

ZÁPADOČESKÁ UNIVERZITA V PLZNI

FAKULTA APLIKOVANÝCH VĚD

KATEDRA KYBERNETIKY

TBD

DIZERTAČNÍ PRÁCE

LEDEN 2020

AUTOR: ING. PETR STANISLAV

ŠKOLITEL: PROF. ING. JOSEF PSUTKA, CSc.

OBOR: KYBERNETIKA

Obsah

1	Úvod	1
2	Motivace a cíle disertační práce	3
3	Příčiny ztráty hlasu a možnosti jeho rehabilitace	4
3.1	Příčiny ztráty hlasu	4
3.1.1	Rakovina hrtanu	5
3.2	Rehabilitace hlasu po totální laryngektomii	9
3.2.1	Foniatické metody	10
3.2.2	Chirurgicko-protetická metoda	14
3.2.3	Hrtanu podobné struktury	17
3.2.4	Transplantace hrtanu	19
3.2.5	Shrnutí	20
4	Automatické rozpoznávání řeči	23
4.1	Parametrizace řečového signálu	25
4.1.1	Modelování produkce řeči	25
4.1.2	Modelování procesu slyšení	30
4.2	Akustické modelování	39
4.2.1	Struktura skrytého Markovova modelu	41
4.2.2	Trénování parametrů HMM s Gaussovskými směsmi	43

4.2.3	Využití neuronových sítí	45
4.3	Jazykové modelování	53
4.4	Dekódování	54
5	Konstrukce ASR systému pro uživatele po totální laryngektomii hovořící pomocí elektrolarynxu	58
5.1	Vytvoření řečového korpusu EL promluv	58
5.2	Analýza akustického signálu a jeho parametrizace	62
5.2.1	Analýza získaných dat	63
5.3	Aplikace obecného systému rozpoznávání a dosažené výsledky	68
5.3.1	Hledání optimálních parametrů baseline modelu	69
5.3.2	Redukce fonetické sady	73
6	Návrh a realizace úprav ASR	80
6.1	Doplnění řečového korpusu o specifická data - vliv nových dat na kvalitu AM	80
6.1.1	Vliv nových dat na kvalitu modelů	84
6.1.2	Eliminace vlivu kanálu	86
6.2	Poslechový test a porovnání výsledků člověka a stroje	91
6.2.1	Izolovaná slova	92
6.2.2	Slovní bigramy	94
6.2.3	Výsledky porovnání	96
6.3	Augmentace dat	99
6.3.1	Protažení na příznacích	101
6.3.2	Protažení na zvuku	104
6.3.3	Aktualizace výsledků porovnání	110
6.3.4	Reálně protažená data	111
6.4	Model akcentující protažení dat	114

6.4.1	Princip explicitních duration modelů	115
6.4.2	Duration model se softmax vrstvou	117
6.4.3	Dosažené výsledky	118
6.4.4	Aktualizace výsledků porovnání	121
6.5	Trenažér	123
7	Závěr	126
	Seznam použité literatury	129

Kapitola 1

Úvod

Lidská řeč je jedním z hlavních dorozumívacích prostředků užívaných člověkem. Její ztráta způsobuje řadu komplikací. Příčinou ztráty může být chirurgický zákrok v důsledku nádorovitého onemocnění nebo poškození hrtanu vlivem traumatické nehody. S ohledem na zvýšení kvality života se už od konce 19. století lékaři snaží o rehabilitaci pacientova hlasu. Mezi nejpoužívanější přístupy patří chirurgicko-protetické a foniatrické metody.

První snahy o navrácení schopnosti mluvit nebyly příliš úspěšné a svým způsobem i životu nebezpečné. Přesto neutuchající snaha lékařů postupně vyústila v bezpečné a běžně používané metody. Mezi nejpoužívanější se řadí využití elektrolarynxu, jícnového hlasu a tracheoezofageální pštěle. Bohužel žádná z používaných metod není univerzálním řešením pro každého pacienta. V mnoha případech je navíc případné používaní spojené s nemalou psychickou zátěží mluvčího, který se může například ostýchat mluvit na veřejnosti. Z tohoto důvodu je této problematice věnována nemalá pozornost a další pomoc mohou přinést řečové technologie.

V polovině 20. století se s rozvojem číslicových počítačů začaly objevovat snahy o zpracování přirozené řeči počítačem. Toto úsilí vyústilo v (v dnešní době) hojně užívané systémy automatického rozpoznávání (zkr. ASR) a syntézy 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 videozáznam artikulace rtů. Bohužel multimodální systémy zatím nedosahují požadovaných kvalit a tak se vývoj ASR systému ubírá zejména směrem vývoje komplexnějších modelů.

Běžné systémy rozpoznávání řeči jsou však trénovány na obecných datech a pro uživatele postižené trvalou ztrátou hlasu jsou nepoužitelné. Hlavním problémem se jeví jiné charakteristiky produkované řeči a ztráta určitého množství informace v ní obsažené. Tato ztráta je důsledkem aktuálně používaných metod rehabilitace hlasu, které se snaží nahradit chybějící buzení hlasivek jiným, ale ve své podstatě konstantním, zdrojem buzení. Obecné ASR systémy pak nejsou bez adaptace schopné obstojně tuto řeč zpracovávat.

Většina doposud vyvíjených metod se tuto ztracenou informaci snaží získat pomocí zapojení dalšího 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.

Kapitola 2

Motivace a cíle disertační práce

1. Seznamte se s přístupy, které umožňují alespoň částečnou obnovu schopnosti řečové komunikace u pacientů po totální laryngektomii.
2. Pro účely konstrukce systému automatického rozpoznávání řeči lidé po totální laryngektomii využívajících pro komunikaci elektrolarynx navrhněte a poříďte vhodný korpus řečových nahrávek.
3. Natrénujte základní systém rozpoznávání řeči pro jednoho řečníka - pacienta po totální laryngektomii mluvícího pomocí elektrolarynxu - a porovnejte funkcionality systému (zejména jeho přesnost) se systémem rozpoznávajícím řeč zdravých lidí. Ke konstrukci systému využijte state-of-the-art metod.
4. Analyzujte základní příčiny případné zvýšené chybovosti realizovaného systému rozpoznávání řeči a pokuste se navrhnut vzhodné úpravy v jeho konstrukci, které chybovost sníží. Diskutujte vhodnost navrženého řešení.

Kapitola 3

Příčiny ztráty hlasu a možnosti jeho rehabilitace

Lidská řeč tvoří jeden ze stavebních kamenů lidského dorozumívání. Pro člověka postiženého dočasnou či trvalou ztrátou hlasu představuje běžná lidská komunikace mnohem náročnější úkol než pro člověka zdravého. Takový jedinec se musí dennodenně potýkat s problémy, které by za normálních okolností řešit nemusel. V mnoha případech doprovází ztrátu hlasu i zvýšená psychická zátěž, například strach z reakce okolí, a proto se problematice lidí trpící ztrátou hlasu věnuje nemalá pozornost. V této kapitole si nejprve v části 3.1 přiblížíme možné příčiny ztráty hlasivek, a tedy i trvalé ztráty hlasu, a následně v části 3.2 představíme dostupné metody rehabilitace hlasu.

3.1 Příčiny ztráty hlasu

Nejčastěji je trvalá ztráta hlasu zapříčiněna chirurgickým zákrokem zvaným totální laryngektomie¹ neboli úplné odstranění hrtanu. Odstraněním hrtanu, a tím i hlasivek (glottis), přichází člověk o schopnost rozvibrovat vzduch vycházející z plic, který je dále modulován artikulačním ústrojím. Nejběžnější příčinou vedoucí k totální laryn-

¹laryngektomie: larynx, laryngos - hrtan, ectos, ectomia - odstranění, vynětí

gektomii představuje rakovina hrtanu v pokročilém stádiu. V mnohem nižší míře je na vině rakovina hltanu či poškození hrtanu automobilovou nebo jinou traumatickou nehodou. Podle [1] přibude v České republice ročně přibližně 400 nových onemocnění rakoviny hrtanu, z toho je přibližně jedna třetina léčena pomocí totální laryngektomie. To představuje více než 100 nových případů trvalé ztráty hlasu každý rok.

3.1.1 Rakovina hrtanu

Jak již bylo zmíněno, rakovina hrtanu je jedním z hlavních důvodů odstranění hrtanu. Tento typ rakoviny postihuje převážně muže ve věku 50-60 let. V posledních letech je však zřejmý trend snižujícího se průměrného věku pacientů [2]. Z celkovém počtu pacientů zhruba 20% představují ženy.

Přesná příčina vzniku nádorovitého onemocnění hrtanu doposud není známa, ale z průzkumů je zřejmá korelace mezi vznikem rakoviny a konzumací alkoholu či kouřením. Jinými slovy, mezi rizikovou skupinu patří lidé, kteří jsou pravidelně vystavováni vlivu kouření, ať již aktivně (sami kouří) či pasivně (vdechují cigaretový kouř) a zároveň si dopřávají nemalé množství alkoholu. Podle [2] 90% pacientů aktivně kuří.

Příznaky onemocnění

Vznikající nádorové onemocnění v oblasti hrtanu se může projevovat různými způsoby. Mezi hlavní faktory ovlivňující počáteční příznaky patří umístění a velikost nádoru.

Již ve velmi raném stádiu, kdy je nádor umístěn přímo na hlasivkách, je příznakem **chrapot**. Ve většině případů se samozřejmě jedná o krátkodobé postižení hlasivek virovou infekcí. Nicméně pokud chrapot trvá déle než tři týdny, je již doporučováno navštívit odborného lékaře. U nádorů nacházejících se ve vchodu do hrtanu a v polykacích cestách se mohou jako příznak objevovat **polykací obtíže**. Mezi další možné příznaky patří **bolesti v krku**, jednostranné bolesti vystřelující do ucha či nepříjemný pocit při polykání. I v tomto případě krátkodobý výskyt nemusí nutně

znamenat rakovinu hrtanu, nicméně při obtížích trvajících déle než měsíc je doporučováno důkladné vyšetření lékařem. Na základě umístění nádoru se může objevovat **dráždivý kašel** s možným vykašláváním krve. Dalším možným příznakem je vznik **zduření na krku**. V tomto případě je vhodné neprodleně vyhledat lékaře.

Z výše uvedených příznaků je zřejmé, že první indicie o vážném onemocnění mohou být podceněny, a tím značně snížena šance na plné uzdravení pacienta. V případě včasného diagnostikování rakoviny hrtanu či hltanu je možnost úplného vyléčení pacienta bez trvalých následků více než 90% [1].

Léčba nádorového onemocnění

U nádorovitých onemocnění se v drtivé většině případů využívá **chirurgické léčby**, **aktinoterapie** neboli ozařování a **chemoterapie**. Nejinak tomu je i v případě rakoviny hrtanu a polykacích cest obecně. Majoritní část pacientů je zpravidla nejprve konfrontována s chirurgickou léčbou. Mezi nejčastější zákroky patří **tracheostomie**, parciální laryngektomie, **totální laryngektomie** a chordektomie. V rámci této práce budou blíže popsány pouze léčebné postupy přímo související s úplným odstraněním hrtanu a hlasivek.

Totální laryngektomie (TL), jak už název napovídá, představuje chirurgický zákrok, při kterém je úplně odstraněn hrtan. V určitých případech může být odstraněna i menší či větší část hltanu. Tento zákrok, ve spojení s léčbou rakoviny hrtanu, poprvé vykonal Dr. Theodor Billroth 31. prosince roku 1873 ve Vídni [3] a do dnešní doby přežil v podstatě nezměněn. Cílem této operace je odstranění orgánu zasaženého rakovinným bujením, a tedy zamezení dalšího šíření nemoci. Součástí hrtanu je také hrtanová příklopka (latinsky epiglottis), která zamezuje vdechnutí potravy nebo tekutin do dýchacích cest. Po odstranění hrtanu by tedy potrava a tekutiny mohly být vdechnuty do plic a z tohoto důvodu jsou jícn a průdušnice trvale odděleny. Rozdíl mezi zdravým člověkem a osobou po totální laryngektomii je znázorněn na obr. 3.1.



Obrázek 3.1: (A) Schéma dýchacích cest zdravého člověka (B) Schéma dýchacích cest po totální laryngektomii

Odstraněním spojení průdušnice a jícnu je zahrzena cesta vzduchu do plic. Z tohoto důvodu je nezbytné společně s TL vykonat také **tracheostomii**. Cílem tohoto zákroku není léčba nádorovitého onemocnění, nýbrž vytvoření vstupu pro vzduch směřující do plic a z nich. Samotný zákrok se využívá i v situacích, kdy dojde k uzávěře hrtanu a postižená osoba se dusí. K tomuto může dojít například při alergické reakci na včelí bodnutí, otoku hrtanu, úrazu apod. Při tracheostomii se provádí řez skrz kůži a průdušnici. Do vzniklého otvoru se zavádí kanya, která slouží k dýchání. Místo výkonu zákroku a princip kanyly je znázorněn na obr. 3.2. Pro lepší názornost je znázorněna tracheostomie se zavedenou kanylou u zdravého člověka. Výsledek operace může být dočasný (například v případě alergické reakce) nebo trvalý.

Další významnou metodou využívanou k léčbě nádorových onemocnění představuje **aktinoterapie** neboli léčba ozařováním. Podstatou postupu je periodické vystavování buňky ionizujícímu záření. Energie z tohoto záření je předávána buňce, která je tím poškozována. Tento postup se opírá o předpoklad, že nádorové buňky jsou náchylnější k poškození. Záření ovšem ovlivňuje i zdravé buňky, a proto je tato léčba pro organismus velkou zátěží. Aktinoterapie je možné využít jednak u případů, kdy



Obrázek 3.2: Tracheostomie

je cílem terapie úplné vyléčení, tak i v případech, kdy naprosté odstranění onemocnění není možné. V druhém případě je metoda využívána k prodloužení a zkvalitnění života [1].

Aktinoterapii je možné využít jako hlavní léčebnou metodu (primární aktinoterapie) nebo i ve spojení s ostatními metodami. V případě primární se k léčbě využívá pouze ozařování a cílem je úplné odstranění všech defektních buněk. Z podstaty metody, zejména dopadů léčby na lidský organismus, je zřejmé, že tímto způsobem je ve většině případů možno léčit pouze malé nádory.

Ve spojení s chirurgickou léčbou rozlišujeme předoperační, pooperační nebo tzv. sandwich (tj. před a po chirurgickém zákroku) aktinoterapii. Předoperační ozařování je užíváno v případech, kdy není možné původní nádor vyoperovat. Cílem je zmenšení tumoru do takové míry, aby jej bylo možné chirurgicky odstranit. Někdy je předoperační ozařování spojeno s chemoterapií. U pooperační aktinoterapie je záměrem odstranění potencionálních mikroskopických zbytků tumoru, které by mohly znova začít růst.

Velmi často se ve spojení s léčbou rakoviny mluví i o proceduře zvané **chemoterapie**. Podstatou je podávání léků zastavujících buněčné dělení, tzv. cytostatik. Zjednodušeně řečeno se jedná o velmi toxiccký koktejl látek sloužící k zahubení buňky tím, že poškodí určitou její část a zastaví tak proces dělení. Na tuto léčbu jsou citlivé převážně rychle se dělící buňky. Právě defektní buňky v tumoru mají obvykle určitým způsobem poškozeny opravné mechanizmy a cytostatiky zasažená rakovinná buňka tak s větší pravděpodobností zahyne. Samozřejmě nelze u chemoterapie hovořit o přesně zacílené léčbě. Cytostatika postihují všechny buňky v lidském těle, a proto je možná namísto srovnání s kobercovým bombardováním. S aplikací cytostatik je tak spojena celá řada vedlejších rizik. Mezi nejzávažnější patří poškození ledvin nebo poškození krvetvorby.

Z výše uvedeného je zřejmé, že postižená osoba má velkou šanci na kompletní vyléčení. V mnoha případech má však pacient trvalé následky (trvalá ztráta hlasu) z důvodu podcenění prvotních příznaků vážného onemocnění.

