FAKULTA
APLIKOVANÝCH VĚD
ZÁPADOČESKÉ
UNIVERZITY
V PLZNI

AUTOREFERÁT
disertační práce

PLZEŇ, 2020

Ing. Petr Stanislav

Ing. Petr Stanislav

Rozpoznávání řeči pacientů po totální laryngektomii komunikujících pomocí elektrolarynxu

obor

Kybernetika

Autoreferát disertační práce k získání
akademického titulu "Doktor"

V Plzni, 31. ledna 2020

Disertační práce byla vypracována v prezenčním/kombinovaném doktorském studiu na katedře mechaniky Fakulty aplikovaných věd Západočeské univerzity v Plzni.

Uchazeč: Ing. Petr Stanislav
Fakulta aplikovaných věd
Katedra kybernetiky
Technická 8, 306 14 Plzeň

Školitel: prof. Ing. Josef Psutka, CSc.
Fakulta aplikovaných věd
Katedra kybernetiky
Technická 8, 306 14 Plzeň

S disertační prací je možno se seznámit na studijním oddělení Fakulty aplikovaných věd Západočeské univerzity v Plzni, Technická 8, UC 133.

prof. Ing. Josef Psutka, CSc.
předseda oborové rady Kybernetika

Anotace

Disertační práce se zabývá problematikou rozpoznávání řeči pacientů, kteří podstoupili totální laryngektomii, a k produkci hlasu využívají elektrolarynx. V první části práce jsou přiblíženy důvody ztráty hlasu a metody, které jsou v současnosti využívány pro jeho rehabilitaci spolu s jejich principy. Významnou pomoc s rehabilitací hlasu mohou poskytnout řečové technologie zpracovávající přirozenou řeč. Z tohoto důvodu jsou v práci popsány metody, které jsou využívány pro konstrukci automatických systémů rozpoznávání řeči (ASR). S ohledem na specifika řeči generované za pomoci elektrolarynxu je v práci prezentován postup pro sestavení speciálního řečového korpusu složeného z nahrávek hlasu pacienta po totální laryngektomii. Specifická řečová data slouží následně pro otestování robustnosti obecného systému rozpoznávání řeči. Získané výsledky však indikují potřebu navrhnout speciální ASR systém s individuálními požadavky vzhledem ke specifikům rozpoznávané řeči. Následně je navrženo několik postupů úpravy akustických dat za účelem zvýšení přesnosti rozpoznávání. Jako nejúčinnější se ukázalo protažení neznělých fonémů, proto byl vyvíjený ASR systém rozšířen o modul zohledňující právě toto protažení. V práci je popsáno nemalé množství experimetů, které byly provedeny za účelem ověření dílčích hypotéz.

Obsah

Úvod

Lidská řeč je jedním z hlavních dorozumívacích prostředků užívaných člověkem, proto ztráta hlasu způsobuje řadu komplikací. Její příčinou může být chirurgický zákrok, který byl proveden za účelem odstranění nádorového onemocnění v oblasti hrtanu, nebo poškození hrtanu vlivem traumatické nehody. Proto se lékaři již od konce 19. století snaží o rehabilitaci pacientova hlasu za účelem zvýšit kvalitu jeho života.

První snahy o navrácení schopnosti mluvit nebyly příliš úspěšné a byly svým způsobem i životu nebezpečné. Přesto neutuchající snaha lékařů postupně vyústila nejen ve vyvinutí bezpečných operačních postupů, ale i metod následně využívaných pro rehabilitaci hlasu. Mezi nejpoužívanější přístupy patří chirurgicko-protetické a foniatrické metody. Nejčastěji postižení pacienti využívají pro rehabilitaci hlasu elektrolarynx, jícnový hlas a tracheoezofageální píštěl. Bohužel žádná z používaných metod není univerzálním řešením pro každého pacienta. U nemalého počtu pacientů je navíc snaha začít opětovně komunikovat s okolím pomocí mluvené řeči doprovázena významnou psychickou zátěží mluvčího, který se například může ostýchat mluvit na veřejnosti. Z toho důvodu je problematice rehabilitace hlasu v současnosti věnována nemalá pozornost. Významnou pomoc mohou v tomto ohledu přinést řečové technologie.

V polovině 20. století se s rozvojem číslicových počítačů začaly objevovat první snahy o zpracování přirozené řeči počítačem. Toto úsilí vyústilo ve vyvinutí v dnešní době hojně užívaných systémů automatického rozpoznávání řeči (zkr. ASR) a systémů pro syntézu řeči (zkr. TTS). Nejmodernější ASR systémy jsou schopné pracovat s obrovskými slovníky v mnoha rozličných situacích. Největší problémy však stále způsobuje okolní hluk ovlivňující výkon těchto systémů. O eliminaci jeho vlivu se výzkumníci snaží už od samých počátků jejich vývoje. V mnoha případech se inspirují schopnostmi člověka, protože ten je schopen relativně úspěšně porozumět promluvě i za velmi ztížených podmínek.

Tyto snahy velmi často vedou k vytvoření multimodálních systémů zpracovávajících nejen akustická data, ale například i data obrazová. Bohužel multimodální systémy zatím nedosahují požadovaných kvalit, proto se vývoj ASR systémů v současnosti ubírá zejména směrem vývoje komplexnějších modelů. Běžně využívané systémy rozpoznávání řeči jsou však trénovány na obecných datech a pro uživatele postižené trvalou ztrátou hlasu jsou nepoužitelné. Jako jeden z hlavních problémů se jeví jiné charakteristiky produkované řeči a ztráta určitého množství informace v ní obsažené. Ke ztrátě části informace dochází v důsledku chybějícího buzení proudu vzduchu hlasivkami. Nejčastěji využívané metody

rehabilitace hlasu se totiž snaží nahradit chybějící buzení jiným zdrojem buzení, které má ale v podstatě konstantní charakter. Obecné ASR systémy pak nejsou bez adaptace schopné obstojně tuto řeč zpracovávat, proto se většina doposud vyvíjených metod snaží získat tuto ztracenou informaci z dalšího doprovodného zdroje dat (např. kamerového záznamu artikulace). Výsledné multimodální systémy však zatím nedosahují konkurence schopných výsledků a ve většině případů předpokládají využití dalšího (prozatím) neergonomického zařízení.

Tato práce si klade za cíl prozkoumání možností rozšíření schopností ASR systému tak, aby se výkon vytvořeného systému co možná nejvíce blížil obecnému na řečníkovi nezávislému ASR systému. Velký důraz je kladen na co možná nejmenší požadavky na samotného řečníka, aby bylo možné navržený systém převést do praxe, a tím tak zlepšit v určitých aspektech život lidí postižených trvalou ztrátou hlasivek.

1 Konstrukce ASR systému pro uživatele po totální laryngektomii hovořící pomocí elektrolarynxu

1.1 Aplikace obecného systému rozpoznávání a dosažené výsledky

Ze získaných výsledků je možné usoudit, že redukce fonetické sady může vést ke zlepšení přesnosti. Nicméně předpoklad, že všechny neznělé fonémy jsou shodné se svými znělými ekvivalenty se ukázala jako mylná. Zároveň také není možné říci, že náhrada např. dvojice /*s*/ a /*š*/ za každých okolností povede k lepším výsledkům. Při hlubší analýze se ukázalo, že velmi záleží na kontextu daného fónemu, ten totiž velmi ovlivňuje jeho podobu. Řeč představuje spojitou formu signálu a při vyslovování různých slov obsahujících stejný foném s odlišným okolím může dojít k odchylkám například v artikulaci, příkladem může být dvojice slov *hrad* a *had*. Toto pozorování ověřil i dodatečný experiment, ve kterém nebyla provedena náhrada /*s*/ za /*z*/ pouze pro případ trifónu *b-s+t*, který je například ve slově *obstát*. Díky vynechání tohoto jediného trifónu byla výsledná nejlepší přesnost u trifónového akustického modelu $83,39$ % (původně $83,28$ %) v případě zerogramového jazykového modelu a $88,44$ % (původně $88,31$ %) v případě trigramového modelu.

