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.

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 16kHz. 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 **??**. 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.