3.2 Rehabilitace hlasu po totální laryngektomii

Nesporná výhoda totální laryngektomie neoddiskutovatelně spočívá v likvidaci primárního nádorového onemocnění. Následky operace však s sebou nesou obrovský zásah do kvality života pacienta. Okem nejviditelnější změnu představuje přítomnost tracheostomie a s ní spojený způsob dýchání. Tato skutečnost má spoustu, na první pohled neúplně očividných, následků. Postižený člověk ztrácí přirozené zvlhčování, ohřev a filtraci vdechovaného vzduchu, jež má za následek vyšší náchylnost k respiračním onemocněním. Příčina spočívá v průchodu vzduchu do průdušnice přes tracheostomii a nikoli přes nosní dutiny.

Pro samotného pacienta je však nejspíše nejobtížnější se vypořádat s trvalou ztrátou vlastního hlasu. Z tohoto důvodu se již samotný autor procedury doktor Billroth

zaobíral otázkou rehabilitace hlasu. Jeho první pokusy s kovovou tracheostomickou kanyoulou sice umožňovaly pacientovi hovořit, ale svou konstrukcí pacienta spíše ohrožovaly na životě [4]. Proto se více uchytila metoda tzv. jícnového hlasu [5]. Ve stejnou dobu, tedy přibližně začátkem minulého století, se začaly objevovat první interní a externí hlasové aparáty. V současnosti je rehabilitace hlasu možná pomocí:

- **foniatrických metod**, mezi které patří jícnový hlas a elektrolarynx,
- **chirurgicko-protetickým způsobem**, který spočívá ve vytvoření kanálku skrze stěnu mezi průdušnicí a jícnem,
- **vytvoření hrtanu podobných struktur chirurgickým způsobem,**
- **transplantace hrtanu.**

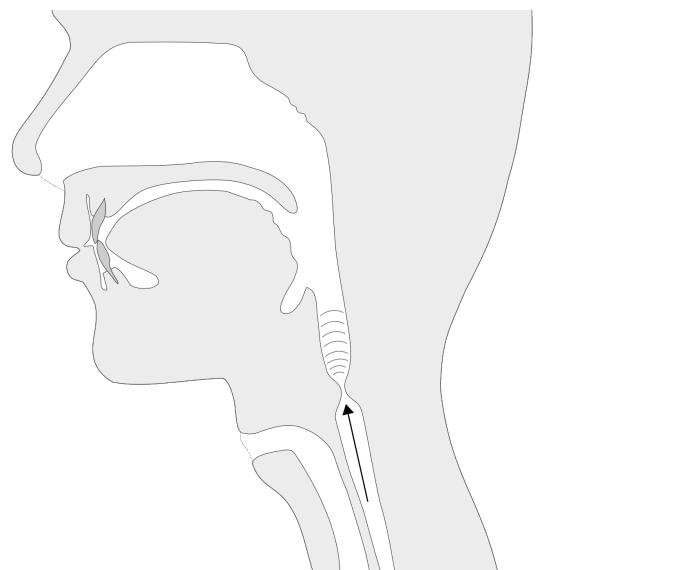
Z uvedeného výčtu se může zdát, že máme k dispozici relativně širokou škálu možností, jak pacientovi vrátit schopnost vyjadřování pomocí mluvené řeči. Ovšem je nutné si uvědomit, že je potřeba volit konkrétní metodu podle stavu a možností pacienta. Jinými slovy, ne každá metoda se hodí pro každého pacienta a žádná z metod není univerzální pro všechny pacienty.

3.2.1 Foniatrické metody

Ačkoli odstranění hrtanu vyústí ve ztrátu hlasu, neznamená to, že byla úplně eliminována schopnost produkovat řeč. V procesu vytváření hlasu zastává odstraněný orgán pouze (i když velmi zásadní) roli generátoru zvuku. Zbylé orgány (hrdelní, nosní a ústní dutina a další) zůstávají nedotčeny a mohou i nadále plnit svou funkci. Logicky se tak nabízí myšlenka nahradit chybějící zdroj zvuku jiným. Mezi nejpoužívanější metody patří jícnový hlas a elektrolarynx.

Jícnový hlas

Počátek této metody se datuje do roku 1922, kdy si prof. MUDr. Miloslav Seeman [6] uvědomil, že funkci štěrbiny mezi hlasivkami (rima glottidis) přebírá tzv. pseudo-glottis, která se vytváří na úrovni horního jícnového svěrače. Zároveň vypracoval a popsal metodiku vytváření jícnového hlasu, při které se vzduch neplní do plic, ale do jíncu. Tato metoda se nazývá **aspirační**. Princip spočívá v aktivním otevření jícnového svěrače, nasáváním a vtlačováním vzduchu do jíncu pomocí polykání. Naplněním jíncu vzduchem si pacient připravuje potřebný vzduch k následné eruktaci² vzduchu a produkci řeči. Vlastní jícnový hlas poté vzniká na přechodu jíncu a hypofaryngu (spodní část hltanu). V této oblasti horního jícnového zúžení dochází k rozkmitání sliznice a podslizniční vrstvy a produkci zvuku, který je následně modulován stejně jako v případě přirozené produkce řeči. Princip tvorby „základního“ tónu jícnového hlasu je znázorněn na obr. 3.3.



Obrázek 3.3: Princip tvorby jícnového hlasu. Průchodem vzduchu přes zúžení vzniká základní tón jícnového hlasu.

²eruktace - latinsky název pro proces říhání (popřípadě krkání), při kterém dochází k úniku plynů pocházejících ze žaludku dutinou ústní.

Kromě aspirační metody je ještě možné se setkat s metodou **injekční**. Hlavní rozdíl spočívá v principu plnění vzduchu do jícnu. Při aspirační metodě se využívá polykání, zatímco v tomto případě je využito kořene jazyka, kterým je vzduch vtlačován do jícnu. Následný princip produkce hlasu je již shodný s původní metodou. S tímto principem se můžeme setkat u pacientů, kterým byla při laryngektomii odstraněna jazylka a aspirační náplň není možná.

Proces učení jícnového hlasu by měl začít co možná nejdříve po operaci. Pokud je to možné, tak se s výukou začíná ještě za pobytu pacienta na ORL klinice nebo krátce po propuštění. V první fázi se pacient učí pouze slabiky sestávající z explosivy a souhlásky. Postupně se však přidávají slabičné shluky, které sice nedávají smysl, ale pomáhají v osvojení potřebné techniky. V případě úspěšného zvládnutí se přistupuje k nácviku frází a souvislé řeči. Potřebnou dobu k nácviku jícnového hlasu nelze přesně určit, protože je závislá na mnoha faktorech. V literatuře se uvádí, že je potřeba 30 až 50 hodin velmi intenzivního tréninku k osvojení jícnového hlasu.

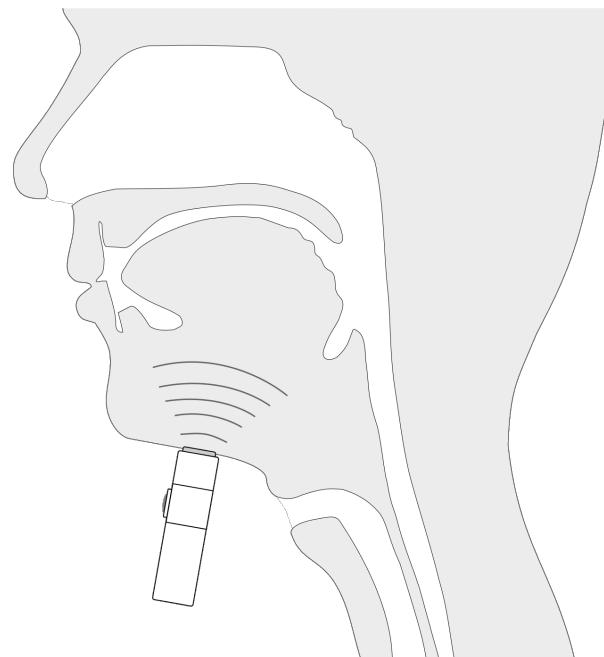
Míra úspěšnosti nácviku srozumitelného hlasu se uvádí v rozsahu od 14% do 75%. Takto obrovský rozsah značí o mnoha faktorech, které mohou ovlivnit úspěšné osvojení jícnového hlasu. Mezi možné příčiny neúspěchu patří fyziologické nebo anatomické problémy, psychologické problémy, nebo jednoduše neadekvátní podpora při řečové terapii [7]. Velkou roli také hraje snaha a odhodlání samotného pacienta.

Nepopíratelnou výhodou této techniky rehabilitace je nezávislost pacienta na lékaři po úspěšném osvojení jícnového hlasu a permanentní oddělení dýchacích a polylakacích cest bez rizika vniknutí potravy do dýchacích cest. Mezi nesporné výhody také patří volné ruce při vytváření řeči. Za nevýhody se obecně považuje srozumitelnost produkovaného hlasu. Je to způsobeno jednak „břišním“ zabarvením, které je už z podstaty metody přítomné, a dále také nízkou intenzitou a krátkou výdrží při tvorbě tónu. Za negativum se dá také považovat množství pacientem vynaloženého úsilí potřebného k osvojení techniky. Velmi často se také mluvčí ostýchají jícnový

hlas používat, protože mají pocit, že je společensky nevhodné dorozumívat se formou blízkou říhání. Z tohoto důvodu se odhaduje, že v běžném životě využívá jícnový hlas pouze 20 až 30% pacientů, kteří se začali tuto techniku učit [8].

Elektrolarynx

Rehabilitace hlasu pomocí elektrolarynxu se řadí mezi tzv. elektromechanické metody. Princip spočívá v přikládání zařízení, které obsahuje generátor zvuku nazývaný elektrolarynx. Přiložením do oblasti spodiny úst a aktivací zařízení se generovaný zvuk a vibrace přenášejí do dutiny ústní a dalších přilehlých artikulačních orgánů. Následnou artikulací je pacient schopen hovořit. Znázorněno na obr. 3.4.



Obrázek 3.4: Princip rehabilitace hlasu pomocí elektrolarynxu.

Takto generovaná řeč se vyznačuje několika charakteristickými rysy. V první řadě řeč budí velmi mechanický dojem. Důvodem je samozřejmě samotný elektrolarynx, jelikož se jedná o elektromechanický generátor zvuku s konstantním buzením, je také základní frekvence produkovaného hlasu více či méně konstantní. Řečník tak má velmi

omezené možnosti, jak řeč emotivně zabarvovat. V průběhu času se objevily snahy průběžně měnit frekvenci zařízení a tím ovlivňovat základní frekvenci produkované řeči [9, 10, 11]. Hlavním problém všech těchto zařízení je docílit změnu fundamentální frekvence na základě toho, co chce řečník říci. V současné době existují pouze experimentální zařízení, která umožňují ve velmi omezené míře změnu frekvence [12]. Další charakteristický rys představuje nižší srozumitelnost řeči, která se ještě snižuje s rostoucím okolním hlukem. Velmi často se stává, že posluchač, který se s takto produkovanou řečí setkává poprvé, není schopen plně porozumět. Se srozumitelností souvisí i další charakteristický rys, kterým je přítomnost zvukového podkresu produkovaného samotným přístrojem.

Za hlavní výhodu elektrolarynxu se považuje rychlosť osvojení schopnosti produkovat řeč. Zároveň je tato metoda vhodná pro téměř všechny pacienty postižené ztrátou hlasu způsobenou léčbou karcinomu hrtanu. Z tohoto důvodu se hojně užívá u pacientů, kteří si neosvojili jícnový hlas nebo u nich není možné využít ostatních chirurgických metod. Za nevýhody se obecně pokládá kvalita produkované řeči, tedy monotonní a mechanicky znějící hlas. Dále potom zaměstnání jedné ruky držením nebo spouštěním zařízení.

Samostatnou kapitolou může být psychologický dopad na pacienta. Stejně jako u jícnového hlasu se řeč produkovaná promocí elektrolarynxu jeví odlišně od řeči přirozené. Navíc se ještě přidává potřeba využití nějakého zařízení. Člověk proto v mnoha případech cítí ostých a bojí se na veřejnosti mluvit.

3.2.2 Chirurgicko-protetická metoda

Další možnost rehabilitace hlasu představuje tracheoezofageální (zkr. TE) protéza. První zmínka o vytvoření fistule³ mezi průdušnicí a jícnem pochází z roku 1932.

³fistule (česky píštěl) je abnormální otvor mezi dvěma dutými orgány, nebo mezi dutým orgánem a kůží.

V tomto roce doktor Guttman poprvé vytvořil tracheoezofageální shunt⁴ („umělá píšťel“). Hlavní myšlenka spočívá ve vytvoření cesty prostřednictvím píštěle, pomocí které u tracheostomovaného člověka může proudit vzduch z plic do úst. Za normálních okolností vzduch proudí skrze tracheostomii a do úst se tak nedostane. Zacepe-li si pacient stomu, může proud vzduchu proudit skrze píštěl do úst. Vzduch procházející přes fistuli naráží do stěn jícnu a je rozvibrován. Tyto vibrace jsou následně modulovány pomocí artikulačních ústrojí a tak vzniká řeč. Tento ojedinělý zákrok otevřel cestu k chirurgické hlasové rehabilitaci. Vzniklo několik operačních metod, které se navzájem lišili víceméně jen umístěním fistule [4].

Hlavní snahou chirurgů bylo vytvoření bezpečné, správně nasměrované píštěle umožňující tvorbu hlasu. Bohužel v mnoha případech byly tyto zákroky spojené s vážnými komplikacemi (infekce, zápaly či těžká krvácení). Důležitým problémem, se kterým se jednotlivý tvůrci museli vypořádat, byla stálost vytvořeného otvoru tak, aby jím neprotékaly tekutiny špatným směrem a nedocházelo k zatékání do dýchacích cest a orgánů. Jelikož se jednalo o velmi náročné techniky, a bylo s nimi spojeno velké množství rizik, došlo v 80.letech 20.století k opadnutí snah tyto metody aplikovat.

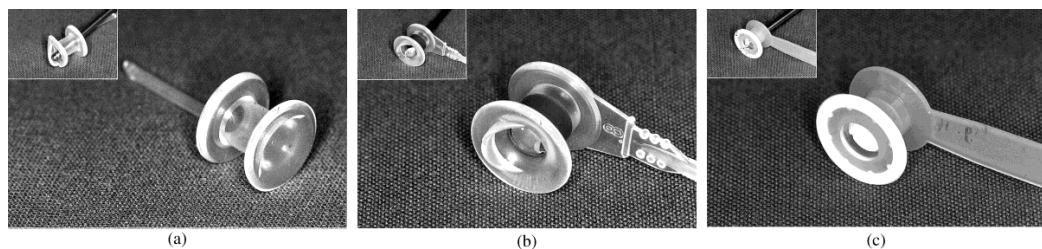
Svou renesanci zažily s vložením jednocestného ventilu, který umožňoval pouze jednosměrný průchod tekutin skrze píštěl, jak je ilustrováno na obr. 3.5. První komerčně dostupná protéza se objevila v 80.letech 20.století v USA. Na obr. 3.6 jsou zobrazeny příklady různých typů protéz. Na používané protézy jsou kladené přísné nároky a musí vyhovovat určitým požadavkům. Předně se musí vyrábět z biokompatibilního materiálu, který odolává biodegradaci. Tím je zaručena dlouhodobá trvanlivost a správná funkce. Potřebný tlak k otevření faryngoezofageálního segmentu by měl být co nejnižší, aby bylo možné vytvářet plynulou řeč. První vyráběné protézy měly tento tlak příliš vysoký a omezovaly tak množinu potencionálních pacientů. Nejmodernější

⁴shunt - kanál, kterým je tekutina odkloněna z přirozené dráhy. Tento kanál může být vytvořen chirurgicky nebo pomocí syntetické trubice.

protézy se již vyznačují velmi nízkým otevíracím fonačním tlakem. V neposlední řadě by měla být protéza samofixační a snadno vyměnitelná.