2 Návrh a realizace úprav ASR

2.1 Doplnění řečového korpusu o specifická data - vliv nových dat na kvalitu akustického modelu

2.1.1 Vliv nových dat na kvalitu modelů

2.1.2 Eliminace vlivu kanálu

K tomu, aby bylo možné použít všechna dostupná data, je tedy potřeba eliminovat vliv kanálu. Standardně se k tomuto účelu používá Cepstral Mean Normalisation (CMN). Principem této metody je odstranění vlivu kanálu na základě střední hodnoty kepstrálních koeficientů.

Výsledky modelů po eliminaci vlivu kanálu

Aplikací CMN dosáhl HMM-GMM model srovnatelných výsledků s výsledky dosaženými v části **??**. Dalším krokem bylo natrénování HMM-DNN modelu. Trénovaná neuronová FF síť měla 5 skrytých vrstev, výstupní vrstva byla typu softmax s dimenzí rovnou počtu HMM stavů. Postupně byla natrénována síť s 1024, 2048 a 4096 neurony v každé skryté vrstvě. Vstupní data byla parametrizována pomocí PLP s 12 kepstrálními, delta a delta-delta koeficienty a CMN počítané ze všech nahrávek dané etapy. Byl využit fonémový zerogramový model, který minimalizoval vliv jazykového modelu na přesnost rozpoznávání. Jazykový model je fonémový zerogramový tak, aby byl co nejvíce amplifikován vliv akustického modelu. V tab. 2.1 jsou zapsány dosažené výsledky všech natrénovaných variant. Nejvyšší přesnosti dosáhl model s 4096 neurony v každé vrstvě, ale rozdíl od ostatních variant s menším počtem neuronů v každé vrstvě nebyl významný. Nejlepší HMM-DNN model dosáhl $Acc_p = 84,66$ % To je zlepšení o $6,97$ % absolutně oproti HMM-GMM na kompletní testovací sadě.

Počet neuronů	Acc_p [%]
1024	84,26
2048	84,51
4096	84,66

Tabulka 2.1: Dosažená přesnost neuronové sítě s monofónovým zerogramovým jazykovým modelem.

2.2 Poslechový test a porovnání výsledků člověka a stroje

2.2.1 Izolovaná slova

Rozpoznání slova, které bylo vysloveno v klidném prostředí se jeví jako velice jednoduchý úkol. Pokud jej ale vyslovil řečník používající EL, tak už to tak snadné být nemusí. Účastníci poslechového testu na izolovaných slovech mají za úkol postupně vyslechnout 320 nahrávek izolovaných slov a vybrat jednu z předem definovaných odpovědí:

a) slovo A *(např. kosa)*,

b) slovo B *(např. koza)*,

c) nemohu rozhodnout.

Ve výčtu možností je vždy skutečně pronesené slovo a k němu pak varianta lišící se pouze znělostí jednoho fonému. První dvě možnosti jsou vždy v abecedním pořadí. Nahrávky použité v rámci poslechového testu pocházejí z 2. etapy nahrávání. Poslechového testu se účastnilo 19 subjektů z řad kolegů.

Celková přesnost rozpoznávání byla vypočtena podle vzorce

$$Acc_w^{human} = \frac{1}{N} \sum_{i=1}^{N} f_i * 100, \tag{2.1}$$

kde $N = 320$ a f_i se rovná relativní četnosti správných odpovědí na otázku i v poslechovém testu s izolovanými slovy. Dosáhla hodnoty $Acc_w^{human} = 70,47$ %.

2.2.2 Slovní bigramy

V druhém poslechovém testu mají posluchači za úkol vyslechnout 333 nahrávek slovních bigramů[1] a vybrat jednu z předem definovaných odpovědí. Ty mají vždy tento formát

[1]Nahrávky obsahují dvě po sobě vyslovená slova.

	Relativní četnost odpovědí [%]		
Slovo	***a)***	***b)***	***Nevím***
borce *(a)*	**57,90**	36,84	5,26
porce *(b)*	21,05	**52,63**	26,32
kosa *(a)*	**0,00**	100,00	0,00
koza *(b)*	0,00	**100,00**	0,00
přibít *(a)*	**94,74**	5,26	0,00
připít *(b)*	10,52	**89,48**	0,00

Tabulka 2.2: Ukázka výsledku poslechového testu na izolovaných slovech.

a) slovo A + slovo A *(např. kosa + kosa)*,

b) slovo A + slovo B *(např. kosa + koza)*,

c) slovo B + slovo A *(např. koza + kosa)*,

d) slovo B + slovo B *(např. koza + koza)*.

Je zřejmé, že to představuje všechny kombinace, které lze z dvojice slov vytvořit. Což není velký problém, každá nahrávka izolovaného slova obsahuje minimálně $0,5$ s ticha na svém začátku a konci. Pokud jsou tyto nahrávky spojeny[2], vznikne jediná nahrávka obsahující dvě zájmová slova oddělena krátkou pauzou. Z každé dvojice slov vznikly vždy dvě nahrávky lišící se pořadím slov.

Vyšší počet položek v testu je zapříčiněn faktem, že pro určitá slova existuje více než jedna kombinace s jiným slovem[3]. Ve snaze zkrátit, už tak docela náročný poslechový test, byly vygenerovány bigramy odpovídající pouze možnostem *b)* a *c)*. Účastníci poslechového testu o tom však nebyli informováni. Přesto tento poslechový test dokončilo pouze 12 účastníků.

Stejně jako v předchozím případě jsou správné odpovědi zvýrazněny tučně. Účastníci v žádném z provedených poslechových testů nebyli omezeni počtem opětovného přehrání promluvy.

[2]Ke spojení je možné použít nástroj *ffmpeg* nebo *sox*.

[3]Ve valné většině se jedná o slova obsahující písmena *i/y*, která jsou v akustické formě identická. Příkladem může být dvojice *nebyli + nepili* a *nebili + nepili*.

Poslední prezentovaný příklad zastupuje množinu odpovědí, kdy účastníci naprosto správně určili význam slov. Průměrná dosažená přesnost rozpoznávání člověka, počítána pomocí rovnice (2.1), dosáhla hodnoty $Acc_w^{human} = 66,24\ \%$.

	Relativní četnost odpovědí [%]			
Slovní bigram	***a)*** *A + A*	***b)*** *A + B*	***c)*** *B + A*	***d)*** *B + B*
borce + porce *(b)*	16,67	**50,00**	0,00	33,33
porce + borce *(c)*	8,33	0,00	**66,67**	25,00
kosa + koza *(b)*	0,00	**8,33**	0,00	91,67
koza + kosa *(c)*	0,00	0,00	**0,00**	100,00
přibít + připít *(b)*	0,00	**100,00**	0,00	0,00
připít + přibít *(c)*	0,00	0,00	**100,00**	0,00

Tabulka 2.3: Ukázka výsledku poslechového testu na dvojicích slov.

2.2.3 Výsledky porovnání

Byl využit model HMM-DNN, konkrétně se jednalo o DNN síť s 6 vrstvami (5 skrytých vrstev, každá s 4096 neurony), přičemž výstupní vrstva byla typu softmax s dimenzí rovnou počtu HMM stavů.

Provedené poslechové testy poskytly dva typy výsledků. První reprezentuje schopnost určit význam izolovaného slova, druhý schopnost rozeznat dvě velmi podobná slova. Pro potřeby porovnání přesnosti rozpoznávání člověka s výsledky dosaženými ASR systémem, byly navrženy celkem 3 experimenty využívající výše popsaný akustický model.

