ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA ELEKTROTECHNICKÁ
Katedra teorie obvodů
TECHNICKÁ 2, 166 27 PRAHA 6
TEL.: +420 224 352 236

Oponentský posudek na disertační práci

Téma:	Rozpoznávání řeči pacientů po totální laryngektomii komunikujících pomocí elektrolarynxu
Doktorand:	Ing. Petr Stanislav, Západočeská univerzita v Plzni, FAV
Oponent:	Prof. Ing. Roman Čmejla, CSc., ČVUT v Praze, FEL

Význam práce pro obor

Rehabilitace hlasu pacientů po laryngektomii je aktuální a společensky velmi důležité téma. Předložená práce se zabývá problematikou rozpoznávání řeči pacientů, kteří podstoupili totální laryngektomii a k produkci hlasu využívají elektrolarynx. Předložená práce, která je svým zaměřením ojedinělá, přináší nové poznatky pro výzkum v oblasti rozpoznávání mluvené řeči ve ztížených podmínkách.

Řešení problému, použité metody a splnění určeného cíle

Cílem práce byl výzkum a ověření možností využití systémů automatického rozpoznávání řeči k realizaci systému usnadňující mluvenou komunikaci skupině pacientů, která je postižena trvalou ztrátou hlasivek a používají elektrolarynx.

Pro dosažení cíle autor zvolil původní cestu, kdy s využitím systémů automatického rozpoznávání řeči a ve spolupráci s mluvčí po totální laryngektomii komunikující pomocí elektrolarynxu, navrhl celou řadu původních přístupů a realizoval velké množství vlastních experimentů. Po vytvoření unikátní databáze elektrolaryngových promluv provedl autor nejprve akusticko-fonetickou analýzu signálu s ohledem na automatické rozpoznávání, na jejímž základě se zabýval redukcí fonémů a zaměřil se na úpravu automatického systému rozpoznávání řeči. Provedl řadu experimentů a porovnání za různých podmínek, s různými akustickými modely a porovnal přesnost rozpoznávání mezi strojem a člověkem. Pro zvýšení přesnosti rozpoznávání dále upravoval data a v závěru práce realizoval trenažer, který pomáhá řečníkovi s výukou prodlužování vybraných fonémů a na základě reálných dat adaptuje akustické modely. Vytýčeného cíle disertační práce bylo dosaženo.

Za účelem zvýšení přesnosti navrhl autor protažení vybraných fonémů. Zatímco praktické protažení frikativ řečníkem si lze snadno představit, tak protažení neznělých explozív je pro čtenáře neseznámeného s elektrolarynxovými promluvami obtížné. Proto bych se rád zeptal:

- Vzniká protažení neznělých explozív tak, že řečník udělá delší pauzu a pak následuje akcent na (iniciální) explozi?
- Jak ovlivňuje protažení neznělých explozív řečníkem jednotlivé segmenty – okluzi, závěr (explozi) a tranzient (přechod do následující samohlásky)?
- Na obrázcích 4.4 a 4.5 vpravo je vidět přítomnost periodického (přístrojového) šumu během elektrolarynxových okluzí. Je možné a případně účelné, potlačit tento šum, např. adaptivní filtrací?
- Na obr. 5.4 je zobrazeno protažení /s/ na dvojnásobek. Jak je to však při protažení neznělých explozív fázovým vokodérem? Kde je začátek a konec "protahování"? Jaká je délka okna, případně překrytí u realizovaných experimentů?

Výsledky a přínos práce

Výsledky práce významně přispívají k realizaci systému, který může zlepšit život lidí postižených trvalou ztrátou hlasivek. Vytvoření unikátního korpusu, trenažéru a upraveného robustního automatického systému rozpoznávání elektrolarynxové řeči s přesností rozpoznávání okolo 90 % jsou nezpochybnitelnými původními výsledky a přínosem této práce.

Formální stránka

Disertační práce je psaná obvyklým vědeckým stylem. Text práce logicky sleduje postup výzkumné činnosti v jednotlivých etapách. Některé drobné formální nedostatky nijak nesnižují celkový přínos práce, např.:

- U obrázků 4.4 a 4.5, ilustrujících analýzu okluzí u velárních explozí, by bylo vhodné srovnatelné měřítko. Takto se měřítko amplitud liší až 40x a průběhy jsou v délkách od 20 do 140 ms, čímž jsou letmá porovnání zavádějící.
- Na str. 126 je foném /v/ nesprávně zahrnut pod označení "neznělé fonémy".
- Do seznamu literatury pronikly nedbale uvedené citace, např. z [28] není vůbec zřejmé, kde byla práce publikována a ve [29] je mix vlastních jmen a jejich iniciál.
- Pro rychlou orientaci je v práci uveden seznam zkratek, ze kterého však vypadla v práci jedna z nejčastějších zkratek: EL (*elektrolarynx*, *elektrolarynxový*, apod.).
- Seznam tabulek by měl být v obsahu uveden na str. 140 (nikoliv 139).
- Práce má poměrně málo překlepů (např.: str. 13 - 4.ř. shora "*u nejen pacientů*", str. 63 - 3.ř. shora "*nahrávanám*", str.129 - 3.ř. shora "*stávacíji*", str.126 - 9.ř. zdola "*demostruji*").