Obrázek 3.5: Průchod vzduchu tracheoezofageální protézou.



Obrázek 3.6: Ilustrace používaných TE protéz (a) Gronigenova nízkotlaká protéza, (b) Provox2 a (c) Blom-Singer protéza.

V praxi se používá několik druhů protéz. Hlavním rozdílem mezi nimi však je zda se pacient přímo účastní výměny ventilu, jehož fundamentální funkcí je vytvoření průchodu pro vzduch proudící z průdušnice do jícnu. U protéz, které jsou vyměňovány operačně, se doba používání pohybuje od 3 do 6 měsíců. Tento interval velmi významně ovlivňuje tvorba biofilmu na povrchu náhrady. K tvorbě dochází následkem přímého kontaktu protézy s tělními tekutinami a potravou. Rychlosť tvorby biofilmu

ovlivňuje tvar a materiál, ze kterého je náhrada vytvořena [13]. U typů, které si nositel může měnit sám, se předpokládá, že budou čištěny nebo měněny přibližně jednou za dva týdny.

Samotný zákrok zavedení protézy je možné provést zároveň s výkonem totální laryngektomie (tzv. primární zavedení hlasové protézy) nebo až po zotavení pacienta z náročné léčby nádorového onemocnění (tzv. sekundární zavedení). Primární zavedení umožňuje začít s hlasovou rehabilitací krátce po odstranění hrtanu. Zároveň pacient nemusí v krátké době podstupovat druhou operaci, při které by se vkládal jednocestný ventil do vytvořené fistule.

V praxi se ukázalo, že úspěšnost rehabilitace je více než 80% [14]. Důležitým faktorem, stejně jako u jícnového hlasu, je funkčnost faryngoezofageálního segmentu. Dále také otvírací tlak horního jícnového svěrače. Hlas tvořený protézou se vyznačuje vysokou kvalitou, dobrou srozumitelností, individuálním zabarvením a relativně dlouhou fonační dobou dosahující průměrně 20 sekund [15]. Oproti jícnovému hlasu není potřeba tak intenzivní edukace pacienta k plnému osvojení hlasu. V současnosti se jedná o nejpoužívanější metodu rehabilitace hlasu.

3.2.3 Hrtanu podobné struktury

S rozvojem mikrovaskulárních⁵ transplantátů se začaly objevovat postupy, které umožňovaly rehabilitovat hlas pouze pomocí chirurgického zákroku. Tyto techniky umožňují permanentní spojení hypofaryngu s tracheou pomocí vlastní tkáně pacienta.

První takovou metodu představil v roce 1984 doktor Ehrenberger [4], který popsal tzv. „řečový sifón“ (angl. **speech siphon**). Tento sifón je vytvořen z části tenkého střeva zvané lačník (jejunum). Spojení mezi hrtanem a hltanem je dvakrát esovitě zahnuto tak, aby bylo minimalizováno riziko sekundární aspirace. Schéma „řečového sifónu“ podle Ehrenberga je znázorněno na obr. 3.7 A. Již na první pohled

⁵mikrovaskulární - část oběhového systému složeného z nejmenších cév, jako jsou kapiláry, žilky aj.

je zřejmé, že se jedná o velmi náročný chirurgický zákrok. První články publikované autorským kolektivem prezentovaly velmi dobré funkční výsledky metody. Podle [5] bylo doposud operováno přibližně 60 pacientů.

V roce 1990 byla popsána laryngoplastika podle Hagena. V tomto případě se vytváří tzv. **neolarynx**, k jehož vytvoření se používá štěp z předloktí. Vnitřek neolaryngu je kryt kůží. Neoglottis je vyztužen chrupavkou a překrývá vchod do neolaryngu tak, aby nedocházelo k sekundární aspiraci. Laryngoplastika podle Hagena je znázorněna na obr. 3.7 B. Doposud bylo operováno přibližně 300 pacientů [5].



Obrázek 3.7: A) Schéma „řečového sifónu“ tak jak jej představil Ehrenberg. B) Laryngoplastika podle Hagena

Bohužel v současné době tyto metody nenacházejí širší uplatnění. Především je to způsobeno chirurgickou náročností samotných metod, kvůli které se velmi těžko prosazují na dalších pracovištích. Dalším aspektem, který limituje tyto metody, je vliv na samotného pacienta. Metody předpokládají další chirurgický zákrok vykonaný po totální laryngektomii. Tento zákrok představuje další zátěž pro pacienta nemluvě o možných komplikacích. I přes nedostatky těchto metod je pochopitelná snaha lékařů o intenzivní výzkum v této oblasti. Při úspěšné léčbě je pacient schopen produkovat

hlas velmi dobré kvality a ve většině případů nepotřebuje žádnou péči ze strany lékařů ORL.

3.2.4 Transplantace hrtanu

Nejkomplexnější možnost rehabilitace hlasu představuje transplantace hrtanu. V tomto případě pacient obdrží implantovaný hrtan od dárce. Pokud je transplantace úspěšná, přebírá transplantovaný orgán plně funkci původního orgánu a velmi významně zvyšuje šance pacienta na plné zotavení bez trvalých následků.

První informace spojené s výzkumem možností provedení transplantace hrtanu se objevují již v 60. letech 20. století⁶. Přesto byla první totální hrtanová transplantace provedena až profesorem Marshalllem Stromem v roce 1998 [16] a do dnešních dnů byly provedeny pouze 2 kompletní transplantace.

Prvním pacientem, který podstoupil transplantaci, byl čtyřicetiletý muž z USA. K laryngektomii v jeho případě vedla motocyklová nehoda, při které si pacient rozdrtil hrtan. K incidentu došlo 20 let před transplantací. Před zákrokem používal k produkci řeči elektrolarynx. Dárcem orgánu byl taktéž čtyřicetiletý muž, který zemřel na mozkové aneurysma. Úspěch transplantace se na příjemci projevil již třetí den po operaci, kdy poprvé po 20 letech promluvil (vyslovil anglické slovo „hello“). Přibližně po 36 měsících od transplantace byl produkovaný hlas srovnatelný s hlasem zdravého člověka. Podle vlastních slov pacienta se po operaci jeho kvalita života „nesmírně“ zlepšila. [17] Doposud poslední úspěšně vykonaná transplantace byla zaznamenána v říjnu 2010.

Mezi hlavní důvody takto malého počtu zákroků patří množství pacientů vhodných pro tuto proceduru. Jelikož se jedná o transplantaci dárcovského orgánu je nutné použít imunosupresiv, tedy medikamentů zabranující odmítnutí orgánu. Imunosupresiva jsou však v současné době nepoužitelná u lidí trpících rakovinou hrtanu z důvodu

⁶Vůbec první úspěšná transplantace orgánu (ledvin) se uskutečnila v roce 1954.

velmi vysokého rizika rozšíření rakoviny [16]. Další problém představuje náročnost samotného zákroku. Předně je potřeba provést reinervaci a obnovení krevního oběhu v implantovaném orgánu. U první provedené transplantace se nepodařilo dosáhnout kompletní reinervace. Výsledkem tak byl velmi kvalitní generovaný hlas, ale zároveň nebylo možné pomocí hrtanu zabezpečit bezproblémové dýchání a bylo proto nutné ponechat tracheostomii.

Poslední výzkum v oblasti imunosuprese však naznačuje, že by v dohledné době mohlo dojít k pokroku a umožnit transplantaci hrtanu i u lidí trpících rozsáhlou rakovinou v oblasti krku [16]. Prozatím je však tato metoda vhodná pro pacienty netrpící rakovinou, případně ty, u kterých převažovaly benigní nádory a již 5 let nedošlo k recidivě.

3.2.5 Shrnutí

Rehabilitaci pacientů, kteří prodělali chirurgické odstranění hrtanu, je ve vyspělých zemích věnována značná pozornost, jelikož následky této operace, oproti jiným druhům léčby, velmi významně ovlivňují kvalitu života pacientů. V první řadě se léčený musí vyrovnat se ztrátou hlasu. Tato situace je již sama o sobě velmi náročnou psychickou zkouškou. Ztráta hlasu je však pouze jedním z vícero problémů, se kterými je potřeba se vypořádat. Mezi další patří možná ztráta čichu či vyšší náchylnost k respiračním onemocněním. Neméně významnou roli sehrává i fyzická odlišnost a z toho pramenící psychická zátěž pacienta po absolvované léčbě.

V současnosti nejpoužívanějšími metodami rehabilitace hlasu jsou **tracheoeozafageální píštěl** (popsáno v části 3.2.2), **jícnový hlas** (3.2.1) a použití **elektrolarynxu** (3.2.1). Existují samozřejmě i další a přehled v současnosti používaných je uveden v tab. 3.1.

Většina pacientů je tedy rehabilitována pomocí tracheoeozafageálního píštěle, který principiálně vychází z jícnového hlasu, jehož negativa se snaží eliminovat. O úspěchu

	Kvalita	Výhody	Nevýhody
Tracheoezofageální píštěl	Vysoká	Vysoká míra osvojení, dlouhá fonační doba	Zanášení píštěle a s ním spojené čištění, případně dodatečná lékařská péče
Jícnový hlas	Dobrá	Volné ruce při mluvení, není potřeba dodatečné lékařské péče	Velmi náročná metoda k naučení, nepřirozený hlas
Elektrolarynx	Nízká	Snadné k naučení	Monotonní až robotický hlas, nutné nosit externí elektrické zařízení
Hrtanu podobné struktury	Vysoká	Nezávislost pacienta na pravidelné lékařské péči	Velmi náročná chirurgická procedura, která pacienta vystavuje dalším možným rizikům
Transplantace hrtanu	Velmi vysoká	Transplantovaný hrtan přejímá funkci odstraněného orgánu	Velmi náročná chirurgická procedura, která je vhodná jen pro malé procento pacientů

Tabulka 3.1: Přehled dostupných metod rehabilitace hlasu

rehabilitace, stejně jako u jícnového hlasu, tak především rozhodují vlastnosti faryngoezofageálního segmentu. Pokud pacient není schopen si osvojit jícnový hlas, případně nemá voperován píštěl, je použit elektrolarynx. Bohužel tyto metody neřeší další problémy spojené s odstraněním hrtanu, a proto se lékaři stále snaží zdokonalovat rehabilitační metody. Za nejkomplexnější se dá považovat úplná transplantace hrtanu, která řeší víceméně všechny problémy spojené s odstraněním hrtanu. Bohužel tento zákrok je velmi náročný a vhodný pouze pro malou část pacientů. I když je tedy

v současné době lékařská věda schopna rehabilitovat hlas, tak zde zůstává otevřený prostor pro inovace, a tím zlepšení kvality života lidí postižených ztrátou hrtanu.

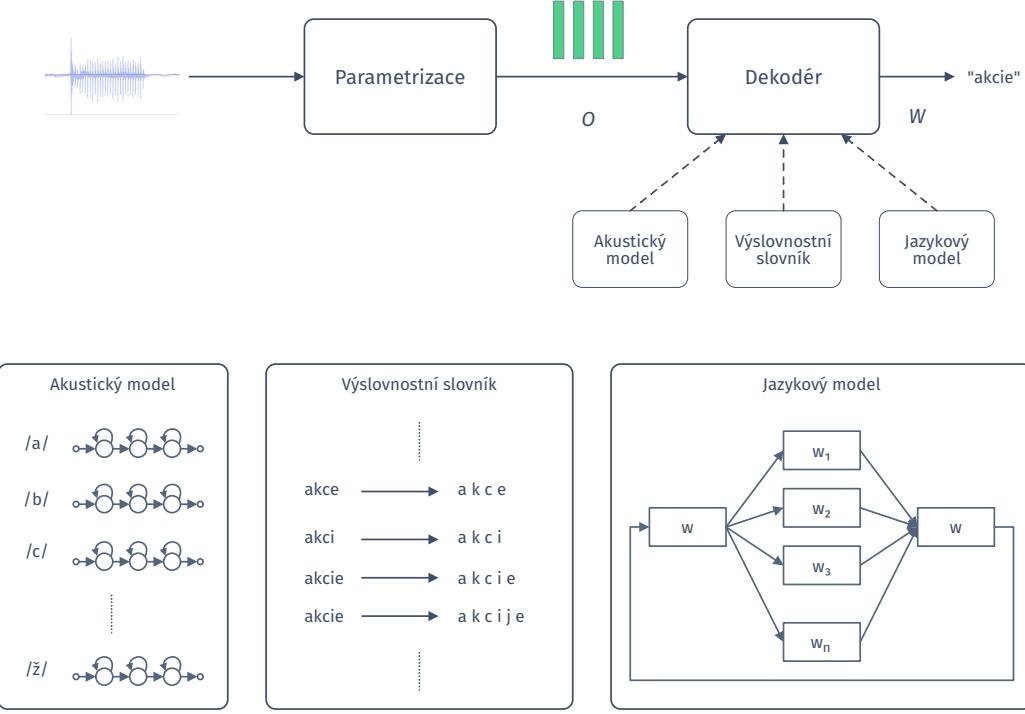
Kapitola 4

Automatické rozpoznávání řeči

Úlohou systému automatického rozpoznávání řeči (ASR) je převedení mluvené řeči na posloupnost slov, které řečník vyslovil. První takovéto systémy se začaly objevovat v první polovině 20. století. Jejich funkce spočívala v analýze akustického signálu a jeho porovnávání se vzorem. Byly tak schopny rozpoznávat jen velmi omezené množství slov. Významný zlom nastal v polovině 80. let minulého století, kdy se začaly používat systémy založené na statistickém přístupu, konkrétně skryté Markovovy modely (HMM) [18]. Princip je u takového systému znázorněn na obr. 4.1. Řečový signál je analyzován a pomocí parametrizace převeden na sekvenci vektorů pozorování $\mathbf{O} = \{\mathbf{o}_1 \ \mathbf{o}_2 \ \dots \ \mathbf{o}_T\}$. Tyto vektory jsou u většiny systémů získávány s periodou 10 ms pro segmenty řeči mající nejčastěji délku 20 až 40 ms. Vlastní rozpoznávání pak probíhá v dekodéru, kde se systém snaží vybrat k vektorům pozorování \mathbf{O} takovou posloupnost slov $W = \{w_1 \ w_2 \ \dots \ w_N\}$, která maximalizuje aposteriorní pravděpodobnost (MAP)

$$\hat{W} = \underset{W}{\operatorname{argmax}} \ P(W|\mathbf{O}). \quad (4.1)$$

Pomocí Bayesova pravidla je možné podmíněnou pravděpodobnost $P(W|\mathbf{O})$ přepsat do tvaru



Obrázek 4.1: Schéma automatického systému rozpoznávání řeči pracující na statistickém přístupu

$$P(W|\mathbf{O}) = \frac{P(\mathbf{O}|W)P(W)}{P(\mathbf{O})}. \quad (4.2)$$

Jelikož pravděpodobnost $P(\mathbf{O})$ je z principu nezávislá na sekvenci slov W , je možné rovnici (4.1) upravit do tvaru

$$\hat{W} = \operatorname{argmax}_W P(\mathbf{O}|W) P(W). \quad (4.3)$$

Takto upravená rovnice představuje obecné pravidlo dekódování a její členy pak základní stavební prvky ASR systému. Podmíněná pravděpodobnost $P(\mathbf{O}|W)$ odhaduje sekvenci pozorování \mathbf{O} na předpokladu slov W , tento výpočet je realizován **akustický modelem** (viz obr. 4.1). K určení \hat{W} je ještě nezbytné mít pravděpodobnost výskytu požadované posloupnosti slov $P(W)$, o určení této pravděpodobnosti se stará **jazykový model**. Pro doplnění je nutné dodat, že **slovník** obsahuje seznam všech slov, se kterými je systém schopen pracovat. Tento seznam je včetně fonetické

transkripce. Všechny tyto části jsou uvnitř **dekodéru**, který realizuje prohledávací strategii. V následujícím textu jednotlivé stavební prvky ASR systému blíže popíšeme.