Experiment s izolovanými slovy. Jeho základem je zerogramový LM obsahující více než 1 milion slov. Všechny položky stejnou pravděpodobnost, tím je zaručeno, že nebudou preferována četnější slova. Slovník potřebný pro tento LM je sestaven z textů pocházejících z novinových článků, webových zpravodajských serverů, filmových titulků a přepisů televizních pořadů. Tento test je pojmenován jako „one-mil“.

Za účelem kompenzace a priorní znalosti člověka, je vytvořen druhý experiment, který pracuje s redukovaným LM. Obsahuje pouze slova, která se opravdu vyskytla v rámci poslechového testu ($N = 320$). Tento experiment je nazván jako „reduced“.

Poslední navržený experiment spočívá v tom, že je ke každé nahrávce se slovním bigramem vygenerován speciální (zerogramový) LM. Ten obsahuje vždy pouze všechny 4 kombinace slov. Tento experiment je nazván jako „bigrams“.

Pro upřesnění je nutné zmínit, že *i/y* na výsledném ohodnocení nehraje roli. V případě experimentu „one-mil“ je úspěšnost rozpoznávání ASR systému významně nižší než úspěšnost rozpoznání člověka. To je způsobeno především enormní perplexitou jazykového modelu. Ta je rovna velikosti slovníku. Po zmenšení slovníku se podařilo získat výsledky srovnatelné s člověkem. Je ale potřeba zdůraznit, že i v případě „reduced“ experimentu vykazuje člověk vyšší úspěšnost rozpoznání, protože čelí pouze perplexitě 3 (kdykoli může podívat na nabízené možnosti). Tento faktor by bylo možno eliminovat tak, že by účastník testu přepsal obsah promluvy a tento přepis by byl následně porovnán se skutečným obsahem. To by ale významně zvýšilo časovou náročnost poslechového testu a bylo by velmi komplivané získat kompletní výsledky od relevantního množství účastníků. Už jen podíl odpadlíků mezi prvním a druhým poslechovým testem dosáhl závratných 30 %.

	Acc [%]		
	one-mil	reduced	bigrams
člověk	70,47	70,47	66,24
stroj	61,24	69,91	54,82

Tabulka 2.4: Porovnání dosažených výsledků člověka a stroje.

2.3 Augmentace dat

Naprosto markantní význam má kontext.

Ať se řečník snaží sebevíc, tak i při využití metod rehabilitace hlasu uvedených v části **??**, se v důsledku ztráty hlasivek část informace z produkované řeči ztrácí. V poslední době bylo prezentováno několik přístupů jak ztracenou informaci obnovit. Souhrn těch nejperspektivnějších je v [1]. Ve valné většině případů se využívá obohacení akustického modelu o artikulační data, nebo dokonce využití jen těchto artikulačních dat [2] . Problém ale spočívá v tom, že ne všechny akustické nuance mezi podobnými fonémy jsou artikulací ovlivněny. Pořízení záznamu artikulačních dat často vyžaduje používaní dalšího zařízení (kamery, ultrazvuku, atp. [3, 4, 5, 6], nebo dokonce nutnost podstoupení dalšího operačního zákroku (magnety [7]). Samozřejmě je nutno říct, že většina těchto vyvíjených systémů si klade za cíl kompletně nahradit současné metody rehabilitace hlasu. Na druhou stranu faktem je,

že ani po dlouholetém vývoji se většina těchto systémů nedostala z rané vývojové fáze. Nepochybně hraje roli i skutečnost, že tato problematika není v ohnisku zájmu řečařské komunity.

Pokud tedy není úplně reálné získat ztracenou informaci pomocí kompletní změny paradigmatu fungování systémů rozpoznávání řeči, tak zbývá jen pracovat s informací, která je k dispozici, a adaptovat současný model. Určitou možností je nahrazení ztracené informace konkrétní cílenou změnou produkované řeči, která je zohledněna modelem. Samozřejmě takovýto přístup nezbaví řečníka elektrolarynxu, ale může mu pomoci v situacích, které jsou pro něj stresující a v konečném důsledku mu velmi komplikují život.

Bohužel získání reálných dat je časově náročný proces (viz **??** a 2.1). Navíc není zřejmé, jestli se vůbec vyplatí taková data pořizovat, protože se jedná o hypotézu. Mnohem snadněji se jeví pro účely testování možnost uměle data protáhnout v místech výskytu zájmových fonémů. Toto protažení je teoreticky možné realizovat dvěma způsoby:

1. protažením na příznacích;
2. protažením na zvuku.

V obou případech je nezbytné získat co možná nejpřesnější fonetické zarovnání. Pokud bude obsahovat chyby, tak mohou být protahovány úplně jiné úseky řeči. K natrénování HMM-DNN modelu na protažených datech je zapotřebí zarovnání získané HMM-GMM nebo HMM-DNN modelem, viz **??**. U obou variant protažení je postup stejný:

1. natrénování akustického modelu na originálních datech;
2. získání zarovnání;
3. protažení zájmových fonémů podle zarovnání;
4. natrénování nového akustického modelu na protažených datech.

Nový model může být otestován a získané výsledky mohou být porovnány s těmi dosavadními. Tyto experimenty navíc pomohou určit vhodné hodnoty parametrů pro případné skutečné protažení dat.

2.3.1 Protažení na příznacích

Protažení na příznacích je založeno na skutečnosti, že při protažení (např. fonému) a následné parametrizaci, budou v inkriminovaných mikrosegmentech po sobě následovat velmi podobné příznakové vektory. V prvním kroku je nahrávka standardně parametrizována. Tyto vektory jsou následně zduplikovány a představují dvojnásobné protažení. Mechanismus by zároveň vyřešil omezení, kdy je kopírováním možné získat pouze protažení odpovídající celočíselnému násobku původní délky.

Dosažené výsledky

Pro prvotní ověřovací experiment bylo zvoleno 2x protažení fonému /s/, tedy všechny vektory odpovídající /s/ byly zduplikovány. Otestování bylo jako v předchozích případech realizováno na testovací sadě s fonémovým zerogramovým jazykovým modelem, aby byl minimalizován vliv LM. Tento nový model dosáhl přesnosti rozpoznávání $85,11$ %.

Další experiment byl realizován na protažených fonémech /k/, /p/, /s/, /t/ a /v/, které reprezentují většinu neznělých zájmových fonémů. Zarovnání bylo identické jako u předchozího experimentu, stejně tak bylo uvažováno jejich 2x protažení, tzn. že všechny vektory inkriminovaných fonémů byly zduplikovány. Znovu byl natrénován HMM-DNN model se stejnými parametry a následně byl otestován s fonémovým zerogramovým jazykovým modelem. Přesnost rozpoznávání na testovací sadě dosáhla hodnoty $87,50$ %, což lze považovat za významné zlepšení.

V další fázi bylo potřeba ověřit, zda jiné hodnoty násobku protažení nemohou poskytnout lepší výsledek. Experiment byl zopakován pro 3x, 4x a 5x jejich původní délky. Protaženy byly fonémy /k/, /p/, /s/, /t/ a /v/. Proces natrénování a otestování modelu korespondoval s předchozími experimenty. Dosažené výsledky byly zaznamenány do tab. 2.5. Pro úplnost byla tabulka doplněna o baseline model neobsahuhjící protažení a model s 2x protažením. Z uvedených výsledků je patrný jasný trend, větší než 2x protažení vede ke zhoršení přesnosti rozpoznávání. Optimální hodnota protažení tak teoreticky leží někde mezi jednonásobkem a trojnásobkem původní délky. Bohužel uvedeným postupem nelze přesně určit hodnotu míry protažení.

