a pátý cíl práce, protože oba tyto cíle mají výzkumný charakter.

V rámci řešení čtvrtého cíle předkladatelka nejprve využila a rozšířila on-line metodu diarizace založenou na testu poměru věrohodností s využitím GMM modelů. Jak si je předkladatelka sama vědoma, jedná se o typ modelů, který měl sice význam v době jejího nástupu do doktorského studia, ale z dnešního pohledu je již překonaný. Tomu nakonec odpovídají i dosažené výsledky, které jsou vyhovující pouze na některých datových sadách. V další části řešení se pak předkladatelka zabývala zejména úlohou off-line diarizace s využitím novějších typů modelů založených na i-vektorech nebo aktuálnějších x-vektorech či kombinaci obou přístupů. Pozitivně zde hodnotím zejména skutečnost, že se předkladatelka podílela na návrhu dvou diarizačních systémů pro mezinárodní evaluaci off-line diarizace. Příprava systémů na tyto evaluace byla jistě pracná a náročná. Zároveň je plusem, že tyto systémy dosáhly solidních výsledků a byly publikovány na dvou konferencích Interspeech, které jsou v rámci oboru považovány za kvalitní. Protože se ale jednalo o týmovou účast a hlavním autorem dvou zmíněných souvisejících publikací je jiný člen řešitelského týmu, bylo by dobré, když by předkladatelka během prezentace v rámci obhajoby svoji roli v tomto týmu, a tedy i při vývoji těchto diarizačních systému, blíže specifikovala. Práce pak v této části obsahuje i experimenty a přístupy rozšiřující funkcionalitu zmíněných off-line systémů i pro práci v on-line režimu. Vy-

hodnocení je ale provedeno spíše jen v rovině postupného ověření na off-line nahrávkách z různých databází - chybí například zhodnocení reálné implementace z hlediska latence a s využitím automatického detektoru řeč/neřeč na různých datových proudech, kde má on-line zpracování největší smysl. Celkově potom čtvrtý cíl disertační práce považuji také za splněný s tím, že za významnější považuji výsledky dosažené v off-line realizaci než v on-line režimu.

Konečně při řešení posledního pátého cíle disertační práce předkladatelka navrhla metodu detekce překrývající se řeči založenou na využití konvoluční neuronové sítě s vyhlazováním výstupu, kterou lze použít pro on-line i off-line režim. Metoda dále využívá uměle připravená a augmentovaná trénovací data. Dosažené výsledky mají zajímavý potenciál zlepšit přesnost diarizace a byly publikovány. I poslední cíl práce proto považuji za splněný.

Celkově pak práci Ing. Marie Kunešové **doporučuji** k obhajobě. Publikační činnost předkladatelky přitom hodnotím celkově jako uspokojivou. I vzhledem k délce studia ale přeci jen trochu postrádám hlavní autorství příspěvku na konferenci typu Interspeech nebo ICASSP nebo hlavní autorství či alespoň spoluautorství článku v časopise.

Do diskuze navrhuji dvě otázky:

1) Navržený systém pro evaluaci DIHARD II na vstupu detekuje typ vstupního korpusu a podle toho pak provádí diarizaci buď vlastním systémem s vytuněnými parametry přímo pro detekovaný typ korpusu nebo systémem Kaldi s obecnými parametry, který je v předložené práci označen za univerzálnější. Uvedený postup s využitím systému Kaldi je částečně pochopitelný z hlediska snahy dosáhnout v rámci evaluace co nejlepších výsledků. Je ovšem problematický z hlediska hodnocení vlastního přínosu navrženého systému jako takového v rámci disertační práce. Nabízí se proto otázka, jaké skóre by navržený diarizační systém dosahoval pouze s jedním globálním nastavením parametrů?
2) Součástí navrženého diarizačního systému se segmentací pomocí fixních oken je i finální modul resegmentace, který provádí redefinici nalezených hranic jednotlivých mluvčích pomocí GMM modelu s následným vyhlazením gaussovským oknem. Podobný způsob vyhlazení plovoucím oknem se pak používá i v rámci metody pro detekci překrývající se řeči. Byly v některém z těchto případů prováděny experimenty i s jiným vyhlazováním, například pomocí Viterbiho dekodéru nebo třeba váženého převodníku s konečným počtem stavů?

V Liberci, 26.8.2021

Ing. Petr Červa, Ph.D.