4.1 Parametrizace řečového signálu

Jako v mnoha jiných odvětvích, tak i při rozpoznávání řeči je v mnoha případech inspirací člověk. Pro získání sekvence pozorování (příznaků) vycházíme z **modelování produkce řeči** a **modelování procesu slyšení**, které se inspirují právě člověkem.

4.1.1 Modelování produkce řeči

Cílem modelování produkce řeči je nalezení matematických vztahů, které poslouží k reprezentaci fyzikálních dějů spojených s produkcí řeči. Základem je parametizační technika **lineárního prediktivního kódování**, známá pod anglickou zkratkou LPC¹ [19]. Vychází z představy, že hlasové ústrojí člověka je schopno vytvářet tři různé typy řečových zvuků:

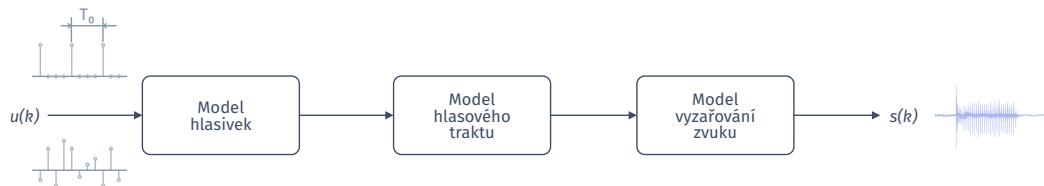
- *samohlásky* - ty se řadí mezi znělé typy zvuků produkované periodickým buzením vznikajícím pulsy vzduchu, které jsou produkovány hlasivkami;
- *frikativy* (např. /f/²) - někdy nazývané jako třené souhlásky, protože vznikají třením vydechovaného proudu vzduchu o překážku v některém místě hlasového ústrojí. Těmito překážkami může být jazyk, zuby ap.;
- *explozivy* (např. /b/, /p/ ap.) - také nazývané jako souhlásky výbuchové, se tvoří úplným uzavřením vydechovaného proudu vzduchu pomocí artikulačních orgánů. To se následně projeví jako krátká pauza (tzv. okluze), po které násle-

¹Linear Predictive Coding

²Zápis /f/ symbolizuje foném, což je akustická reprezentace písmene, *f*. Konkrétní zápisy se mohou lišit podle použité fonetické abecedy. V Čechách se nejčastěji používá abeceda SAMPA či ZČFA.

duje náhlé jednorázové uvolnění a únik nahromaděného vzduchu, tzv. exploze [20];

Snahou modelování je navržení modelu hlasového traktu, který bude dobře popisovat výše popsané zvuky. Nesmí se však zapomenout na možnou přílišnou složitost a nedostatečnou přesnost modelu. Jako ideální se může jevit lineárně časově invariantní model. Bohužel lidská řeč je zástupcem kontinuálního časově variantní a některých situacích dokonce nelineárního proces, takže je téměř nemožné jej přesně namodelovat. Pokud se však, za určitého předpokladu, provedou určitá zjednodušení, tak je možné navrhnout lineární časově invariantní model řeči, který je platný pro krátké časové úseky. Jinými slovy, tímto předpokladem je, že v konkrétním krátkém časovém úseku zůstává buzení a parametry hlasivkového traktu přibližně konstantní. Tuto podmínka lze považovat za platnou pro intervaly od 10 do 30 ms. Odtud také vychází uvažovaná perioda segmentů řeči, viz úvod této kapitoly. Pro tyto segmenty je pak možné proces vytváření řeči modelovat pomocí tzv. **krátkodobého modelu**, který má v krátkých časových intervalech pevné parametry. [18]

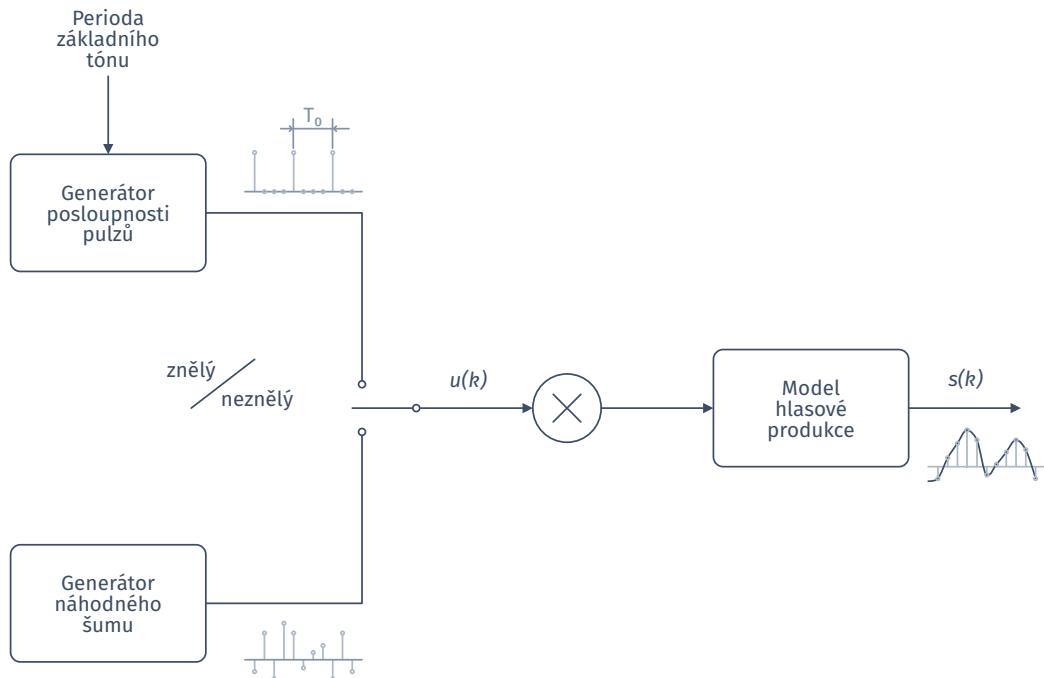


Obrázek 4.2: Blokové schéma modelu produkce řeči

Pro odvození obecného diskrétního modelu hlasivkového traktu se vychází ze zjednodušeného modelu produkce řeči (obr. 4.2). Ten je tvořen modelem hlasivek, modelem hlasivkového traktu a modelem vyzařovaného zvuku, které jsou sériově řazeny. K odvození a popisu vlastností modelu se využívá výhod z-transformace [20]. Po zjednodušení je krátkodobý model produkce řeči approximovat celopólovým modelem (filtrem) $H(z)$ ve tvaru

$$H(z) = \frac{G}{1 + \sum_{i=1}^Q a_i z^{-i}} = \frac{S(z)}{U(z)}, \quad (4.4)$$

kde G představuje celkové zesílení, Q je řád modelu odpovídající $2K + 1$ počtu formantů, které má model postihovat, a_i jsou parametry modelu. Vstupem modelu je buzení $u(k)$ (viz obr. 4.2), tedy pro znělé zvuky sled pulsů s periodou T_0 ³ a pro neznělé zvuky náhodný šum s plochým spektrem. V časové oblasti je pak diskrétní výstupní odezva při fixovaných parametrech hlasového traktu ($10 - 30 ms$) dána konvolucí buzení a impulzní odezvy krátkodobého modelu. Na základě toho je možné model upravit na tvar podle obr. 4.3, kde je $u(k)$ buzení a $s(k)$ je výstupní signál s parametry hlasového ústrojí odpovídající a_i celopólového modelu.



Obrázek 4.3: Blokové schéma upraveného modelu produkce řeči

K odhadu parametrů a_i slouží **lineární prediktivní analýza**. Odhad probíhá přímo z krátkodobého řečového signálu. Přenosové vlastnosti krátkodobého modelu je možné popsat rovnicí (4.4). Myšlenka metody LPC staví na předpokladu, že vzorek

³Perioda základního hlasivkového tónu.

k řečového signálu je možné popsat lineární kombinací Q předchozích vzorků a buzení $u(k)$, což lze zapsat úpravou vztahu (4.4) do tvaru

$$s(k) = - \sum_{i=1}^Q a_i s(k-1) + Gu(k). \quad (4.5)$$

Z něj je patrné, že se LPC snaží parametry modelu a_i a zesílení G odhadnout pomocí známé reálně naměřené posloupnosti $s(k)$. K vyřešení se používá principu minimizace kvadratické chyby krátkodobé energie signálu. Ta je v časové oblasti popsána vztahem

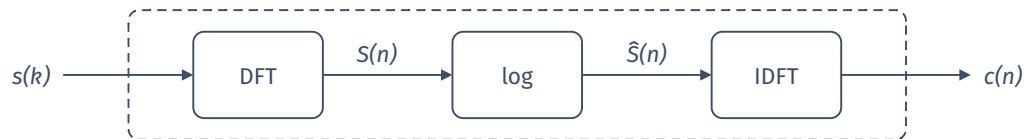
$$E = \sum_k e^2(k) = \sum_k [s(k) - s'(k)]^2 = \sum_k \left(s(k) + \sum_{i=1}^Q a_i s(k-1) + Gu(k) \right)^2, \quad (4.6)$$

kde $s(k)$ jsou vzorky reálného řečového signálu a $s'(k)$ jsou ty predikované LPC filtrem. K získání řešení krátkodobé chyby predikce E , pro konkrétní analyzovaný segment, je použita metoda nejmenších čtverců. K výpočtu konkrétních koeficientů modelu a_i je možné použít rekurzivního Durbinova algoritmu. [18]

Další možností jak popsat hlasový trakt je pomocí **kepstrálních koeficientů lineární predikce**. Kepstrum je definováno jako inverzní diskrétní Fourierova transformace (IDFT) logaritmu velikosti transformovaného vstupního signálu pomocí diskrétní Fourierovy transformace (DFT), matematicky popsáno vztahem

$$c(k) = \mathcal{F}^{-1} \{ \log |\mathcal{F} \{ s(k) \}| \}, \quad (4.7)$$

a graficky znázorněno na obr. 4.4.



Obrázek 4.4: Blokové schéma principu výpočtu kepstra

Pro získání kepstrálních koeficientů lineární predikce logaritmujeme rovnici (4.4) čímž vznikne vztah

$$\log H(z) = \log \left(\frac{G}{A(z)} \right). \quad (4.8)$$

Člen $A(z)$ je polynomem proměnné z^{-1} řádu Q , a pokud všechny jeho kořeny leží uvnitř jednotkové kružnice, tak lze aplikovat Taylorův rozvoj na vztah (4.8) ve tvaru

$$\log \left(\frac{G}{A(z)} \right) = c(0) + c(1)z^{-1} + \dots = \sum_{k=0}^{\infty} c(k)z^{-k}, \quad (4.9)$$

kde $c(k)$ jsou tzv. kepstrální koeficienty LPC. K odstranění logaritmu je potřeba obě strany rovnice derivovat. Po úpravě je výsledný vztah

$$-\sum_{i=1}^Q ia_i z^{-i} = \left(\sum_{k=0}^{\infty} kc(k)z^{-k} \right) \left(\sum_{i=0}^Q a_i z^{-i} \right). \quad (4.10)$$

Pokud se $a_i = 1$, pak je možné roznásobení rovnice (4.10) a porovnáním členů u stejných mocnin z zapsat vztahy pro výpočet kepstrálních koeficientů LPC

$$c(1) = -a_1,$$

$$c(k) = \begin{cases} -a_k - \sum_{i=1}^{k-1} \binom{i}{k} c(i)a_{k-i}, & \text{pro } 2 \leq k \leq Q, \\ -\sum_{i=1}^Q \binom{k-i}{k} c(k-i)a_i, & \text{pro } k = Q+1, Q+2, \dots, \end{cases} \quad (4.11)$$

kde $k = 1, 2, \dots, Q^*$ a Q^* je počet kepstrálních koeficientů a $Q^* \geq Q$.

Kepstrální koeficienty LPC jsou vztaženy ke spektrální obálce mikrosegmentu řeči odvozené LPC analýzou, tu je možné získat dosazením $e^{j\omega}$ za z v rovnici (4.4). Pro uspokojivou reprezentaci se tradičně volí $Q = 7$ až 15 v závislosti na spektrální šířce přenášeného pásma, požadované přesnosti approximace apod. Z toho plyne, že pro popis mikrosegmentu řeči by mohl stačit příznakový vektor o 15 koeficientech.

4.1.2 Modelování procesu slyšení

Zvuk představuje mechanické vlnění hmotných částic, které se šíří v plyném, kapáném nebo tuhém prostředí. Z fyziologického pohledu je však zvukem považováno pouze slyšitelné vlnění. To je takové, které je schopno vnímat sluchové ústrojí člověka. Zpravidla se jedná o frekvence $16 \text{ Hz} - 20 \text{ kHz}$. Pro každého člověka je, ale toto rozmezí individuální a mění se s věkem. S přibývajícím věkem a sluchovou zátěží klesá hlavně horní mezní kmitočet. [20]

To, zda je člověk schopen daný zvuk slyšet však není závislé pouze na frekvenci zvuku. Velmi podstatná je i intenzita zvuku, která se rovná energii zvukového vlnění, která projde za jednotku času jednotkovou plochou kolmou ke směru šíření vln. Zároveň je úměrná akustickému tlaku zvukové vlny, tj. tlaku, kterým zvukové vlny působí na nějakou překážku. V případě člověka tedy ušní bubínek. Závislost mezi intenzitou zvuku $I \text{ [W m}^{-2}\text{]}$ a akustickým tlakem $p \text{ [Pa]}$ je vyjádřen vztahem

$$I = \frac{p^2}{z}, \quad (4.12)$$

kde z je měrná akustická impedance prostředí, kterým se zvuk šíří. Lidské ucho je schopno vnímat akustický tlak v rozsahu od $2 \cdot 10^{-5}$ až $2 \cdot 10^2 \text{ Pa}$, tj. v rozsahu sedmi řádů. Z praktického důvodu se tedy používá logaritmické stupnice. K vyjadřování pak slouží logaritmus poměru uvažované veličiny a mezinárodně normované referenční hodnoty téže veličiny. [20] Hladina intenzity je pak definována vztahem

