OPONENTSKÝ POSUDEK DISERTAČNÍ PRÁCE

Kandidát: Mgr. David Klusáček, MFF UK
Název: New methods in statistical speech recognition
Recenzent: Doc. Dr. Ing. Jan Černocký, FIT VUT v Brně

Předložená disertační práce Mgr. Klusáčka má 226 stran včetně příloh a obsahuje 10 kapitol.
Tento posudek se v sekci 1 nejprve zabývá jednotlivými kapitolami včetně poznámek a v sekci 2 obsahuje zhodnocení technické stránky práce. Sekce 3 hodnotí formální stránku a sekce 4 obsahuje závěr, celkové zhodnocení práce a doporučení. Posudek je doplněn otázkami k obhajobě.

1 Obsah práce a poznámky ke kapitolám

Práce se věnuje velmi aktuální problematice rozpoznávání řeči. Její jádro je v exaktním matematickém popisu technik zpracování signálu, strojového učení a obsahuje také rozsáhlý souhrn současných poznatků o lidském slyšení. V závěru práce je definováno nové schéma pro extrakci příznaků z řečového signálu.

První kapitole uvádí do problematiky rozpoznávání řeči a obsahuje jeho krátkou historii a kritiku obvyklých nešvarů jeho výzkumu, jako je nerespektování state-of-the-art. Obsahuje také shrnutí požadavků na rozpoznávací systém.

Kapitola druhá uvádí do pravděpodobnostního rozpoznávání řeči a definuje základní bloky systému jako je feature extraktor (FE), akustický model, jazykový model a dekodér. Zatímco u FE je popsáno obvyklé schéma a jazykový model je rovněž standardní, v případě akustického modelování definuje autor model pracující s množinou diskrétních akustických jednotek, což v dnešní době s dominujícími Markovovými modely se spojitými rozloženími pravděpodobnosti není úplně obvyklé. Pro uvedený akustický model jsou odvozeny základní algoritmy pro dekódování a trénování. Na závěr je zmíněno diskriminativní trénování, zde bych doporučoval podívat se na novější práce než [46] a [67] z let 1986 a 1991, především na disertaci Dana Poveyho a na jeho následné práce.

Kapitola 3 je první ze sady učebnic obsažených v disertační práci. Zavádí rigorózním způsobem základní pojmy z pravděpodobnosti a teorie informace, a jde až ke kódům použitelným pro zašuměný kanál. Kapitola je psána velmi fundovaně a definovaný matematický formalismus je dokonalý, otázkou je, nakolik má být cílem disertace pečlivý popis teorie a nakolik algoritmický či experimentální posun. K této otázce se vrátím v závěru.

Kapitola čtvrtá obsahuje souhrn poznatků o zpracování signálu, opět podaný vyčerpávajícím způsobem ve formě učebnice. Z hlediska využitelnosti této kapitoly pro praktiky bych přivítal pečlivější práci s frekvencemi – na mnoha místech není odlišena běžná frekvence v Hz od normované, rozdíly jsou někdy i v používaných znaménkách (Fourierova řada – (156) a dále). Kapitola se rozbíhá do značné šíře a opět není zcela jasné, zda jsou některé sekce (např. dvousměrných filtrech, implementaci filtrů, či o číslicově-analogovém převodu) používány dále v práci.

Kapitola 5. se zabývá tradičními FE: MFCC a PLP, a nepomíjí velmi důležitou normalizaci koeficientů pomocí odečítání průměru. Zde bych doporučil poznámku o zásadní důležitosti kvalitního detektoru řečové aktivity (VAD) – špatný VAD může normalizaci naprosto rozbourat. U schématu PMVDR-BISN by byla na místě opatrnost při citování „the most powerful one“ – oba zmiňované články jsou mírně letité (2004) a bylo by dobré uvést typ dat a systémy, se kterými byly tyto parametry srovnávány. U podobných proklamací se navíc často jedná o závěr ze srovnání s nekvalitními baseliny, což autor v úvodu sám kritizuje. Rovněž u TRAPS bych doporučil poohlédnout se po novějších publikacích – od disertace Sangity Sharma uběhla řada let a TRAPS se mezitím staly důležitým komponentem FE velkých rozpoznávacích systémů.