Publikační činnost

Výsledky své práce doktorand publikoval jako hlavní autor především na oborově prestižních mezinárodních konferencích INTERSPEECH a TSD. Doktorand také publikoval na dalších mezinárodních a studentských vědeckých konferencích. Řada publikací mezi aplikačními výsledky svědčí o zapojení doktoranda i do dalších výzkumných projektů. U neveřejných impaktovaných publikací, které doktorand v seznamu svých prací uvádí, je obtížné posoudit autorský podíl a vztah k tématu disertace.

Závěr

Stanovené cíle disertace byly splněny. Předložená disertační práce přináší původní výsledky a přispívá k rozvoji vědy. Doktorand projevil schopnost samostatně vědecky pracovat, publikoval výsledky své vědecké práce na mezinárodní úrovni a podílel se na řešení výzkumných projektů.

Disertační práce, podle mého názoru, odpovídá obecně uznávaným požadavkům k udělení akademického titulu Ph.D. a **doporučuji ji k obhajobě**.

V Praze, 14. dubna 2020

Prof. Ing. Roman Čmejla, CSc.

FILOZOFICKÁ FAKULTA
Univerzita Karlova

Fonetický ústav Filozofické fakulty Univerzity Karlovy

Univerzita Karlova
Filozofická fakulta
Fonetický ústav
doc. Mgr. Radek Skarnitzl, Ph.D.

Posudek disertační práce

Rozpoznání řeči pacientů po totální laryngektomii komunikujících pomocí elektrolarynxu,

kterou předkládá

Ing. Petr Stanislav

Disertační práce Ing. Petra Stanislava se věnuje zajímavému a zároveň bezpochyby prospěšnému tématu. Její výsledky mají potenciál zlepšit kvalitu života pacientům, kteří z nejrůznějších důvodů přišli o hrtan a tím o schopnost produkovat fonovanou řeč.

Po krátké první kapitole, v níž jsou představeny cíle práce, se druhá kapitola zabývá možnými příčinami ztráty hlasu a rehabilitací hlasu. Tématika je zde přehledně a čtivě představena, autor jednotlivá řešení tvorby hlasu bez přítomnosti hrtanu doprovází efektivně a jednotným způsobem zpracovanými obrázky. Všechny možnosti, s jejich výhodami a nevýhodami, jsou přehledně shrnuty v závěrečné tabulce. U popisu některých metod však postrádám více či méně podstatné detaily. Např. na str. 17 u jícnového hlasu se hovoří o krátkém trvání produkované řeči a zároveň o nácviku slov a vět. Jaké a jak dlouhé jednotky je mluvčí užívající tuto metodu schopen produkovat? Na str. 19 autor zmiňuje obtížnou srozumitelnost elektrolaryngální (EL) řeči a uvádí pro to několik důvodů. Zarazilo mě, že zde jako jeden z hlavních důvodů není uvedeno setření rozdílu mezi neznělými a znělými obstruenty – a to zejména v souvislosti s tím, že je tomuto rozdílu věnována většina experimentální části předložené disertační práce. Na stranách 19 a 20 je mechanismus tvorby hlasu s protézou popsán tak, že vzduch z trachey projde skrz voperovanou fistuli do jícnu, kde naráží do jeho stěn, čímž dojde k jeho rozkmitání. Na straně 22 se pak uvádí, že se výsledný hlas vyznačuje vysokou kvalitou a dlouhou fonační dobou. Je zcela zřejmé, že takto tvorba hlasu fungovat nemůže. V popisu chybí klíčová

charakteristika voperované fistule, která by takové vlastnosti hlasu umožnila; v práci je však popsána v podstatě jako pouhý průchod mezi tracheou a jícnem. Konečně v oddílu 2.2.3 je zmíněn řečový sifon a neolarynx, z textu však není patrný samotný mechanismus produkce hlasu. Rád bych proto autora poprosil o objasnění.

Třetí kapitola předložené disertační práce se věnuje automatickému rozpoznávání řeči (ASR). Jako čtenář se spíše všeobecnou orientací v této problematice oceňuji čtenářsky příjemný styl popisu, k technickým aspektům se však kompetentně vyjadřovat nemůžu.