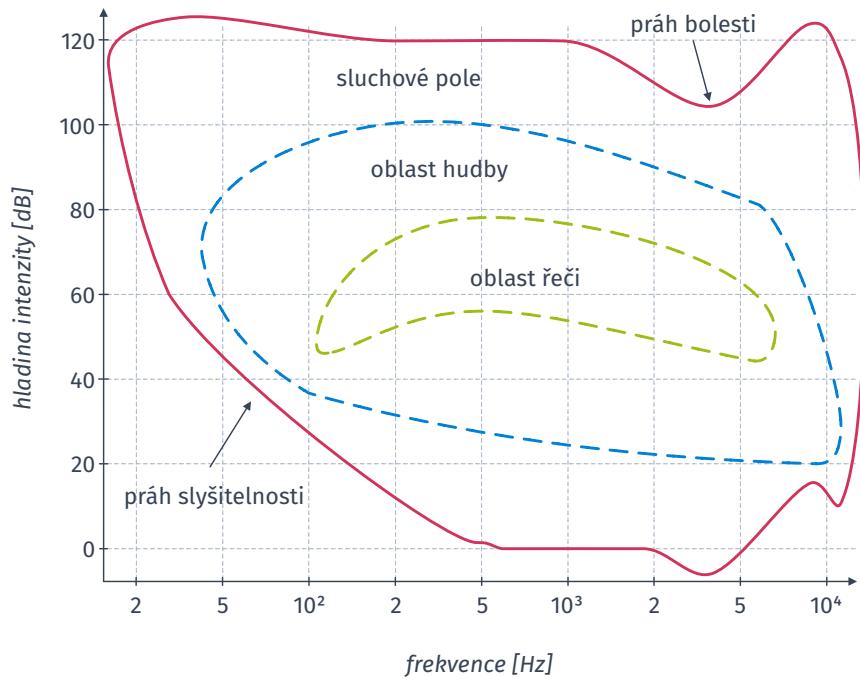
$$L_I = 10 \log_{10} \frac{I}{I_0}, \quad (4.13)$$

kde I představuje intenzitu zvuku a $I_0 = 10^{-12} \text{ W m}^{-2}$ referenční hodnota intenzity. Pro hladinu akustického tlaku platí

$$L_p = 20 \log_{10} \frac{p}{p_0}, \quad (4.14)$$

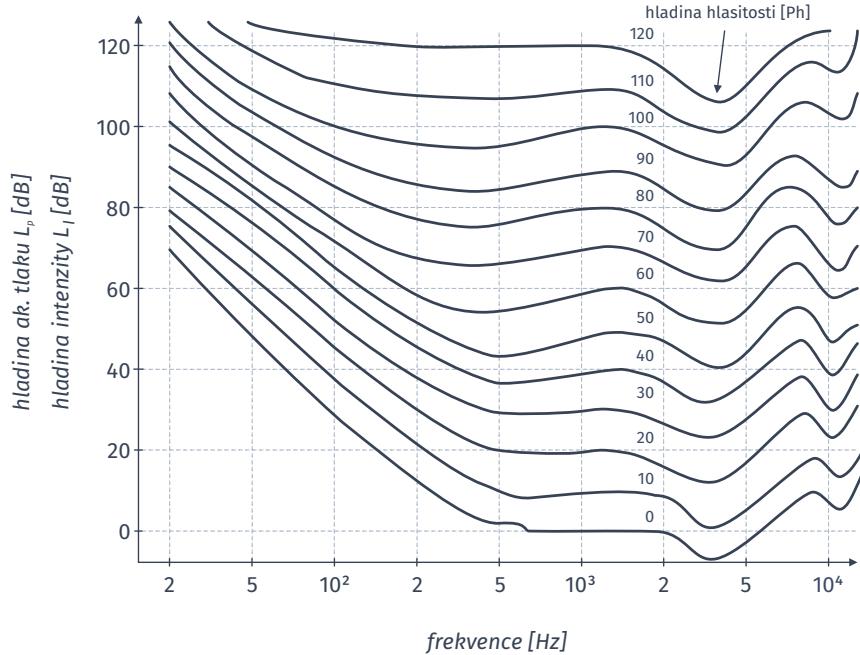
kde p je akustický tlak a $p_0 = 2 \cdot 10^{-5} \text{ Pa}$ je referenční hodnota akustického tlaku. Jak L_I , tak L_p je udáváno v decibelech [dB].

Důležitým pojmem je pak **práh slyšitelnosti**, který představuje minimální intenzitu zvuku potřebnou k tomu, aby jej šlověk mohl slyšet, viz obr. 4.5. Tento práh je zcela subjektivní a je závislý na frekvenci. Obecně je lidský sluch nejcitlivější na frekvence $3 - 4 \text{ kHz}$. Směrem k nižším a vyšším kmitočtům citlivost sluchu klesá. **Práh bolesti** představuje horní mez intenzity sluchového pole (viz obr. 4.5), při níž již posluchač pocítuje bolest. Překročení této meze může vést k poškození sluchu. [18]



Obrázek 4.5: Oblasti vnímání akustického signálu lidským sluchem.

Hlasitost zvuku je závislost intenzity na frekvenci a je zcela subjektivní pocit, kterým člověk posuzuje intenzitu daného zvuku. Na obr. 4.6 jsou vyznačeny hladiny hlasitosti, které vznikly spojením bodů ve sluchovém poli (obr. 4.5) odpovídající tónům, které člověk vnímá stejně hlasitě. Z křivek je patrné, že subjektivní hlasitost se mění s frekvencí zvuku. Zvuky s nižší frekvencí vnímáme méně hlasitěji než zvuky s vyšší frekvencí, zejména pak zvuky v rozmezí $3 - 4 \text{ kHz}$. [20]



Obrázek 4.6: Oblasti vnímání akustického signálu lidským sluchem.

Principem medelování procesu slyšení je pak právě kompenzace nelineárního vnímání frekvencí lidským sluchem (viz obr. 4.6). Dále pak i respektování maskování zvuků včetně tzv. kritických pásem slyšení, což je přirozená vlastnost lidského sluchu. Maskováním se rozumí jev, kdy vnímání jednoho zvuku je ovlivněno přítomností jiného zvuku. Jinými slovy lze říci, že přítomnost jednoho zvuku zvyšuje prah slyšitelnosti pro jiný zvuk. Ten buď zní současně nebo s drobným časovým odstupem od toho prvního. Tento jev je jakýsi „psychologický filtr“, který ignoruje veškerý šum ležící mimo určité kritické pásmo. Šířka takového kritického pásma je přitom závislá na frekvenci poslouchaného tónu.

Typickým příkladem metod modelující proces slyšení jsou **melovská kepstrální filtrace** a **perceptivní lineární prediktivní analýza**.

Melovské kepstrální koeficienty

Metoda melovských frekvenčních kepstrálních koeficientů (MFCC) se snaží respektovat výše zmíněné vlastnosti lidského sluchu. Zejména se snaží dodržet kritická pásma slyšení a vliv subjektivního vnímání výšky tónů.

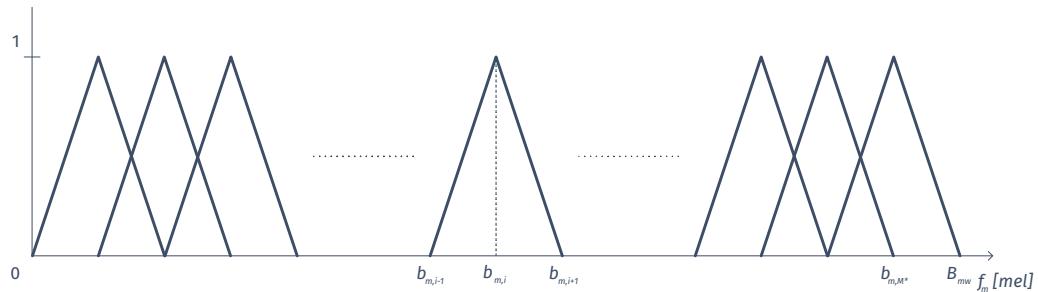
Základem MFCC je využití banky filtrů a lineárním rozložením frekvencí v tzv. **melovské frekvenční škále** definované vztahem

$$f_m = 2595 \log \left(1 + \frac{f}{700} \right), \quad (4.15)$$

kde $f [Hz]$ je frekvence v lineární škále a $f_m [mel]$ je odpovídající frekvence v melovské stupnice. Melovský filtr má trojúhelníkový tvar. Banka obsahuje filtry rozmístěné lineárně v melovských frekvenčních souřadnicích, a to tak, že dva sousední filtry se navzájem o polovinu překrývají. Pro střední frekvence jednotlivých filtrů $b_{m,i}$ platí v melovské škále vztah

$$b_{m,i} = b_{m,i-1} + \Delta_m, \quad (4.16)$$

kde $b_{m,0} = 0 mel$, $i = 1, 2, \dots, M^*$, a $\Delta_m = B_{m,w}/(M^* + 1)$. Ukázka banky filtrů je na obr. 4.7. Pro výpočet odezvy filtrů je však nezbytné přepočítat všechny koeficienty FFT do melovské frekvenční škály. Vhodnější je vyjádření trojúhelníkových filtrů ve frekvenční škále s měřítkem v herzích.

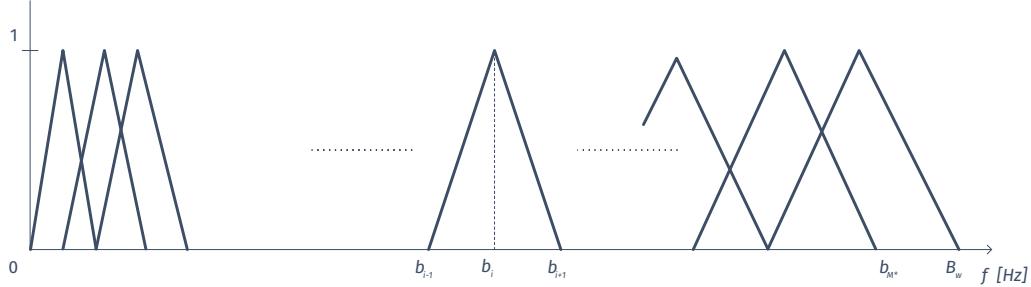


Obrázek 4.7: Rozložení banky trojúhelníkových filtrů v melovské frekvenční škále

K přepočtu středních frekvencí $b_{m,i}$ se využívá inverzního vztahu k (4.15) tedy

$$f = 700 \left[\exp(0,887 \cdot 10^{-3} f_m) - 1 \right]. \quad (4.17)$$

Střední frekvence b_i , $i = 1, 2, \dots, M^* + 1$ jsou vyjádřené v jednotce [Hz]. Filtry jsou rozmístěny nelineárně viz obr. 4.8.



Obrázek 4.8: Rozložení banky trojúhelníkových filtrů ve frekvenční škále

Při výpočtu melovských kepstrálních koeficientů jsou na vstup systému přivedeny mikrosegmenty (10 až 30 ms) řečového signálu $s(k)$. Na vzorcích $s(k)$ byla ještě předtím provedena preemfáze⁴. Pro jednotlivé mikrosegmenty je pomocí FFT vypočteno amplitudové spektrum $|S(f)|$ a následuje klíčová část celého procesu, melovský filtrace. Odezvy filtrů ve frekvenční oblasti lze vyjádřit vztahem

$$y_m(i) = \sum_{f=b_{i-1}}^{b_{i+1}} |S(f)| u(f, i), \quad i = 1, 2, \dots, M^*, \quad (4.18)$$

kde frekvence f jsou vybírány ze souboru frekvencí využívaných při FFT výpočtu a $u(f, i)$ je vyjádření konkrétního trojúhelníkového filtru i . Průchod filtrem tedy znamená, že každý koeficient FFT je násoben odpovídajícím ziskem filtru a výsledky jsou pro příslušné filtry akumulovány. Logaritmováním akumulovaných koeficientů $y_m(i)$ se provede převod do kepstrální oblasti. Tento krok příznivě omezí dynamiku signálu [19].

⁴Preemfáze znamená zdůraznění amplitud spektrálních složek řečového signálu s jejich vzrůstající frekvencí. [20]

Posledním krokem při výpočtu melovských kepstrálních koeficientů $\{c_m(j)\}_{j=1}^M$ je provedení IDFT (viz (4.7)). V případě MFCC se, ale používá diskrétní kosinova transformace (DCT), protože spektrum je reálné a symetrické. K výpočtu slouží vztah

$$c_m(j) = \sum_{i=1}^{M^*} \log y_m(i) \cos \left(\frac{\pi j}{M^*} (i - 0, 5) \right), \quad \text{pro } j = 0, 1, \dots, M, \quad (4.19)$$

kde M^* je počet pásem melovkového pásmového filtru a M je počet melovských kepstrálních koeficientů. Počet těchto koeficientů M se volí podstatně menší, než je počet pásem melovkového pásmového filtru M^* , obvykle se uvažuje prvních $M = 10$ až 13 koeficientů. Velmi často se také používá 1. a 2. těchto koeficientů, protože svým způsobem zohledňují dynamickou složku řeči.

Perceptivní lineární prediktivní analýza

Stejně jako MFCC, tak také i **perceptivní lineární prediktivní analýza (PLP)** vychází z lidského vnímání a slyšení zvuků. Snaha je postihnout z psychofyziky slyšení zejména kritická pásma spektrální citlivosti, vztah mezi intenzitou a vnímáním hlasitosti a také křivky stejné hlasitosti. [20] PLP (podobně jako LPC) pak approximuje získané sluchové spektrum koeficienty autoregresního celopólového modelu.

Prvním krokem PLP analýzy je **výpočet výkonového spektra řečového signálu**. Pro konkrétní mikrosegment řečového signálu $s(k)$ aplikujeme⁵ DFT. Krátkodonné spektrum je pak definováno vztahem

$$P(\omega) = |S(\omega)|^2 = [Re S(\omega)]^2 + [Im S(\omega)]^2. \quad (4.20)$$

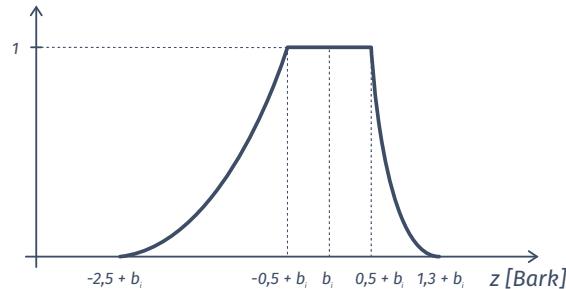
⁵Ještě před výpočtem je, stejně jako u MFCC aplikována preemfáze.

Poté následuje kompenzace nelineárního vnímání změn ve výšce zvuku. Vnímání je logaritmické, proto je nutné provést nelineární transformaci frekvenční osy pomocí vzorce

$$\Omega(\omega) = 6 \ln \left(\frac{\omega}{1200\pi} + \sqrt{\left(\frac{\omega}{1200\pi} \right)^2 + 1} \right), \quad (4.21)$$

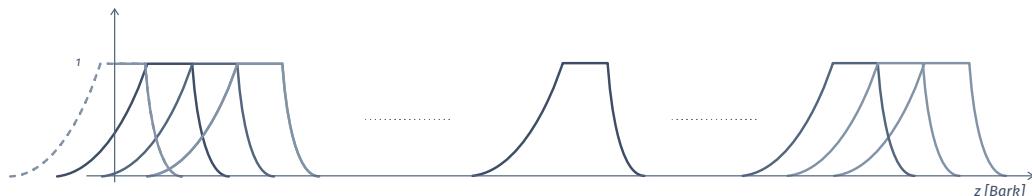
kde $\omega = 2\pi f$ [rad/s] a $\Omega(\omega)$ [Bark].

Zahrnutí kritických pásem slyšení (tzv. maskování zvuku) je realizováno navržením vhodného filtru typu pásmová propust šířky jednoho kritického pásma. Stejně jako v případě MFCC se jedná o banku filtrů, kde na sebe jednotlivé filtry ve frekvenční oblasti navazují. Na Barkově frekvenční ose (viz (4.21)) mají všechny filtry šířku 1 a jsou lineárně rozmístěny. Na obr. 4.9 je zobrazen průběh jednoho takového filtru. Filtr má strmost $+20 \text{ dB/Bark}$ směrem k nižším frekvencím a -50 dB/Bark směrem k vyšším frekvencím.



Obrázek 4.9: Ukázka filtru umístěného na Barkově frekvenční ose

Rozmístění filtrů na Barkově frekvenční ose je pak znázorněno na obr. 4.10.



Obrázek 4.10: Rozmístění filtrů na Barkově frekvenční ose

Jelikož člověk vnímá intenzitu zvuku v závislosti na frekvenci, tak je potřeba aplikovat **přizpůsobení křivkám stejné hlasitosti**. Na začátku je důležité definovat referenční hlasitost, tj. hlasitost, na kterou budeme normalizovat. Obvykle se volí 40 Ph [20], což přibližně odpovídá hlasitosti běžné řeči. K normalizaci je použit inverzní filter popsaný vztahem

$$E(\omega) = K \frac{\omega^4 (\omega^2 + 56,9 \cdot 10^6)}{(\omega^2 + 6,3 \cdot 10^6)^2 (\omega^2 + 379,4 \cdot 10^6) (\omega^6 + 9,6 \cdot 10^{26})}, \quad (4.22)$$

kde $\omega = 2\pi f$ a K je konstanta nastavená podle požadovaného zesílení. Přizpůsobení křivce stejné hlasitosti je pak možné například přenásobením celého výkonového spektra mikrosegmentů podle vztahu

$$P'(\omega) = E(\omega) P(\omega), \quad (4.23)$$

kde $P'(\omega)$ je spektrum transformované na stejnou hlasitost. Případně lze upravit tvar jednotlivých filtrů pomocí vztahu

$$\Phi(\omega, i) = E(\omega) \Psi(\omega - \omega_i, i), \quad (4.24)$$

kde $\Phi(\omega, i)$ je nový tvar filtru i v závislosti na frekvenci ω , $\Psi(\omega - \omega_i, i)$ je odezva filtru i se středovou frekvencí ω_i .