Kapitola pokračuje souhrnem poznatků o lidském slyšení. Tuto část práce velmi oceňuji, kompilace (často protichůdných) článků, konsultací s odborníky a formulace vlastních hypotéz je podle mého názoru velmi zdařilá.

Šestá kapitola definuje nový přístup k FE: NUFIBA. S argumentací na začátku kapitoly, která pokládá FE za nejslabší část řetězce ASR, by se dalo polemizovat, ale je jisté, že pokrok v FE je potřebný. Na počátku kapitoly jsou definovány požadavky na FE a jsou vhodným způsobem využity poznatky o slyšení z předcházející kapitoly, v sekci 6.4 je pak NUFIBA FE skutečně popsán. Je škoda, že tato sekce neobsahuje blokové schéma celého procesu extrakce – usnadnilo by její pochopení. Parametry banky filtrů i následných Hilbertových filtrů byly numericky optimalizovány, což je v pořádku, škoda, že jako kriteriální funkce byly použity odezvy filtrů a ne „konečný cíl" rozpoznávače, tedy WER nebo PER. Další sekce popisují zajímavý návrh na konstrukci inverzního filtru (pro odstranění šumu a impulsní charakteristiky rušení) pomocí zaměřování na glotální pulsy v řeči. Jedná se o zajímavý přístup a je škoda, že nebyl dále experimentálně vyhodnocen. Kapitola končí definicí akustické abecedy pro diskrétní HMM modely.

Kapitola 7 obsahuje návrh akustického modelu, který pracuje nad diskrétní abecedou. Pracuje s difóny, což je opět mimo „mainstream", kde jsou preferovány trifóny, případně delší kontexty.

Kapitola 8 pojednává o jazykovém modelování a zaměřuje se na diskriminativní tvorbu tříd na class-based LM. Jedná se o zajímavou práci rozšiřující autorovu publikaci [44], je ovšem poněkud mimo hlavní zaměření práce a pravděpodobně není ani použita v experimentální části.

Kapitola 9 obsahuje návrh experimentů, které do doby psaní posudku zůstaly nedokončeny. Experimenty jsou plánovány na databázi Switchboard pouze v režimu rozpoznávání fonémů. Kapitola 10 o rozsahu necelé stránky pak uzavírá práci.

2 Zhodnocení technické stránky práce

Práce dokládá, že kandidát pronikl do problematiky, byl schopen nastudovat velmi složitou teorii a popsat ji koherentním matematickým formalismem, který v obvyklých inženýrských pracích není obvyklý. Otázkou je, zda má být cílem disertační práce právě taková studie nebo definice, pečlivé testování a diskuse výsledků nových postupů (nebo alespoň nového postupu). V této souvislosti mě poněkud děsí věta „…letting me explore whichever direction in my reseaerch I liked to", která předznamenává následující poznámky:

- V práci je definován mocný a značně široký teoretický aparát, který ale je ale z velké části nepoužitý v experimentální práci.
- Práce neobsahuje experimentální výsledky a jejich diskusi.
- Kandidát postupoval metodou „všechno změnit" (FE, akustický model), která nedovoluje systematickou práci a srovnávání se state-of-the-art. Výsledný systém pak funguje nebo nefunguje, produkuje nějakou chybovost, ale je velmi složité analyzovat, kde přináší co dobrého nebo špatného. Doporučoval bych postupovat systematicky od zavedeného systému (např. standardní MFCC, standardní modely), měnit postupně jednotlivé bloky (banka filtrů, odšumování, dereverberace, jiné modely, atd.) a pečlivě studovat účinky. Na závěr je pak akceptovatelné prostudovat exotičtější systémy, kde je změněno více „klasických" funkčních bloků.
- Experimenty nejsou dokončeny, ale pokud se práce zabývá FE a robustností, bylo by vhodné poohlédnout se po standardních datových setech, s množstvím publikovaných prací i analýz, např. AURORA.