	Míra protažení				
	1x	2x	3x	4x	5x
Acc_p [%]	84,66	**87,50**	86,73	85,12	83,65

Tabulka 2.5: Vliv míry protažení na přesnost modelu.

2.3.2 Protažení na zvuku

Protažení na příznacích vedlo sice k významnému zlepšení přesnosti rozpoznávání, ale tento přístup není bohužel reálně použitelný. Vhodnou alternativou může být model pracující s fonémy protaženými přímo v audio signálu. Taková data budou teoreticky více odpovídat reálným datům získaným od řečníka. Stejně jako v předchozím případě je k protažení potřeba zarovnání. To s určitou mírou přesnosti určuje počáteční a koncové hranice

jednotlivých fonémů. S ohledem na stanovené hranice je možné určitý úsek protáhnout například pomocí

- převzorkování signálu,
- TD-PSOLA algoritmu,
- fázového vokodéru.

Asi nejjednodušší metodou je převzorkování dat, pro jehož realizaci stačí načíst všechny vzorky odpovídající vybranému fonému a změnit jejich vzorkovací frekvenci. Pokud je cílem tento úsek protáhnout, je nová vzorkovací frekvence menší než originální. Hlavním problémem této metody je tonální posun[4].

Zbylé dva uvažované přístupy využívají sofistikovanější úpravy signálu v časové oblasti. Obě metody využívají *analýzu* signálu, pro jeho následné *zpracování*, které je zakončené *syntézou*.

Pomocí těchto dvou zmíněných metod je možné upravit nejen délku, ale i fundamentální frekvenci F_0 signálu.

Dosažené výsledky s DNN

K ověření schopností modelu pracovat s uměle protaženými daty byl použit stejný HMM-DNN model jako v předchozích případech. V datech jsou protaženy všechny výskyty fonémů /k/, /p/, /s/, /t/ a /v/. Uvažováno je protažení 1,25x, 1,50x, 1,75x a 2,00x. Jazykový model je stejně jako v případě protažení na příznacích fonémový zerogramový. Dosažené výsledky jsou shrnuty v tab. 2.6. Nejlepšího výsledku dosáhl *baseline* model s hodnotou 84,66 %. S libovolným protažením dochází k poklesu přesnosti.

	Míra protažení				
	1,00x	1,25x	1,50x	1,75x	2,00x
Acc_p [%]	**84,66**	84,48	84,15	83,12	82,55

Tabulka 2.6: Vliv míry protažení fonému na přesnost DNN modelu.

[4]Mění se fundamentální frekvence F_0. Pokud dojde ke zrychlení, frekvence se zvýší. Při zpomalení naopak sníží.

Upravené zarovnání a time delay neural network

Při analýze výsledků se ukázalo, že zarovnání v mnoha případech není zrovna nejpřesnější, a to zvláště u inkriminovaných neznělých fonémů.

S přesnějším zarovnáním je možné přistoupit k protažení fonémů /k/, /p/, /s/, /t/, /v/ a vytvoření nového modelu pracujícího s těmito daty. Jako další model je tedy použita TDNN síť. K otestování modelu je využit standardní fonémový zerogramový jazykový model. Uvažováno je protažení od 1, 25x do 3, 00x s krokem 0, 25. Výsledky experimentu jsou uvedeny v tab. 2.7. Oproti hodnotám rozpoznávání přesnosti v tab. 2.6 je vidět výrazné zlepšení přesnosti oproti baselinu modelu ($Acc_p = 85, 41$ %). Nejvyšší přesnosti 87, 90 % dosáhl model pracující s 2, 5x protaženými daty, navíc modely pracující s protažením od 1, 75x do 2, 75x dosahují velmi blízkých hodnot přesnosti rozpoznávání. To poskytuje relativně široký pracovní interval pro případné skutečné protažení dat řečníkem.

	Míra protažení								
	1, 00x	1, 25x	1, 50x	1, 75x	2, 00x	2, 25x	2, 50x	2, 75x	3, 00x
Acc_p [%]	85,41	86,42	87,05	87,58	87,71	87,69	**87,90**	87,39	87,11

Tabulka 2.7: Vliv míry protažení fonému na přesnost TDNN modelu.

Podstatnou otázkou je robustnost modelu. S ohledem na hodnoty přesnosti rozpoznávání uvedené v tab. 2.8 se dá říci, že duration model je v rámci možností robustní v poměrně širokém rozsahu protažení. Očekávaným výsledkem je nejnižší hodnota přesnosti rozpoznávání pro neprotažená data (konkrétně 78, 61 %). Většina chyb v rozpoznávání nastala v důsledku neznalosti inkriminovaných neprotažených fonémů. Tento výsledek je předpokladem pro funkci trenažéru prezentovaného v části 2.5.

	Míra protažení								
	1, 00x	1, 25x	1, 50x	1, 75x	2, 00x	2, 25x	2, 50x	2, 75x	3, 00x
Acc_p [%]	78,61	80,72	82,56	84,98	86,47	87,53	**87,90**	87,26	86,24

Tabulka 2.8: Robustnost nejlepšího TDNN modelu (2, 5x) na míru protažení.

Experimenty s uměle protaženými daty potvrdily správnost uvažované hypotézy. Protažením jednoho z párových fonémů dojde k dostatečnému odlišení velmi podobných zvukových reprezentací. Tím dojde k natrénování odlišných modelů fonémů v HMM. Model

pracující s fonémy protaženými přímo ve zvuku nakonec dosáhl lepších výsledků než model s uměle protaženými daty na příznacích.

2.3.3 Aktualizace výsledků porovnání

V části 2.2 byla prezentováno srovnání schopností člověka a stroje. Posloužily k tomu dva poslechové testy a celkem 3 ASR experimenty „one-mil“, „reduced“ a „bigrams“. S využitím nového modelu je možné aktualizovat hypotetické výsledky stroje. Hypotetické z toho důvodu, že použitá data jsou uměle protažena.

K aktualizaci výsledků byl použit model z části 2.3.2, tedy ten s $2,5$x protaženými daty.

2.3.4 Reálně protažená data

Nahrávání se zhostil stejný řečník jako v 1. a 2. etapě. Tedy žena v důchodovém věku, která používá EL v běžném životě již více než 15 let. Nahrávání se uskutečnilo v průběhu 5 měsíců od července 2018 do listopadu 2018. Texty určené k nahrávání obsahovaly většinu izolovaných slov z poslechového testu a věty, které doposud neobsahuje řečový korpus složený z 1. a 2. etapy nahrávání. Řečník byl instruován, aby slova, která obsahují zdvojená písmena (např. „kossa“), adekvátně prodloužil. Celkový čas promluv v 3. etapě dosáhl 2 hodin a 28 minut.

Hlubší analýza pořízených slov ve 3. etapě nahrávání ukázala, že proces umělého protažení produkuje svými charakteristikami velmi podobné nahrávky těm reálným. Pro vytvoření modelu pouze z reálně protažených dat se však nepodařilo získat dostatečné množství dat. Pokud jsou reálně protažená data opravdu velmi podobná uměle protaženým datům, tak by mělo být možné dosáhnout dobrých výsledků s modelem, který je natrénovaný na uměle protažených datech, ale otestovaný těmi reálně protaženými.

V případě, že se k trénovací sadě přidaly věty z 3. etapy, které neobsahují protažené fonémy, výsledná přesnost TDNN modelu s fonémovým zerogramovým jazykovým modelem dosáhla hodnoty $Acc_p = 85,88$ %. To je lepší hodnota než v případě baseline TDNN model ($Acc_p = 85,41$ %, viz tab. 2.7). Tyto experimenty podporují ideu trenažéru prezentovaného v části 2.5.