Po přizpůsobení následuje **výpočet energie jednotlivých filtrů**, to je obdobné jako u MFCC. Výpočet se provádí pro jednotlivé filtry a výsledky se pak sčítají. Matematicky to je zapsáno vztahem

$$\zeta_m = \sum_{\Omega=\Omega_m-2,5}^{\Omega_m+1,3} P(\Omega) \Phi(\Omega, m), \quad m = 1, 2, \dots, M-2, \quad (4.25)$$

kde M je počet použitých filtrů (kritických pásem).

Dalším krokem výpočtu je uplatnění „**zákona slyšení**“. Ten popisuje závislost mezi intenzitou a vnímanou hlasitostí. Na energie ζ_m je aplikována nelineární transformace vyjádřena vztahem

$$\xi_m = (\zeta_m)^{0,3}, \quad m = 1, 2, \dots, M-2, \quad (4.26)$$

kde M je opět počet filtrů. Díky této operaci dojde také k redukci proměnlivosti „výstupů“ kritických pásmových filtrů a výsledný hledaný celopólový model může být relativně nízkého řádu.

Finálním krokem je **aproximace celopólového modelu**. Ta vychází z výpočtu koeficientů celopólového modelu metody LPC, kde je model popsán vztahem (4.5). Pro chybu predikce pak platí

$$e(k) = \sum_k \left(s(k) + \sum_{i=1}^Q a_i s(k-i) \right). \quad (4.27)$$

Aplikací z-transformace a uvážením rovnice (4.4), je možné (4.27) upravit do tvaru

$$E(z) = \left[1 + \sum_{i=1}^Q a_i z^{-i} \right] S(z) = A(z) S(z), \quad (4.28)$$

kde $A(z)$ je inverzní filtr a $E(z)$ a $S(z)$ jsou z-transformace $e(k)$ a $s(k)$. Celkovou chybu predikce je pak možné vyjádřit vztahem

$$E(z) = \frac{1}{2\pi} \int_{-\pi}^{\pi} P(\omega) A(e^{j\omega}) A(e^{-j\omega}) d\omega, \quad (4.29)$$

kde $P(\omega)$ je vypočtené výkonové spektrum. Podobně jako u LPC je řešením nalezení minima celkové chyby autokorelační funkce $R(i)$. Pro konečný počet známých frekvencí je tato funkce definována vztahem

$$R(i) = \frac{1}{N} \sum_{n=0}^{N-1} P(\omega_n) \cos(i\omega_n), \quad (4.30)$$

kde $i = 0, \dots, Q$ a Q je řád autoregresního modelu, a N je počet bodů spektrální charakteristiky. Frekvence ω_n jsou ty, pro které jsou známé spektrální hodnoty. Pro dobrou approximaci se volí $Q = 5$.

Výpočet kepstrálních koeficientů PLP lze pak pro známé hodnoty $R(i)$, podobně jako u LPC, určit Durbinovým algoritmem. Nalezené koeficienty lze již využít jako příznaky při návrhu parametrizátoru řeči, ale častěji se používají kepstrální koeficienty PLP. [18]

K vytvoření parametrizátoru je možné použít libovolnou metodu představenou v 4.1.1 a 4.1.2. V současnosti, ale převládají metody postavené na principu fungování lidského sluchu, protože amplifikují podstatnou informaci zakódovanou v řeči.

4.2 Akustické modelování

Akustický model představuje v rovnici (4.3) podmíněnou pravděpodobnost $p(O|W)$. Úkolem akustického modelu je poskytnout co nejpřesnější odhad této pravděpodobnosti pro libovolnou posloupnost vektorů příznaků $O = \{o_1 o_2 \dots o_T\}$. Velmi vhodným způsobem modelování řeči se ukázalo využití tzv. **skrytých Markovových modelů (HMM)**. Ty vycházejí z principu vytváření řeči člověkem. V průběhu produkce řeči se hlasové ústrojí vždy nachází, v krátkém časovém úseku, v nějaké konfiguraci. Množina všech možných konfigurací je konečná. V tomto mikrosegmentu je pak hlasovým ústrojím generován krátký signál, který závisí právě na aktuální konfiguraci. Tento vyprodukovaný zvuk je metodami (popsanými v 4.1) převeden na vektor příznaků O .

Skrytý Markovův model je model stochastického procesu. Na ten je možné nahlížet jako na pravděpodobnostní konečný automat, který v diskrétních časových okamžicích generuje náhodnou posloupnost vektorů příznaků $O = \{o_1 o_2 \dots o_T\}$. Model v každém časovém kroku změní stav svůj s_j podle předem daných pravděpodobností

přechodu a_{ij} . Přechod ze stavu s_i do stavu s_j má za následek vygenerování výstupního vektoru pozorování o_t , a to podle rozdělení výstupní pravděpodobnosti $b_j(o_t)$ příslušné k tomuto stavu [20].

Podmíněná pravděpodobnost přechodu a_{ij} určuje, s jakou pravděpodobností přechází model ze stavu i v čase t , do stavu j v čase $t + 1$. Platí tedy

$$a_{ij} = p(s(t+1) = s_j | s(t) = s_i), \quad (4.31)$$

kde $s(t)$ je stav modelu v čase t . Další podmínkou je, že pro všechny stavy i , $i = 1, 2, \dots, N$, platí

$$\sum_{j=1}^N a_{ij} = 1. \quad (4.32)$$

Funkce rozdělení výstupní pravděpodobnosti $b_j(o_t)$ popisují rozdělení pravděpodobnosti pozorování o_t produkovaného ve stavu s_j v čase t . Pro tuto funkci platí

$$b_j(o_t) = P(o_t | s(t) = s), \quad (4.33)$$

kde P značí pravděpodobnost, pro kterou u spojitého rozdělení platí

$$\int_o b_j(o) do = 1, \quad (4.34)$$

kde toto platí pro všechny stavy HMM, které mohou generovat výstupní vektor.

Rozdělení výstupní pravděpodobnosti musí být při modelování řečových zvuků dostatečně specifické, aby bylo možné od sebe oddělit různé zvuky, a zároveň dostačně robustní, aby zahrnulo značnou variabilitu řečového signálu. Toto rozdělení je možné modelovat

- spojitym normálním rozdělením se směsí hustotních funkcí,
- neuronovými sítěmi.

4.2.1 Struktura skrytého Markovova modelu

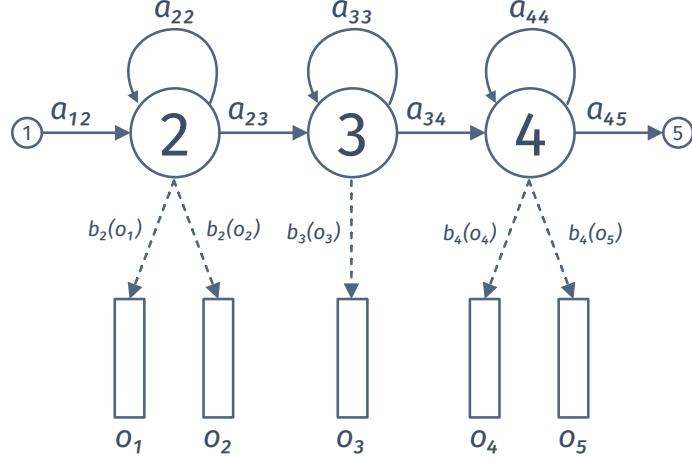
Z pohledu rozpoznávání řeči se nejčastěji využívá tzv. levo-pravá struktura Markovova modelu. V průběhu let bylo testováno mnoho různých struktur HMM, např. modely s počtem stavů odvozených od průměrné délky slova pro nějž byl model konstruován, až po pevnou strukturu stavů pro každé slovo. Tyto modely sloužily hlavně pro rozpoznávání izolovaných úseků řeči, nejčastěji slov. V současnosti, kdy je většina systémů konstruovaných pro zpracování souvislé řeči a počet slov ve slovníku může přesahovat 1 milion slov, převažují modely odvozené od menších jednotek, než jsou slova. Takovými jednotkami mohou být například fonémy anebo specifické trifóny. Trifón je svým způsobem kontextově závislý foném, který bere v potaz svůj levý a pravý kontext, tj. levý a pravý sousední foném. Přepis slova do fonémové, resp. trifónové struktury, lze ukázat na příkladu izolovaného slova „akcie“, které má přepis „**sil a k c i j e sil**“, v trifónové podobě je pak zápis následující

sil sil-a+k a-k+c k-c+i c-i+j i-j+e j-e+sil sil,

kde **sil** má význam pauzy před, případně za vyslovenou promluvou slova „akcie“.

Oproti slovním modelům, u fonémů (monofónů), resp. trifónů, bývá struktura relativně jednoduchá a často je vyjádřena 5 stavovým modelem (znázorněn na obr. 4.11). Jedná se o 5 stavový levo-pravý Markovův model, jehož první a poslední stav jsou tzv. neemitující. Jejich primární úlohou je zřetězování jednotlivých HMM modelů trifónů (monofónů) do rozsáhlejších modelů, např. slov, vět ap. Při zřetězení se tyto neemitující stavy vypouštějí. Ostatní stavy modelu jsou emitující a vztahují se k nim odpovídající rozdelení pravděpodobnosti $b_j(\cdot)$.

Pokud předpokládáme, že posloupnost slov W je modelována zřetězeným skrytým Markovovým modelem λ , kde dílčí modely odpovídají fonetickým jednotkám, pak je možné určit pravděpodobnost generování posloupnosti O modelem λ jako



Obrázek 4.11: Příklad levo-pravého Markovova modelu trifónu

$$P(O|\lambda) = \sum_{\forall S} P(O, S|\lambda) P(S|\lambda) = \sum_{\forall S} a_{s(0)s(1)} \prod_{t=1}^T b_{s(t)}(o_t) a_{s(t)s(t+1)}, \quad (4.35)$$

kde posloupnost stavů $S = \{s(0), s(1), \dots, s(T+1)\}$ je chápána tak, že $s(0)$ je vstupní a $s(T+1)$ výstupní neemitující stav modelu Θ dané promluvy [20]. Přitom tento model lze značit trojicí

$$\lambda = \left[\{a_{ij}\}_{k,s=1}^I; \{b_s(\cdot)\}_{s=1}^I; \{\pi_s\}_{s=1}^I \right], \quad (4.36)$$

kde a_{ij} je přechodová a $b_s(\cdot)$ výstupní pravděpodobnost. Dále π_s je rozložení pravděpodobnosti počátečního stavu a I je počet stavů modelu.

Přímé vyčíslení pravděpodobnosti $P(O|\lambda)$ podle vztahu (4.35) je z hlediska počtu operací často nerealizovatelné, protože se jedná řádově o $2TN^T$ operací násobení. Z tohoto důvodu se proto využívá výpočetně efektivnější tzv. **algoritmus forward-backward (FB)** s přibližně N^2T operací násobení.

Při výpočtu odpředu (forward) se určuje pravděpodobnost $\alpha_j(t)$ definovaná vztahem

$$\alpha_j(t) = P(o_1 o_2 \dots o_t, s(t) = s_j | \lambda), \quad (4.37)$$

pro výpočet odzadu (backward) se určuje pravděpodobnost $\beta_j(t)$ definována vztahem

$$\beta_j(t) = P(o_{t+1} o_{t+2} \dots o_T | s(t) = s_j | \lambda). \quad (4.38)$$

Podle [20] lze snadno dokázat, že výsledná pravděpodobnost $P(O|\lambda)$ může být vyčíslena vztahem

$$P(O|\lambda) = \sum_{s=1}^N P(O, s(t) = s | \lambda) = \sum_{i=1}^N \alpha_i(t) \beta_i(t) \quad (4.39)$$

pro $1 \leq t \leq T$.

4.2.2 Trénování parametrů HMM s Gaussovskými směsmi

Volba struktury skrytého Markovova modelu je spíše expertní úlohou návrhu. Stanovení hodnot parametrů modelu je uskutečněno trénováním (odhadem, estimací) na základě trénovacích akustických dat a jejich textových anotací (tzv. korpus). Pro trénování parametrů se využívá tzv. Baum-Welchův iterativní algoritmus, což je speciální případ EM algoritmu. Více o něm v [18]. Základem je vyčíslení $\gamma_j(t)$, což vyjadřuje pravděpodobnost, že proces generování posloupnosti O je v čase t ve stavu j . Pro její vyjádření je možné využít rovnice (4.39), výsledný vztah má poté tvar

$$\gamma_j(t) = \frac{P(O, s(t) = j | \lambda)}{P(O | \lambda)} = \frac{\alpha_j(t) \beta_j(t)}{P(O | \lambda)}, \quad (4.40)$$

kde $j = 1, \dots, N$ a $t = 1, \dots, T$. Pravděpodobnost, že proces generování posloupnosti O je v čase t ve stavu j a generuje složku m gaussovské hustotní směsi, pak určuje vztah

$$\gamma_{jm}(t) = \frac{P(O, s(t) = j, m(j, t) = m | \lambda)}{P(O|\lambda)} = \frac{\alpha_j(t) \beta_j(t)}{P(O|\lambda)} \frac{c_{jm} \mathcal{N}(o_t; \mu_{jm}; C_{jm})}{\sum_{i=1}^M c_{ji} \mathcal{N}(o_t; \mu_{ji}; C_{ji})}. \quad (4.41)$$

Pro odhad střední hodnoty rozložení μ_{jm} , tj. složky m gaussovské směsi ve stavu j slouží vztah

$$\hat{\mu}_{jm} = \frac{\sum_{t=1}^T \gamma_{jm}(t) o_t}{\sum_{t=1}^T \gamma_{jm}(t)}, \quad (4.42)$$

Ten se vyčísluje pro $1 \leq j \leq N$ a $1 \leq m \leq M$, kde N je počet stavů a M počet složek modelu. Odhad kovarianční matice C_{jm} , tj. složky náležící m -té složce gaussovské směsi ve stavu j

$$\hat{C}_{jm} = \frac{\sum_{t=1}^T \gamma_{jm}(t) (o_t - \hat{\mu}_{jm})(o_t - \hat{\mu}_{jm})^T}{\sum_{t=1}^T \gamma_j(t)}, \quad (4.43)$$

kde $1 \leq j \leq N$ a $1 \leq m \leq M$. Odhad váhové složky hustotní směsi c_{jm} , tj. složky náležící složce m gaussovské směsi ve stavu j se provádí vztahem

$$\hat{c}_{jm} = \frac{\sum_{t=1}^T \gamma_{jm}(t)}{\sum_{t=1}^T \gamma_j(t)}, \quad (4.44)$$

kde $1 \leq j \leq N$ a $1 \leq m \leq M$.

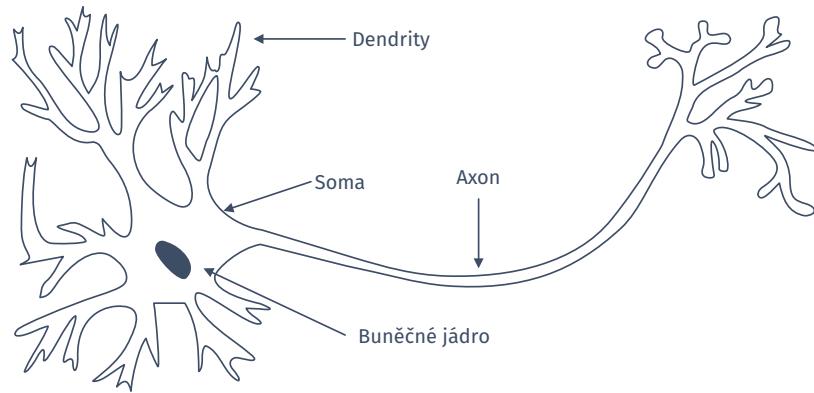
Rozdělení výstupní pravděpodobnosti $b_j(o_t)$ pro emitující stav j pak má tvar

$$b_j(o_t) = \sum_{m=1}^M \hat{c}_{jm} \mathcal{N}(o_t; \hat{\mu}_{jm}; \hat{C}_{jm}). \quad (4.45)$$

Celkový počet složek hustotních směsí se u modelů postavených na kombinaci skrytých Markovových modelů a gaussovských směsí nejčastěji pohybuje v rozmezí od 10 tisíc do 100 tisíc směsí. Při dimenzi příznakového vektoru (viz 4.1) vektoru například 15 je často nutné provést odhad až 10 miliónů parametrů.