3 Formální stránka

Formálně je práce velmi pěkně zpracována, je psána velmi kvalitní angličtinou s minimem překlepů a chyb a je velmi čtivá. Matematická sazba je kvalitní, přimlouval bych se pouze za značení vektorů a matic tlustou sazbou tak, jak je to obvyklé v anglosaské literatuře: výsledek je pak mnohem srozumitelnější. Obrázky jsou z velké části rukou kreslené, což není na závadu, pouze v sekci o slyšení bych doporučoval reprodukovat obrázky ze standardní literatury s patřičným uvedením zdroje, přece jen by byly názornější.
„Literární" prezentaci výsledků (tabulky, grafy) nejsem schopen zhodnotit, protože práce až na charakteristiky banky filtrů a několik spektrogramů výsledky neobsahuje.

Literatura použitá v práci by mohla být mladší (kromě standardních učebnic) a její seznam by měl být podstatně pečlivěji zpracován – u mnoha referencí není uveden zdroj.

4 Celkové zhodnocení

Předložená disertační práce prezentuje na jedné straně téměř dokonalé dílo: pečlivý matematický popis a rozbor technik zpracování signálu, teorie informace a rozpoznávání řeči, na druhé straně trpí značnými nedostatky v experimentální části, který je nesystematický a především nejsou k disposici experimentální výsledky, srovnání s klasickými technikami a diskuse. Z tohoto pohledu je diskutabilní, zda práci ponechat její původní název „New methods in statistical speech recognition" – zajímavá nová technika FE je sice navržena, ale měla by být dokumentována její funkčnost, jinak by mohl být za „novou techniku" označen libovolný soubor funkčních bloků.

Neměl jsem k disposici seznam publikovaných prací doktoranda, z databáze ÚFAL MFF vyplývá, že je autorem/spoluautorem dvou článků na mezinárodních konferencích, techniky pro FE popsané v disertační práci zřejmě publikovány nebyly. Poslední publikace je z roku 2007. Nesdílím názor, že k obhajobě je nutná časopisecká publikace, rozsah publikací kandidáta ovšem považuji za minimální akceptovatelný.

Celkově mám z disertační práce dojem značné nevyrovnanosti: excelentní pasáže jsou střídány kapitolami s minimálním rozsahem, velkou mírou nesystematičnosti, ignorováním standardních vědeckých postupů a nedostatečnou orientací ve state-of-the-art. Teoretické části práce mají potenciál vřelého přijetí vědeckou komunitou.

Závěr

Teoretická část práce **překračuje obecně uznávané požadavky k udělení akademického titulu Ph.D., doporučuji proto tuto práci k obhajobě. Celkové hodnocení a rozhodnutí o udělení titulu Ph.D: ponechávám na úvaze komise.**

Pro obhajobu navrhuji následující otázky:

1. Presentujte experimentální výsledky (byť i částečné) a jejich diskusi.
2. Uveďte, zda by bylo možné použít navržený FE se standardním CD-HMM systémem pro rozpoznávání.
3. Uveďte, proč trénování HMM probíhá na párech slov (sekce 9.2.1) a ne na celých promluvách ?

V Brně, 31.8.2012

Doc. Dr. Ing. Jan Černocký
Speech@FIT, Ústav počítačové grafiky a multimédií
Vysoké učení technické v Brně, Fakulta informačních technologií
Božetěchova 2, 612 66 Brno
Tel: +420 5 41141284 Fax: +420 5 41141270, mailto:cernocky@fit.vutbr.cz
http://www.fit.vutbr.cz/~cernocky http://www.fit.vutbr.cz/speech