2.4 Model akcentující protažení dat

S ohledem na princip fungování HMM a jeho topologie je délka fonému modelována pomocí přechodových pravděpodobností, které vedou na funkce geometrické distribuce pravděpodobnosti [8]. Bohužel skutečná podoba těchto distribucí odpovídá spíše gamma nebo logaritmicko-normálnímu rozdělení [9]. Správné modelování délky může být realizováno úpravou přechodových funkcí v HMM nebo změnou topologie modelu. Další možností je

vytvoření speciálního modelu pracujícího s délkou jednotlivých fonémů (duration model) a reskórováním výstupních N-best hypotéz či celé rozpoznávací mřížky [9, 10, 11].

2.4.1 Princip explicitních duration modelů

Další možností je vytvoření explicitního modelu pracujícího s délkou fonémů. Aby bylo možno takový model využít, je často potřeba problém rozpoznávání přeformulovat na úlohu nalezení nejlepší sekvence slov W^* a jim odpovídajících délek D^* [9]. Za předpokladu, že je dána sekvence slov W a vektory pozorování $\boldsymbol{O}$ lze považovat za nezávislé na délkách D, je možné rovnici **??** upravit jako

$$\begin{aligned} W^*, D^* &= \underset{W,D}{\operatorname{argmax}} P\left(W, D|\boldsymbol{O}\right) \\ &= \underset{W,D}{\operatorname{argmax}} P\left(\boldsymbol{O}, D|W\right) P\left(W\right) \\ &= \underset{W,D}{\operatorname{argmax}} P\left(\boldsymbol{O}|W\right) P\left(D|W\right) P\left(W\right). \end{aligned} \tag{2.2}$$

Úkolem duration modelu je tedy odhadnout pravděpodobnosti $P\left(D|W\right)$, přičemž délku D je možné dekomponovat na m délek jednotlivých fonémů d_i

$$P\left(D|W\right) = P\left(d_1, \ldots, d_m|W\right). \tag{2.3}$$

Tuto pravděpodobnost je dále možné upravit pomocí tzv. chain pravidla do tvaru

$$\begin{aligned} P\left(d_1, \ldots, d_m|W\right) &= \prod_{i=1}^{m} P\left(d_i|d_1, \ldots, d_{i-1}, W\right) \\ &\approx \prod_{i=1}^{m} P\left(d_i|d_{i-n-1}, \ldots, d_{i-1}, W\right). \end{aligned} \tag{2.4}$$

Model tedy odhaduje pravděpodobnost $P\left(D|W\right)$ na základě délek n předchozích fonémů a předpokládaného slova W. Některé duration modely navíc pracují i s tempem řeči [12], tento efekt je však u modelu využívajícího vztah 2.4 zohledněn prostřednictvím délek n předchozích fonémů. Ve skutečnosti je vhodné vytvořit model, který bere v potaz nejen délky předchozích fonémů, ale i příznakové vektory těchto fonémů [9]. Zohlednění délek předchozích fonémů v závislosti je provedeno prostřednictvím pravděpodobnosti $P\left(d_i|x_i\right)$, kde x_i představuje příznakový vektor obsahující délky n předchozích fonémů, jejich vektory pozorování a případně další hodnoty. K odhadu této pravděpodobnosti se jako vhodné ukázaly neuronové sítě [9], [13].

Na odhad $P(d_i|x_i)$ je možné nahlížet ze dvou pohledů. V prvním případě je cílem odhadnout parametry pravděpobnostní distribuce pomocí conditional density estimation network (CDEN) [9]. V tomto případě se předpokládá, že délky fonémů odpovídají určitému pravděpodobnostnímu rozdělení, nejčastěji logaritmicko-normálnímu. Konkrétní hodnota pravděpodobnosti je pak stanovena s využitím příslušného vztahu pro výpočet hustoty pravděpodobnosti.

Druhou možností je stejně jako v případě HMM-DNN akustického modelu odhad pseudo-pravděpodobností za pomocí NN mající jako poslední vrstvu tzv. softmax vrstvu. Tento přístup nevnáší do modelu žádné předpoklady o podobě pravděpodobnostního rozdělení. Experimenty v [13] ukazují, že tento přístup je vhodnější[5].

2.4.2 Duration model se softmax vrstvou

Úlohou modelu je predikovat sekvenci délek na základě sekvence fonémů. To implikuje možnost použití levého kontextu (L) i pravého kontextu (R) fonému i. Do vstupního vektoru sítě ale mohou přijít pouze délky fonémů L kontextu nebo R kontextu. Jestliže je zvolen levý (L) kontext pro délky, pak příznakový vektor obsahuje následující položky:

- Pro každý foném kontextu $-L \leq i \leq R$ je použito kódování 1 z n (1 pro správný foném, 0 pro ostatní, angl. one-hot encoding). Celková dimenze kontextu je tak $N_p \times (L + R + 1)$, kde N_p je počet fonémů ve slovníku.

- Druhou množinu příznaků reprezentují otázky použité u fonetických rozhodovacích stromů (viz část **??**). U těchto otázek je opět použito one-hot encoding. Dimenze těchto příznaků je $N_q \times (L + R + 1)$, kde N_q odpovídá celkovému počtu otázek.

- Poslední skupinu příznaků představují délky fonémů L kontextu na pozicích $-L \leq i < 0$. Celková dimenze je L. Neuronová síť nejlépe pracuje s hodnotami v intervalu $(0, 1)$. Jako vhodné se ukázalo normalizovat hodnotu délky $d = 1, 2, \ldots, D$ pomocí sigmoid funkce

$$d' = \frac{2}{1 + e^{-0{,}01d}} - 1, \tag{2.5}$$

  která transformuje hodnoty do požadovaného intervalu $(0, 1)$ [9]. Pokud není kontext k dispozici (krajní případy), tak $d = 0$.

Celková dimenze výsledného příznakového vektoru je pak $I = (L + R + 1) * (N_p + N_q) + L$.

[5] Při vytváření duration modelu byly otestovány oba přístupy a i naše experimenty ukazují, že NN se softmax vrstvou poskytuje lepší výsledky, protože u EL řeči CDEN model přinesl zanedbatelné zlepšení a v některých případech dokonce reskórování způsobilo zhoršení výsledků.

2.4.3 Dosažené výsledky

Stejně jako v případě augmentace dat (viz 2.3.2) je potřeba k natrénování duration modelu kvalitní zarovnání. Jednou z hlavních částí příznakového vektoru modelu je totiž délka L kontextu modelu. K získání co možná nejpřesnějšího zarovnání je použit nejlepší TDNN model natrénovaný na uměle protažených datech[6], viz tab. 2.7.

Samotný duration model (popsaný v předchozí části 2.4.2) je typu feedforward. Počet skrytých vrstev sítě se odvíjí od konkrétní řešené domény, ale standardně se uvažují 2 případně 3, viz [13]. Velikost těchto skrytých vrstev je volena jako násobek dimenze příznakového vektoru, v tomto případě byl zvolena hodnota $3I$. Aktivační funkce je typu RELU. Velikost výstupní vrstvy odpovídá maximálnímu počtu mikrosegmentů D, v [13] bylo dosaženo nejlepších výsledků s $D = 50$. K vytvoření duration modelu posloužil framework Kaldi. Ten představuje obecný framework pro vytváření HMM a DNN řečových modelů.