4.2.3 Využití neuronových sítí

Neuronové sítě se inspirují neuronem v mozku člověka. Ukázka stavby neuronové buňky je znázorněna na obr. 4.12. Dendrity jsou krátké výběžky, které slouží k příjímání vstupních informací od ostatních neuronů nebo nervů. V tělo neuronu (soma) dochází k reakci na vstupní signály a vytvoření příslušné odezvy. Ta se dále šíří pomocí výběžku nazvaného axon. Jeho délka může dosahovat až 100 cm. Axon je přes synapse spojen s jinými neurony nebo dalšími buňkami v těle.



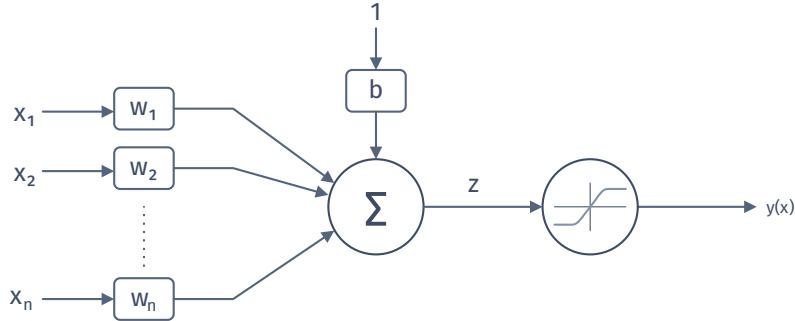
Obrázek 4.12: Ukázka neuronové buňky

Umělý ekvivalent s názvem perceptron byl vytvořen Frankem Roseblattem v první polovině 60. let 20. století [21]. Schématicky je zobrazen na obr. 4.13. Matematicky lze princip neuronu popsat vztahem

$$\hat{y}(x) = \sigma(z) = \sigma(w^T x + b) = \sigma\left(\sum_{j=1}^n w_j x_j + b\right), \quad (4.46)$$

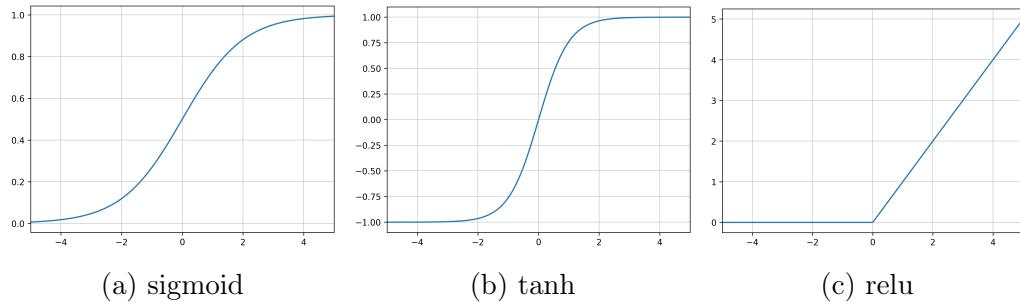
kde x představuje vstupní vektor, w váhový vektor a b práh. Výsledek lineární kombinace je vstupem aktivační funkce $\sigma(.)$, jejíž výstup je zároveň výstupem neuronu. Neuronová síť⁶ (NN, viz obr. 4.15) je složena z jedné či více vrstev neuronů. V případě více vrstev NN jsou vždy propojeny neurony mezi vrstvami l a $l+1$.

⁶Popisovaná neuronová síť je typu feedforward (FF). Dalšími typy sítí jsou konvoluční a rekurentní neuronové sítě. Oproti FF sítí se liší hlavně svou strukturou. Princip propojení neuronových buněk je však stejný.



Obrázek 4.13: Schéma perceptronu

Zmíněná aktivační funkce hraje velmi významnou roli, protože umožňuje řešení i nelineárních problémů. Pokud by NN nevyužívala aktivační funkce, jednalo by se de facto stále o lineární kombinaci vektorů, a tím pádem by bylo možné řešit jen lineární problémy. Mezi nejčastěji používané patří *sigmoid* ($\sigma(z) = (1 - e^{-z})^{-1}$), *tanh* ($\sigma(z) = \tanh(z)$) a *relu* ($\sigma(z) = \max(0, z)$). Průběhy těchto aktivačních funkcí jsou vidět na obr. 4.14.



Obrázek 4.14: Příklady používaných aktivačních funkcí

Pro výpočet výstupu neuronové sítě, tzv. **forward propagation**, je použit iterativní postup matematicky zapsán jako

$$\begin{aligned} \mathbf{Z}^{[l]} &= \mathbf{W}^{[l]} \mathbf{a}^{[l-1]} + \mathbf{b}^{[l]}, \\ \mathbf{a}^{[l]} &= \sigma^{[l]}(\mathbf{Z}^{[l]}), \end{aligned} \tag{4.47}$$

kde $a^{[l]}$ představuje výstup l -té vrstvy ($a^{[0]} = x$), $\mathbf{W}^{[l]}$ představuje váhovou matici l -té vrstvy, $\mathbf{b}^{[l]}$ vektor prahů l -té vrstvy a $\sigma^{[l]}(\cdot)$ aktivační funkci l -té vrstvy. Pro l platí $l = 1, \dots, L$, kde L je počet vrstev neuronové sítě. Výsledkem iterativního výpočtu (4.47) je výstup sítě $y = a^{[N]}$.

Trénováním neuronové sítě je myšleno určení hodnot váhových matic $\mathbf{W}^{[l]}$ a prahů $\mathbf{b}^{[l]}$. Tento proces se iterativně sestává ze 3 kroků (viz obr. 4.15)

1. výpočet výstupu sítě (4.47),
2. vypočtení chyby predikce $J(y, \hat{y})$,
3. aktualizace vah pomocí algoritmu backpropagation.

Výpočet výstupu NN je realizován pomocí (4.47). Následně je nezbytné vypočítat chybu predikce $J(y, \hat{y})$. Ta je definována vztahem

$$J(y, \hat{y}) = \frac{1}{M} \sum_{i=1}^M \mathcal{L}_i(y^i, \hat{y}^i), \quad (4.48)$$

kde M je počet prvků trénovací množiny a $\mathcal{L}_i(y^i, \hat{y}^i)$ je funkce výpočtu chyby predikce i -tého prvku trénovací množiny. Konkrétní funkce závisí na typu řešené úlohy, ale často se používá cross-entropie definované vztahem

$$\mathcal{L}_i(y^i, \hat{y}^i) = - \sum_{j=1}^m y_j^i \log \hat{y}_j^i, \quad (4.49)$$

kde m je dimenze výstupního vektoru.

Samotná aktualizace parametrů sítě je realizování **backpropagation** algoritmem. Cílem tohoto algoritmu je vypočtení parciálních derivací $\partial J / \partial \mathbf{W}^{[l]}$ a $\partial J / \partial \mathbf{b}^{[l]}$. Tyto parciální derivace je potřeba vypočítat pro všechny vrstvy sítě. Chyba ve vrstvě l je závislá na chybě v předchozí vrstvě $l - 1$. Tato skutečnost znamená, že je možné použít tzv. chain pravidlo. Parciální derivace pak mají následující podobu

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{W}^{[l]}} &= \frac{\partial J}{\partial \mathbf{a}^{[l]}} \frac{\partial \mathbf{a}^{[l]}}{\partial \mathbf{z}^{[l]}} \frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{W}^{[l]}}, \\ \frac{\partial J}{\partial \mathbf{b}^{[l]}} &= \frac{\partial J}{\partial \mathbf{a}^{[l]}} \frac{\partial \mathbf{a}^{[l]}}{\partial \mathbf{z}^{[l]}} \frac{\partial \mathbf{z}^{[l]}}{\partial \mathbf{b}^{[l]}}\end{aligned}\tag{4.50}$$

Vzorce pro výpočet aktualizací parametrů sítě jsou pak následující

$$\begin{aligned}\delta^{[L]} &= \nabla_a J \odot \sigma'(\mathbf{z}^{[L]}), \\ \delta^{[l]} &= \left((\mathbf{w}^{[l+1]})^T \delta^{[l+1]} \right) \odot \sigma'(\mathbf{z}^{[l]}), \\ \frac{\partial J}{\partial \mathbf{W}^{[l]}} &= \mathbf{a}^{[l-1]} \delta^{[l]}, \\ \frac{\partial J}{\partial \mathbf{b}^{[l]}} &= \delta^{[l]},\end{aligned}\tag{4.51}$$

kde $\nabla_a J = \partial J / \partial \mathbf{a}^{[L]}$ a \odot představuje Hadamardův součin. Samotná aktualizace parametrů je realizována vztahy

$$\begin{aligned}\mathbf{W}^{[l]} &= \mathbf{W}^{[l]} - \alpha \frac{\partial J}{\partial \mathbf{W}^{[l]}}, \\ \mathbf{b}^{[l]} &= \mathbf{b}^{[l]} - \alpha \frac{\partial J}{\partial \mathbf{b}^{[l]}},\end{aligned}\tag{4.52}$$

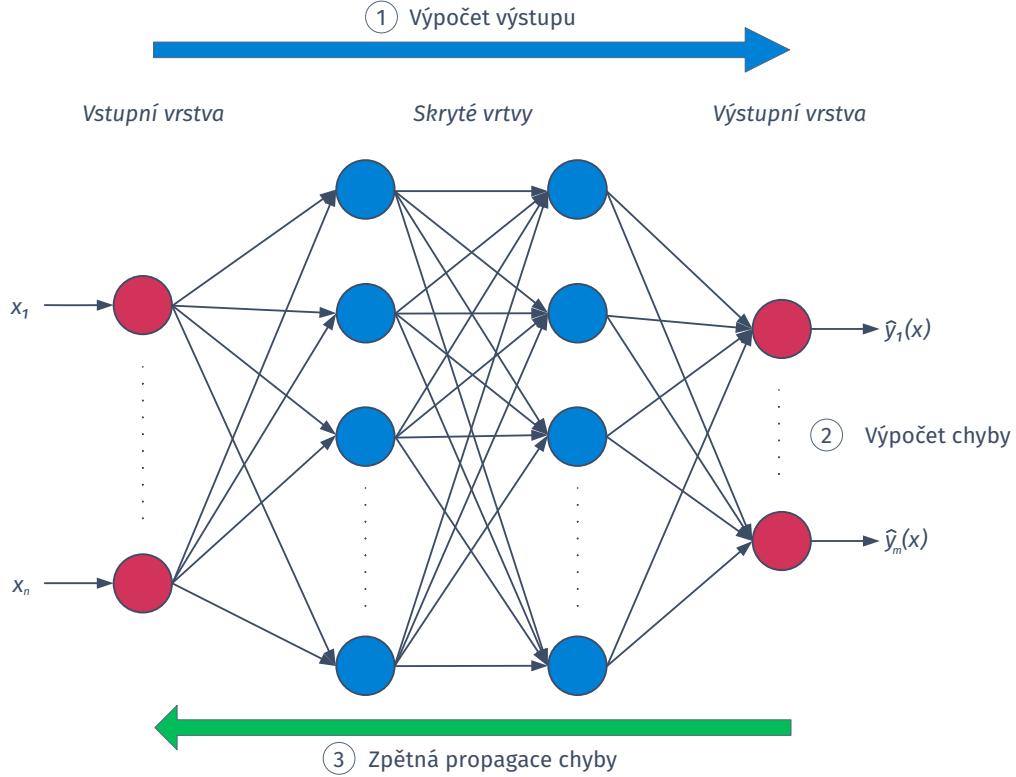
kde α reprezentuje koeficient učení.

Spojení skrytých Markovových modelů a neuronových sítí

Rozvoj výpočetní techniky, zejména GPU⁷ s možností provádět obecné maticové operace, zapříčinil masivní využití tzv. hlubokých neuronových sítí (DNN). Ty se vyznačují vyšším počtem skrytých vrstev, což umožňuje řešit sofistikovanější problémy. Jedním takovým je rozpoznávání souvislé řeči. Bohužel DNN end-to-end⁸ systém je

⁷Graphics Processing Unit

⁸Systém, který kompletne řeší rovnici (4.3) pomocí jediné DNN sítě. Tyto systémy jsou většinou postaveny na rekurentních neuronových sítích (RNN).



Obrázek 4.15: Schéma a princip učení neuronové sítě

zatím velmi komplikované vytvořit a provozovat zejména, protože k úspěšnému natrénovaní je potřeba řádově více dat, než u GMM [22]. Z tohoto důvodu jsou v současné době nejčastější systémy postavené na kombinaci HMM a DNN (HMM-DNN). Rozdíl oproti end-to-end systému je v tom, že cílem DNN není odhad \hat{W} , ale stejně jako v případě HMM-GMM, určit $b_j(o_t)$.

V případě HMM-GMM je odhad $b_j(o_t)$ realizován gaussovskými hustotními směsmi podle vzorce (4.45). Těchto směsí je tolik, kolik je unikátních stavů HMM. U DNN však žádné směsi k dispozici nejsou. Pokud je však výstupní vrstva typu **softmax**, kde výstup j -tého neuronu je definován vztahem

$$y_j = a_j^{[L]} = \frac{e^{z_j}}{\sum_{i=1}^m e^{z_i}}, \quad (4.53)$$

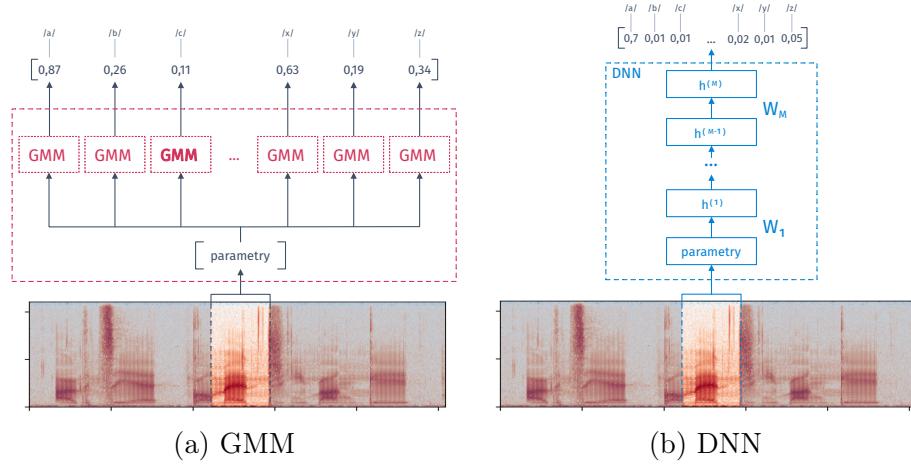
kde m je počet neuronů v poslední vrstvě. Zároveň platí

$$\sum_{j=1}^m y_j = 1. \quad (4.54)$$

Hodnoty výstupního vektoru y mají pseudo-pravděpodobnostní charakter. Pokud tedy bude m rovno počtu stavů HMM , pak výstupní pravděpodobnost $b_j(o_t)$ pro emitující stav j má, podle (4.53), tvar

$$b_j(o_t) = y_j = \frac{e^{z_j}}{\sum_{i=1}^m e^{z_i}}. \quad (4.55)$$

Principiální rozdíl ve funkci HMM-GMM a HMM-DNN je znázorněn na obr. 4.16.