Čtvrtá a pátá kapitola představují empirickou část disertační práce. Ing. Stanislav nejprve ve čtvrté kapitole popisuje sestavení korpusu EL řeči získané z nahrávek jedné mluvčí, následně na několika ukázkách ilustruje hlavní vlastnosti EL řeči a porovnává je s řečí standardně fonovanou. Protože slovní přesnost běžného ASR systému byla pouhých 18,5 %, přistoupil autor nejprve k ladění parametrů systému a následně k redukci fonetické sady, konkrétně k nahrazení neznělých hlásek jejich znělými protějšky. Zde však není zřejmé, proč z těchto záměn byla vynechána dvojice /p – b/ (viz tabulku 4.4 na str. 78) nebo podle jakého klíče byly znělostní dvojice v dílčích testech seskupovány (tabulka 4.5 na str. 80). Pátá kapitola pokračuje ve snahách ještě zlepšit výsledek rozpoznávacího procesu. Autor se nejprve zaměřuje na doplnění řečového korpusu o nové nahrávky, především ty obsahující slova s minimálními páry založenými na fonologické znělosti, poté se věnuje vyrušení vlivu odlišného způsobu nahrávání databází EL řeči pomocí optimalizace parametrů kepstrální průměrové normalizace. Samostatný oddíl je věnován porovnání výsledků poslechového testu respondenty s výsledky ASR. Výsledky poukázaly na problémy s rozpoznáním právě v kontextech založených na znělostním protikladu. Proto autor v navazujícím kroku navrhuje „augmentaci dat", která spočívá v prodloužení (nazývaném poněkud zvláštně „protažení") neznělých hlásek, a to na úrovni příznaků a na úrovni akustického signálu pomocí metody TD-PSOLA. Výrazné zlepšení přesnosti systému ASR vedlo k pokusu s mluvčí, která byla požádána o prodlužování neznělých hlásek ve vlastní řeči. V závěrečném oddílu pak autor rozvinul myšlenku trenažéru, který by mluvčím pomáhal s nácvikem prodlužování cílových hlásek.

Předložená disertační práce je na velmi vysoké odborné úrovni z hlediska automatického zpracování řečového signálu. Nemohu se však nevyjádřit k nepřesnostem v popisu některých fonetických aspektů. Ty nacházíme především ve třetí kapitole, kde je definice fonému v poznámce pod čarou na str. 29 zcela chybná; hlasivkový trakt je termín velmi nevhodný, protože má s hlasivkami pramálo co do činění (nadhrtanové rezonanční dutiny se označují jako vokální trakt); frekvence se měří v hertzech, ne herzech; vztah definovaný na str. 35 jako závislost intenzity na frekvenci neodpovídá hlasitosti, ale hladinám hlasitosti (vnímaná hlasitost je velmi složitý koncept). Další problematické pasáže se nacházejí v páté kapitole. V poznámce pod čarou č. 7 na str. 95 je zmíněn rozdíl mezi *i* a *y* tím, že „z akustického hlediska jsou oba fonémy identické"; to je tvrzení zcela nesmyslné, a to na několika úrovních. Obecně platí, že fonémy žádné akustické hledisko či akustické vlastnosti nemají. Především ale *i* a *y* v češtině nepředstavují fonémy, ale písmena (grafémy) – která se ve výslovnosti realizují totožně, jako

[ɪ]. Stojí za zmínku, že v pozn. č. 10 na str. 98 se již o písmenech hovoří. Na str. 107 je mezi neznělými fonémy uvedeno i znělé [v].

Rád bych, aby se Ing. Stanislav při obhajobě kromě výše uvedených dotazů vyjádřil i k následujícím otázkám:

- Měli respondenti v oddílu 5.2 předchozí zkušenost s poslechem EL řeči?
- Na str. 116 autor srovnává umělé a autentické prodloužení [s] ve slově *kose* a jako hlavní rozdíl uvádí „slabší zastoupení frekvencí kolem 2 kHz". Je to jev, který se opakoval u více položek? O čem by to svědčilo? Nemůže být významnější – a to jak percepčně, tak i pro samotný experiment s prodlužováním a následně pro funkčnost zamýšleného trenažeru – slábnoucí šum během produkce [s] v autentické realizaci mluvčí?
- Můj hlavní dotaz se týká možnosti generalizace výsledků na základě experimentů s EL řečí jedné mluvčí. Jakou změnu v přesnosti systému ASR by autor očekával při použití stávajícího, nejlepšího nalezeného systému pro jiného mluvčího s elektrolarynxem, např. pro mluvčího mužského pohlaví?

Na závěr shrnuji, že Ing. Petr Stanislav ve své disertační práci důkladně představil problematiku využití ASR pro velmi specifickou populaci mluvčích, tedy mluvčích po totální laryngektomii. Disertační práce přináší zcela nové poznatky v oblasti rozpoznávání řeči, autor logickým a strukturovaným způsobem identifikoval možnosti zlepšení přesnosti systému ASR, tyto možnosti adekvátně otestoval a výsledky přinášejí oproti baseline systému významné zlepšení. Disertační práce je napsána čtivým a zároveň vysoce odborným stylem, i přes občasný výskyt jazykových chyb, neobratností či překlepů (poněkud neobvyklé až matoucí je např. kladení mezer za desetinnou čárku).

Na základě výše uvedených skutečností uzavírám, že autor jednoznačně prokázal schopnost samostatné vědecké práce. O tom svědčí i poměrně bohatá publikační činnost, a to v oblasti základního i aplikovaného výzkumu; autorovy publikace většinově přesahují disertační téma.

Navrhuji klasifikaci „prospěl" a disertační práci doporučuji k obhajobě.

V Praze dne 11. května 2020

doc. Mgr. Radek Skarnitzl, Ph.D.
Fonetický ústav
Filozofická fakulta Univerzity Karlovy