Ověření funkčnosti duration modelu je provedeno na $2x$ uměle protažených datech[7]. Kontextuální okénko má hodnotu $(L, R) = (3, 3)$, $N_p = 42$ a $N_q = 6$. Velikost vstupního vektoru $I = 339$. Model má 2 skryté vrstvy o velikosti 1017 neuronů. Výstupní vrstva typu softmax má dimenzi $D = 50$. Model je trénován a otestován pomocí stejné trénovací a testovací sady jako modely v části 2.3.2. Jazykový model je fonémový zerogramový. Tento model dosáhl $Acc_p = 88,54$ %, což představuje zlepšení o $0,83$ % absolutně a $7,24$ % relativně oproti TDNN $2x$ modelu pro uměle vytvořenou množinu trénovacích dat ($Acc_p = 87,71$ %). Duration model tedy relativně významně zlepšuje přesnost modelu. Na reálně protažených datech pak tento model dosáhl přesnosti $Acc_p = 85,68$ % (původní TDNN $2x$ model dosáhl $Acc_p = 84,51$ %). Pokud vstupem duration modelu byla neprotažená data, tak přesnost modelu byla pouze $Acc_p = 80,73$ %. Z analýzy chyb pak plyne, že v takovém případě významně přibylo chyb u vybraných neznělých fonémů. Tento výsledek, ale přesně kopíruje očekávání, protože je model natrénován na protaženou podobu.

Mezi hyperparametry modelu patří zejména velikost L a R kontextu, počet vrstev sítě a maximální délka D. Zejména hodnota maximální délka D teoreticky poskytuje největší možnost pro zlepšení výsledků modelu, protože hodnota $D = 50$ byla zvolena na základě experimentů provedených v [13], kde se však pracovalo se standardní neprotaženou řečí. Tab. 2.9 ukazuje vliv maximální délky na přesnost modelu. Speciální je hodnota $D = 189$, která je určena automaticky na základě zarovnání před samotným trénováním. Model s $D = 189$ zároveň dosáhl nejvyšší přesnosti $Acc_p = 88,58$ %, což představuje drobné zlepšení oproti původnímu modelu s $D = 50$.

Dalším hyperparametrem, který může ovlivnit kvalitu modelu, je počet vrstev neuro-

[6] Natrénování TDNN modelu pomocí reálně protažených dat nebylo vhodné, protože nebylo k dispozici dostatečné množství reálně protažených dat.

[7] Hodnota $2x$ je zvolena, protože se nejvíce blíží reálně protaženým datům.

	D				
	50	100	150	189	200
Acc_p [%]	88,54	88,57	88,53	**88,58**	88,57

Tabulka 2.9: Vliv maximální délky na přesnosti modelu.

nové sítě. V tab. 2.10 jsou vypsány výsledky jednotlivých modelů. Varianta *1H* představuje model s 1 skrytou vrstvou, *2H* model s 2 skrytými vrstvami a *3H* s 3 skrytými vrstvami. Speciálním případem jsou modely obsahující bottleneck vrstvu (*2H (bottleneck)* a *3H (bottleneck)*). Ty místo poslední skryté vrstvy o velikosti $3I$ obsahují vrstvu s pouze 10 neurony. Tato vrstva by měla pomoci v zobecňování [13]. Z dosažených výsledků je patrné, že velikost sítě není úplně zásadním parametrem. Rozdíl mezí přesností sítě s 2 a 3 vrstvami je minimální. Přínos bottleneck vrstvy, oproti výsledkům prezentovaným v [13], je také spíše minimální. Nicméně obecně se dá říci, že tato vrstva má pozitivní dopad na přesnost.

	Model				
	1H	2H	3H	2H (bottleneck)	3H (bottleneck)
Acc_p [%]	88,42	88,58	88,56	**88,60**	88,59

Tabulka 2.10: Vliv počtu skrytých vrstev na přesnosti modelu ($D = 189$[8]).

Posledním hyperparametrem, který může mít vliv na přesnost modelu, je velikost L a R kontextu. Z tab. 2.11 a tab. 2.12 vyplývá, že nejlepších výsledků dosahují modely, které mají délku kontext $L + R = 6$. Úplně nejlepšího výsledku pak dosáhl model mající symetrický kontext, ale oproti modelům s asymetrickým kontextem je rozdíl spíše zanedbatelný.

Nejlepšího výsledku tedy dosahuje model mající $D = 189$, 2 skryté vrstvy, kde poslední skrytá vrstva má pouze 10 neuronů a $L + R = 6$. Výsledky však ukazují, že duration model není významně citlivý na změnu parametrů. V případě rozpoznávání reálně protažených dat, dosáhl model přesnosti $Acc_p = 85,93$ %. Pokud se trénovací sada modelu rozšířila o část reálně protažených dat (10 % a 25 % vět) a natrénoval a otestoval se nový model (pomocí zbytku reálně protažených dat), tak výsledná přesnost dosáhla hodnoty $Acc_p = 87,02$ %. Hodnoty přesnosti však nejsou úplně porovnatelné, protože testovací sada není identická, nicméně lze vyvozovat závěr, že pokud by byl model natrénován z dostatečného

[8] V průběhu určování nejlepších kombinace hyperparametrů byly otestovány všechny kombinace velikosti sítě a maximální délky D. Nejlepších výsledků dosahovaly modely s $D = 189$.

	Kontext *(L, R)*				
	(0, 0)	(1, 1)	(2, 2)	(3, 3)	(4, 4)
Acc_p [%]	87,36	87,98	88,47	**88,60**	88,58

Tabulka 2.11: Porovnání vlivu velikosti symetrického kontextu.

	Kontext *(L, R)*				
	(5, 1)	(4, 2)	(3, 3)	(2, 4)	(1, 5)
Acc_p [%]	88,57	88,58	**88,60**	88,57	88,58

Tabulka 2.12: Vliv levého a pravého kontextu v případě, že celková délka $L + R = 6$.

množství reálně protažených dat, tak by se jeho výsledky blížily výsledkům modelu na uměle protažených datech.

2.4.4 Aktualizace výsledků porovnání

V části 2.2 a 2.3.3 jsou prezentovány výsledky srovnání schopností člověka a stroje. V případě člověka jsou zdrojem dva poslechové testy, které prověřily schopnost posluchače nejprve určit význam izolovaných slov, a následně od sebe rozeznat dvě akusticky velmi podobná slova. V případě stroje jsou použity celkem 3 ASR experimenty „one-mil“, „reduced“ a „bigrams“. V tab. 2.13 jsou pak předchozí výsledky doplněny o hodnoty dosažené duration modelem jehož parametry odpovídají nejlepšímu modelu z předchozí části 2.4.2.

Stejně jako v předchozích experimentech je u „one-mil“ experimentu použit zerogramový jazykový model s 1 milionem slov, „reduced“ obsahuje pouze slova obsažená v poslechovém testu. V případě „bigrams“ je pro každou položku generován speciální LM obsahující 4 kombinace slov. Použití duration modelu nepřineslo významné zlepšení výsledků *augmented* modelu. Tento výsledek, ale není překvapivý, protože *augmented* model dosahuje teoreticky maximálních možných hodnot. Největšího zlepšení dosáhl duration model v případě „one-mil“ experimentu. Zde došlo ke zlepšení o $1,42$ % absolutně, což je o $10,42$ % relativně. Rozhodně se jedná o významné zlepšení.

Použití reskóringu pomocí duration modelu ještě zlepšuje výsledky TDNN modelu natrénovaného na protažených datech. Oba modely navíc dokáží pracovat i s řečníkem reálně protaženými daty, což umožňuje i reálné použití těchto modelů. Hlavní nedostatek aktuální implementace reskóringu pomocí duration modelu spočívá v tom, že se jedná o

	Acc_p [%]		
	one-mil	reduced	bigrams
člověk	*70,47*	*70,47*	*66,24*
stroj (baseline)	*61,24*	*69,91*	*54,82*
stroj (augmented)	*84,95*	*94,36*	*98,80*
stroj (duration model)	86,37	94,42	98,81

Tabulka 2.13: Aktualizované porovnání dosažených výsledků člověka a stroje.

offline přístup. Promluvy jsou nejprve zpracovány pomocí TDNN modelu a až následně jsou kompletní výstupy modelu reskórovány pomocí duration modelu. Změna modelu a principu reskórování tak, aby byl schopen pracovat i v online režimu, je otázkou budoucího výzkumu, ale principiálně tomu nic nebrání.

2.5 Trenažér

V předchozích částech (2.3 a 2.4) byla rozvíjena a ověřována myšlenka doplnění chybějící informace, ke které došlo v důsledku ztráty hlasivek, pomocí protahování určitých vybraných fonémů, zejména pak /k/, /p/, /s/, /š/, /t/, /t′/ a /v/ reprezentující neznělé fonémy. Prezentované výsledky (viz tab. 2.7 a 2.9) prokazují, že daný přístup poskytuje výrazné zlepšení přesnosti ASR systému, zejména u promluv s minimálním kontextem.

Hlavní problém tohoto přístupu spočívá v protažení zájmových fonémů samotným řečníkem. Prezentovaný přístup totiž nepočítá s protažením celého slova, ale pouze nezbytně nutné části, fonému. Nicméně s trochou pomoci je řečník schopen protáhnout požadovanou část slova. Výsledky prezentované v části 2.4.2 demostrují, že model natrénovaný na uměle protažených datech je schopen lépe rozpoznávat i reálně protažené fonémy. V případě rozpoznávání neprotažených dat pak přesnost modelu významně klesá. Tento výsledek je fundamentálním předpokladem pro myšlenku trenažéru. Jeho hlavní funkcí je pomoci řečníkovi naučit se automaticky protahovat inkriminované fonémy tak, aby přesnost rozpoznávání byla maximální. Zároveň je možné pomocí trenažéru postupně adaptovat akustický model na základě reálných dat. Postupem času by tak měly být eliminovány všechny chyby v datech způsobené umělým protažením.

Samotný trenažér si lze představit jako počítačový program, který řečníkovi zobrazuje jednotlivá slova/věty a ten je musí vyslovit. Primární funkcí trenažéru je pomoc řeční-

kovi s učením automatického protahování. Z tohoto důvodu je jeho součástí ASR systém s individuálním modelem, který slouží k rozpoznávání vyřčených promluv. O výsledku rozpoznávání (správně/špatně) je uživatel srozuměn. V případě úspěšného pokusu je promluva uložena a řečník může pokračovat v další promluvě, pokud se ji nerozhodne přeskočit. U protahovaných fonémů je použito zdvojeného zápisu (např. *„kossa“*, viz 2.3.4). Ten se ukázal jako velmi názorný a podvědomě nutící řečníka vyslovit daný foném „jinak“.

Sekundární funkcí trenažéru je adaptace akustického modelu na základě reálně protažených dat. Originální duration model je vytvořen pomocí uměle protažených dat. Tím, jak řečník postupně více a více úspěšně reprodukuje požadované promluvy, je model postupně adaptován reálnými daty. Tím se všechny případné nedostatky, způsobené uměle protaženými daty, postupně odstraňují. Kompletní proces vytvoření adaptovaného duration modelu s pomocí trenažéru je následující:

1. Získání co možná největšího množství řečových dat (v řádech hodin).
2. Vytvoření ASR systému k získání co možná nejpřesnějšího zarovnání.
3. Umělé protažení dat na základě zarovnání.
4. Vytvoření akustického duration modelu, který je použit v trenažéru.
5. Adaptace řečníka a modelu na základě úspěšně rozpoznaných promluv.
6. Použití adaptovaného modelu[9].

Adaptovaný model je pak možné použít, ve spojení s TTS a původním hlasem řečníka, např. při telefonování. Což v počátečních fázích života po TL může rapidně zvýšit kvalitu života i psychický stav pacienta. [14]

Jednou z prerekvizit trenažéru je možnost individuálního použití doma. Zejména proto, že odpadá nutnost použití specializovaného HW či zvukové komory. Nezanedbatelným benefitem je pak flexibilita, kterou řečník má. Může trenažér používat v pro něj, ideální době a prostředí. Pilotní projekt pro získávání dat řečníků (pro účely TTS) pořízených v domácím prostředí na vlastním HW je prezentován v [15], [16].

Při vytváření trenažéru je nepochybně ještě potřeba zodpovědět mnoho otázek, např. jak často adaptovat akustické modely, z jakého množství dat, či zda to provádět na serveru či lokálně. A priori se však jedná spíše o implementační detaily, než nezbytné konceptuální otázky. I přesto, že v současném stavu jsou duration modely vhodné pouze k offline zpracování, tak je de facto možné započít práce na vytváření trenažéru. Principiálně totiž

[9] V případě dostatečného množství reálně protažených dat, pak natrénování nového modelu na reálných datech.

není problém, pokud odpověď na otázku, zda je promluva správně/špatně, bude dostupná až po nezbytně nutné krátké době po skončení promluvy nebo ještě v průběhu.

Fundamentálním předpokladem funkčnosti trenažéru je schopnost určit správnost protažení, jinými slovy správně rozpoznat protažené slovo a neprotažené naopak označit jako špatné. Tento předpoklad podporují výsledky v tab. 2.8, kde je jasně vidět významný pokles přesnosti u neprotažených dat. Z analýz výsledků plyne, že většina chyb (oproti optimální situaci) je právě v protažených fonémech. Vytvořený trenažér by tak měl být schopen ve většině případů určit zda bylo slovo správně protaženo či nikoliv.

Samotný trenažér je samozřejmě jen prostředek k tomu, aby bylo dosaženo ASR systému, který bude schopen co možná nejlépe pracovat s TL řečí, a tím pádem zlepšit kvalitu života řečníků postižených ztrátou hlasivek.

Závěr

Předložená disertační práce se zabývá problematikou rozpoznávání řeči pacientů po totální laryngektomii, kteří komunikují pomocí elektrolarynxu. Motivací pro zpracování tohoto tématu bylo obohatit stávající postupy využívané pro rehabilitaci hlasu o možnosti, které přináší využití moderních technologií, konkrétně možností automatického rozpoznávání řeči. V kapitole **??** byly vytyčeny cíle práce, jejichž naplnění bylo v následujících odstavcích zhodnoceno.

V kapitole **??** jsou přiblíženy nejčastější příčiny ztráty hlasu a metody užívané k jeho rehabilitaci. Pomocí klasických rehabilitačních technik lze pacientům navrátit možnost mluvit, ale kvalita produkované řeči nutně nemusí splňovat očekávání pacientů a požadavky kladené na mluvčího okolím. Například použití elektrolarynxu sice neklade na uživatele vysoké nároky co se týče učení, ale kvalita řeči není vůbec přirozená. Oproti tomu pomocí jícnového hlasu je produkován relativně kvalitní hlas, ale k edukaci je potřeba vynaložit opravdu nemalé úsilí. Jako ideální se může jevit použití tracheoezofageální píštěle, která umožňuje proudění vzduchu z plic do dutiny ústní. Produkovaný hlas se v tomto případě vyznačuje vysokou kvalitou, dobrou srozumitelností, individuálním zabarvením a relativně dlouhou fonační dobou. Za nedostatek se dá považovat nutnost pravidelně čistit a měnit píštěle. Existují i další metody, ale ty jsou zatím používány spíše autorskými týmy a o masovém použití se rozhodně nedá hovořit. Bohužel žádná z technik nepředstavuje univerzální řešení, a proto se je lékaři stále snaží zdokonalovat, a tím zkvalitňovat život pacientů. U všech aktuálně používaných metod rehabilitace hlasu je patrný významný negativní dopad na psychiku pacienta, který se musí vyrovnat nejen se ztrátou vlastního hlasu, ale i s ostychem, který provází opětovné snahy mluvit.

Pomoc s rehabilitací hlasu mohou poskytnout řečové technologie zpracovávající přirozenou řeč. V současnosti využívané obecné systémy automatického rozpoznávání řeči (ASR systémy) ale poskytují spolehlivé výsledky v případě rozpoznávání promluv zdravého řečníka. V případě, že se charakteristiky rozpoznávané řeči příliš líší (např. řeč obsahuje vyšší množství šumu), může se u běžně uživaných ASR systémů projevit jejich nedostačná robustnost. Proto bylo pro potřeby návrhu ASR systému, který bude sloužit pro rozpoznávání řeči pacientů po totální laryngektomii, nutno pořídit řecová data odpovídající kvality. V takovýchto promluvách se ukazuje jako problematická zejména přílišná podobnost produkovaných znělých a neznělých fonémů. Proto byla navázana spolupráce s mluvčí, která prodělala TL a komunikuje pomocí elektrolarynxu. V průběhu pěti let byly pořízeny 3

sady nahrávek, což odpovídá necelým 15 hodinám řečových dat. První sada je složená z vět, které jsou součástí interně využívaného řečového korpusu, 2. sada rozšiřuje řečový korpus o další sadu vět a problematických izolovaných slov. Ve 3. sadě jsou obsaženy další věty a slova respektující protažení problematických fonémů. S ohledem na toto lze i vytyčený cíl č. 2 považovat za naplněný.

Pro otestování robustnosti obecného ASR systému byla využita data poskytnutá řečníkem po totální laryngektomii. Testovaný systém vykázal pro trigramový jazykový model obsahující 1 milion unikátních slov úspěšnost rozpoznávání slov pouze **18, 49 %**. Takto nízká přesnost rozpoznávání indikovala nutnost navrhnout individuální akustický model, který bude reflektovat specifika řeči produkované s využitím EL. Při využití dat z EL korpusu dosáhl systém přesnosti rozpoznávání slov **83, 33 %**.

Následně byl minimalizován vliv jazykového modelu, a byly hledány optimální parametry akustického modelu. Byl ověřen vliv maximálního počtu unikátních stavů HMM modelu a vzorkovací frekvence na přesnost rozpoznávání. Nejlepších výsledků bylo dosaženo pro model pracující s maximálně 4096 unikátními stavy a vzorkovací frekvencí 16 kHz. Pro HMM-GMM bylo dosaženo přesnosti rozpoznávání **81, 20 %**, pro HMM-DNN pak **85, 23 %**. Po provedení analýzy získaných výsledků se jako problematické ukázalo rozpoznávání neznělých fonémů, proto bylo přistoupeno k redukci fonetické sady prostřednictvím různých kombinací náhrady neznělých fonémů za znělé, což ve většině případů nemělo pozitivní dopad. Proto byly navrženy další úpravy ASR systému, konkrétně protaženy neznělé fonémy a následně byl ASR systém rozšířen o tzv. duration model akcentující právě délku fonémů. Na základě provedených experimentů se jako vhodné ukázalo prodloužit neznělé fonémy na dvojnásobek jejich původní délky. Úspěšnost rozšířeného modelu dosahovala **88, 54 %**.

S ohledem na získané výsledky vyvstala potřeba porovnat schopnosti rozpoznávání člověka a stroje. Za tímto účelem byl navržen tzv. poslechový test, jehož princip byl přiblížen v části 2.2. Pro model s výše navrženými optimálními parametry (max. 4096 unikatní stavů a vzorkovací frekvencí 16 kHz) stroj dosál přesnosti rozpoznávání **69, 91 %** pro případ izolovaných slov a **54, 82 %** pro případ bigramů. U člověka bylo dosaženo přesnosti **74, 47 %**, resp. **66, 24 %**. Při zohlednění umělého protažení akustických dat dosáhl stroj přesnosti **94, 36 %** pro izolovaná slova, resp. **98, 80 %** pro bigramy. Po následném rozšíření systému rozpoznávání o duration model dosáhl stroj úspěšnosti **94, 42 %**, resp. **98, 81 %**. S ohledem na výsledky poskytnuté ASR systémem byla nahrána další sada řečových dat respektující protažení neznělých fonémů. Na takto rozšířené testovací sadě bylo dosaženo přesnosti rozpoznávání **87, 02 %** na úrovni fonémů. Tím bylo ověřeno, že model natrénovaný na uměle protažených datech je schopen rozpoznávat i data reálná, a lze ho tedy využít jako základ pro vývoj trenažéru, který bude výukovým nástrojem pro osvojení schopnosti řečníka protahovat neznělé fonémy.

S ohledem na výše uvedené výsledky a z nich vyvozené závěry lze říci, že vytyčené cíle

disertační práce byly naplněny a s ohledem na aktuálnost řešené problematiky lze získané poznatky využít jako základ další práce.

Výběr použité literatury

[1] Bruce Denby et al. "Silent speech interfaces". In: *Speech Communication* 52.4 (dub. 2010), s. 270–287. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.08.002.

[2] Robin Hofe et al. "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing". In: *Speech Communication* 55.1 (led. 2013), s. 22–32. ISSN: 0167-6393. DOI: 10.1016/j.specom.2012.02.001.

[3] Thomas Hueber et al. "Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips". In: *Speech Communication* 52.4 (dub. 2010), s. 288–300. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.11.004.

[4] Michael Fagan et al. "Development of a (silent) speech recognition system for patients following laryngectomy." In: *Medical engineering & physics* 30.4 (květ. 2008), s. 419–25. ISSN: 1350-4533. DOI: 10.1016/j.medengphy.2007.05.003.

[5] Charles Jorgensen a Sorin Dusan. "Speech interfaces based upon surface electromyography". In: *Speech Communication* 52.4 (dub. 2010), s. 354–366. ISSN: 01676393. DOI: 10.1016/j.specom.2009.11.003.

[6] Tatsuya Hirahara et al. "Silent-speech enhancement using body-conducted vocal-tract resonance signals". In: *Speech Communication* 52.4 (dub. 2010), s. 301–313. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.12.001.

[7] Robin Hofe et al. "Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA." In: *INTERSPEECH*. August. 2011, s. 3009–3012.

[8] Lawrence R Rabiner. "A tutorial on hidden Markov models and selected applications in speech recognition". In: *Proceedings of the IEEE* 77.2 (1989), s. 257–286.

[9] Tanel Alumäe. "Neural network phone duration model for speech recognition". In: *Interspeech 2014*. Singapore, 2014.

[10] Anastasios Anastasakos, Richard Schwartz a Han Shu. "Duration modeling in large vocabulary speech recognition". In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Sv. 1. IEEE. 1995, s. 628–631.

[11] VR Rao Gadde. "Modeling word duration for better speech recognition". In: *Proceedings of NIST Speech Transcription Workshop*. 2000.

[12] Janne Pylkkonen a Mikko Kurimo. “Duration modeling techniques for continuous speech recognition”. In: *Eighth International Conference on Spoken Language Processing*. 2004.

[13] Hossein Hadian et al. “Phone Duration Modeling for LVCSR Using Neural Networks.” In: *INTERSPEECH*. 2017, s. 518–522.

[14] Jiří Mertl, Eva Žáčková a Barbora Řepová. “Quality of life of patients after total laryngectomy: the struggle against stigmatization and social exclusion using speech synthesis”. In: *Disability and Rehabilitation: Assistive Technology* 13.4 (2018). PMID: 28447495, s. 342–352. DOI: 10.1080/17483107.2017.1319428. eprint: https://doi.org/10.1080/17483107.2017.1319428.

[15] Markéta Jůzová, Jan Romportl a Daniel Tihelka. “Speech Corpus Preparation for Voice Banking of Laryngectomised Patients”. In: *Text, Speech, and Dialogue*. Ed. Pavel Král a Václav Matoušek. Cham: Springer International Publishing, 2015, s. 282–290. ISBN: 978-3-319-24033-6.

[16] Markéta Jůzová et al. “Voice Conservation and TTS System for People Facing Total Laryngectomy”. In: *Proc. Interspeech 2017*. 2017, s. 3425–3426.