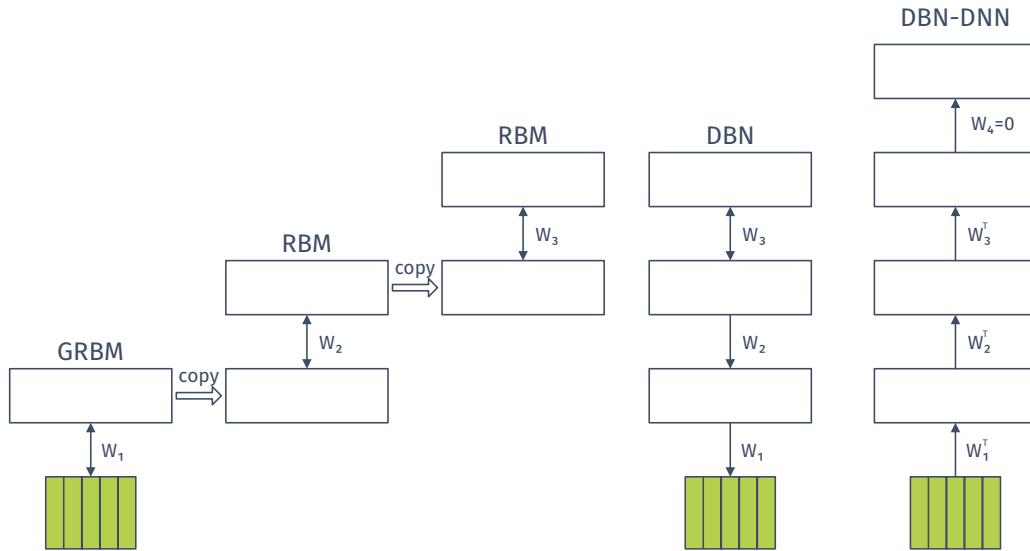


Obrázek 4.16: Principiální rozdíl ve funkci GMM a DNN systému

K natrénování DNN se používá zmíněného backpropagation algoritmu. V poslední době se však prosadilo trénování využívající předtrénování DNN pomocí tzv. restricted Boltzmann machines (RBM) [23]. Předtrénování řeší problém kdy se informace zpětně propagovaná pomocí backpropagation algoritmu úplně neovlivní počáteční vrstvy, protože gradient je příliš malý. Předtrénování pomocí RBM pomáhá lépe určit parametry sítě. Principiálně je tento proces znázorněn na obr. 4.17.

Nejprve je natrénován GRBM (Gaussian-Bernoulli RBM) model na mikrosegmentu řeči složeného minibatche skládající se z mikrosegmentů, každý odpovídající délce promluvy například 10 ms. Stav skrytých jednotek je použit k natrénování RBM.

Tento proces se opakuje dokud není natrénován požadovaný počet vrstev výsledné sítě. Následně jsou jednotlivé RBM spojeny do deep belief sítě (DBN). Následně je přidána výstupní softmax vrstva dimenze rovné počtu HMM stavů (DBN-DNN). Tato DBN-DNN síť je pak diskriminativně trénována na základě zarovnání získaného pomocí HMM-GMM. Více o tomto principu trénování v [23] a [24]. Vstupem neuronové sítě je často mikrosegment t a jeho okolní mikrosegmenty. Velmi často se používá okolí $t - 2$ a $t + 2$.



Obrázek 4.17: Princip předtrénování pomocí RBM s třemi vrstvami [23].

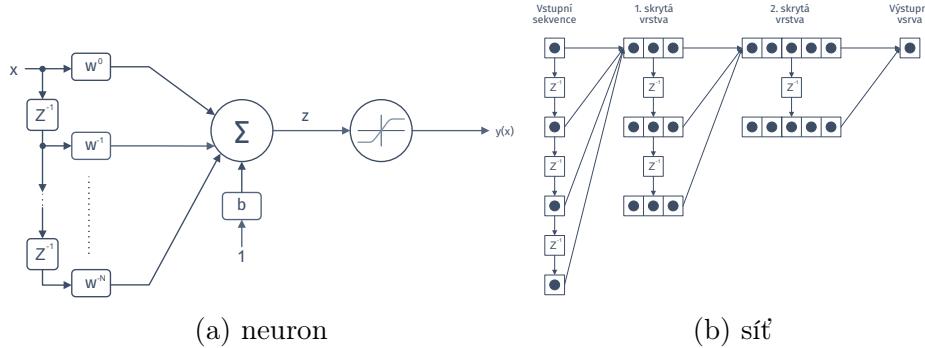
Time-delay neural networks

Nevýhodou DNN sítí spočívá ve faktu, že pracuje pouze se statickými parametry v rámci zpracovávaných mikrosegmentů, protože sumace v perceptronu odpovídá sumě vážených statických vstupů. Po zpracování segmentu t není získaná informace nijak reflektována při zpracování segmentu $t + 1$. Tento nedostatek by řešilo použití rekurrentních neuronových sítí (RNN). Bohužel tyto sítě mají mnohem komplikovanější buňku neuronu než FF sítě s perceptronem [22]. Proto je potřeba řádově více dat k natrénování. Složitost buňky také zvyšuje komputační náročnost výpočtu [25].

Základním rozdílem sítí typu time-delay neural network (TDNN) (představená v [26]) je přidání časové filtrace do sumiční části neuronu, tím je docíleno zahrnutí dynamické složky do výpočtu sítě [27]. Filtrace je implementována jako filtr s konečnou impulzní odezvou (FIR), tedy

$$z_j^{[l]}(t) = \sum_{n=0}^N w_j^{[-n]} f(n) a^{[l-1]}(t-n) + b_j, \quad (4.56)$$

kde t je diskrétní časový index, N je délka FIR filtru, $f(n)$ odezva filtru v bodě n , $w_j^{[-n]}$ příslušná váha, $a^{[l-1]}$ je výstup vrstvy $l-1$ a $z_j^{[l]}(t)$ je výstup sumiční části neuronu j ve vrstvě l . Vztah (4.56) tedy představuje konvoluci. Na obr. 4.18a je principiálně znázorněn neuron pracující s N FIR filtry. Z obr. 4.18a je také zřejmé, že TDNN síť má několik souborů vah W^x , které umožňují lépe pracovat s dynamickou složkou signálu [28].



Obrázek 4.18: Neuron TDNN sítě a zjednodušené blokové schéma TDNN sítě [27]

Stejně jako v případě DNN sítě je vstupem parametrizovaný mikrosegment t a jeho okolí. Z obr. 4.18b je patrné, že hlubší vrstvy postupně zpracovávají větší a větší okolí mikrosegmentu t . Dimenze výstupní vrstvy odpovídá počtu HMM stavů. Přestože je TDNN síť typu FF, tak dokáže pracovat i s dynamickými parametry řeči, protože využívá princip konvoluce.

4.3 Jazykové modelování

Jazykový model (obr. 4.1) je po parametrizaci a akustickém modelu další důležitou částí systému rozpoznávání řeči. Jeho úkolem je poskytnout dekodéru co nejrychleji nejpřesnější odhad apriorní pravděpodobnosti $P(W)$ pro libovolnou posloupnost slov W . Tuto pravděpodobnost je možné vyjádřit vztahem

$$P(W) = \prod_{k=1}^K P(w_k | w_{k-1} \dots w_1), \quad (4.57)$$

kde K je počet slov posloupnosti W . Pokud by byl proveden rozklad (4.57) vyšlo by najev, že pravděpodobnost výskytu slova $P(w_i)$, $i \leq K$ je podmíněna pouze svou historií, tj. posloupností slov $w_1 \dots w_{i-2}w_{i-1}$.

Systémy rozpoznávání řeči pracují obvykle s rozsáhlými slovníky, čítající stovky tisíc až jednotky milionů slov, není možné předpokládat, že by bylo možné pravděpodobnosti v (4.57) dostatečně robustně odhadnout pro libovolnou délku posloupnosti K .

Obvykle se proto provádí approximace vztahu (4.57), při nichž dochází k redukci počtu odhadovaných parametrů. Nejčastějším způsobem je stanovení ekvivalentních tříd slov na základě jejich slovní historie, tj. všechny historie $w_1 \dots w_{i-2}w_{i-1}$, které se shodují v posledních $n - 1$ slovech, jsou zařazeny do stejné třídy. Uvedené modely se nazývají **n -gramové modely**. Přitom n -gramem se rozumí posloupnost n za sebou jdoucích slov v pozorování jejich náhodného výběru, např. trénovacího korpusu obsahujícího textová data. Modely s $n = 0$ se nazývají **zero gramy**, $n = 1$ pak **unigramy**. Nejpoužívanější jsou pak **bigramy** ($n = 2$) a **trigramy** ($n = 3$). Pravděpodobnost $P(W)$ u n -gramového modelu se vypočte vztahem

$$P(W) = \prod_{k=1}^K P(w_k | w_{k-n+1} \dots w_{k-1}). \quad (4.58)$$

V ideálním případě by optimální model měl mít $n > 3$, ale v praxi se tyto modely moc nepoužívají, protože s rostoucím řádem modelu enormně roste potřebná velikost trénovacích dat. Například pro slovník s N položkami existuje stále N^n n -gramových statistik, které je potřeba odhadnout. Jak bylo zmíněno odhad těchto statistik se provádí na základě relativních četností v trénovacích datech. Například u bigramů ($n = 2$) a slovníku o velikosti $N = 10^5$ je zapotřebí odhadnout 10^{10} různých bigramů a k tomu je zapotřebí relativně velké trénovací množiny. Je zřejmé, že většina z těchto 10^{10} bigramů se vůbec neobjeví v datech. Těmto „neviděným“ bigramům tedy odpovídá nulová pravděpodobnost, což vyústí v nulovou pravděpodobnost $P(W)$ (4.58). K řešení tohoto problému se používá technik „vyhlazování“. Jejich cílem je odhad pravděpodobností těchto neviděných jevů s využitím tzv. ústupových, interpolačních a diskontních schémat [20].

Výstupem akustického modelu jsou většinou fonémy ve zvolené fonetické abecedě (např. SAMPA). Nezbytnou součástí systémů rozpoznávání řeči, tak je výslovnostní slovník, který obsahuje kombinace slov a fonetického přepisu těchto slov. Tento slovník umožňuje výpočet $P(W)$ na základě výstupu akustického modelu.

4.4 Dekódování

Hlavní funkcí dekodéru (viz obr. 4.1) je řešení rovnice

$$\hat{W} = \operatorname{argmax}_W p(O|W) P(W), \quad (4.59)$$

kde $p(O|W)$ představuje již popsaný akustický model a $P(W)$ pak ten jazykový. Někdy je úloha dekódování zobecněna na nalezení více než jedné posloupnosti slov \hat{W} . O té se pak mluví jako o hledání **N nejlepších** (N -best) posloupností slov \hat{W} . Řešení této úlohy je netriviální, protože dekodér obvykle nemá informaci o počtu slov v dané promluvě, protože ASR systémy nevyžadují vyslovování pauz mezi jednotlivými slovy.

Navíc, i kdyby tato informace byla k dispozici, tak pro promluvu, která čítá M slov, tak pak se slovníkem čítajícím N slov, je potřeba prozkoumat N^M různých slovních kombinací (hypotéz), tj. například 10^{50} vyhodnocení při $N = 100000$ a $M = 10$. Z toho jasně plyne, že aplikace metody vyčerpávajícího prohledávání je i pro úlohu s malými slovníky a krátkými promluvami nerealizovatelná.

Naštěstí bylo navrženo několik účinných algoritmů, které řeší hledání maxima v rovnici (4.1) bez exponenciálního nárůstu počtu výpočtů. Mezi takové algoritmy patří dekódování podle **kritéria maximální aposteriorní pravděpodobnosti (MAP)**, nebo v současnosti primárně používaného dekódování podle **Viterbiova kritéria**.

Akustický model určuje $p(O|W)$, resp. $p(O|\lambda)$, pomocí forward-backward (FB) algoritmu. Ten pro pozorovanou posloupnost \mathbf{O} určí pravděpodobnosti všech možných cest délky T modelem λ . Výpočet podmíněné pravděpodobnosti lze approximovat pravděpodobností $P_S(O|\lambda)$, jako nejpravděpodobnější posloupností HMM stavů, kterou projde posloupnost \mathbf{O} modelem λ

$$p(O|\lambda) \approx P_S(O|\lambda) = \max_S P(O, S|\lambda) = \max_S a_{s(0)s(1)} \prod_{t=1}^T b_{s(t)}(o_t) a_{s(t)s(t+1)}. \quad (4.60)$$

Tuto pravděpodobnost i optimální posloupnost stavů lze určit tzv. **Viterbiiovým algoritmem** [18]. Algoritmus řeší úlohu s využitím prohledávání typu beam, což je heuristický prohledávací algoritmus, který vždy expanduje pouze několik nejslibnějších uzel. Tím pádem dochází k urychlení výpočtů časově synchronního prohledávání, protože dochází k prořezávání neperspektivních hypotéz.

Pro další urychlení dekódování (zejména u systému pracujících v reálném čase), bylo navrženo několik dalších sofistikovaných postupů. Mezi takové může patřit využití tzv. lexikálních stromů, dalších technik prořezávání případně zjednodušení akustického modelu slova. Více o této problematice v [20].

U reálného systému je často potřeba vyřešit nebo „vybalancovat“ poměr příspěvků pravděpodobností od akustického a jazykového modelu. Z principu ASR systémy upřednostňují při dekódování krátký slova, což způsobuje chybu typu vložení. Ta se kompenzuje tzv. penaltou vložení, která mění měřítka $p(O|W)P(W)$ v závislosti na počtu slovních hypotéz. Jinými slovy penalizuje vložení krátkého slova pokud se jako „lepší“ jeví delší slovo. Vyvážení příspěvku jazykového modelu se ve většině systémů používá tzv. „grammar scale factor“. Zohledněním těchto poznatků, je možné rovnici (4.59) zapsat ve tvaru

$$\hat{W} = \operatorname{argmax}_W [\log p(O|W) + \kappa_1 \log (P(W) + \kappa_2 H)], \quad (4.61)$$

kde κ_1 je faktor změny měřítka, κ_2 je penalta vložení a H celkový počet slov v hypotéze. Parametry κ_1 a κ_2 jsou většinou nastavovány experimentálně.

V úloze rozpoznávání spojité řeči se vyskytují 3 typy chyb

- *substituce (S)* - došlo k rozpoznání špatného slova;
- *deletace (D)* - došlo k vynechání nějakého slova;
- *inzerce (I)* - došlo k vložení slova, které nebylo součástí W ;

K evaluaci schopností systému rozpoznávání řeči se pak využívá vzorce pro výpočet míry chybovosti na slovech (WER)

$$WER = \frac{C(S) + C(D) + C(I)}{N}, \quad (4.62)$$

kde N představuje počet slov v \hat{W} a $C(\cdot)$ je celkový počet chyb konkrétního typu. Čím je WER nižší, tím je systém lepší.

Velmi často se také používá metrika přesnosti rozpoznání udávaná v [%]. Stejně jako WER se počítá na základě chyb systému pomocí vzorce

$$Acc = \frac{N - C(S) - C(D) - C(I)}{N} * 100 = (1 - WER) * 100. \quad (4.63)$$

Oproti WER je systém s vyšší přesností lepší než systém s nižší přesností.

Kapitola 5

Konstrukce ASR systému pro uživatele po totální laryngektomii hovořící pomocí elektrolarynxu

5.1 Vytvoření řečového korpusu EL promluv

Před započetím libovolných prací na vytvoření ASR systému pracující s lidmi po TL je potřeba vytvořit řečový korpus, který poslouží k natrénování a otestování vytvořeného systému. Tato data jsou velmi specifická a je proto potřeba zajistit co možná největší množství kvalitních¹ a přesných dat, které budou součástí řečového korpusu.

V části 3.1 bylo zmíněno, že ročně se objeví více než 100 nových případů trvalé ztráty hlasu. V [2] bylo řečeno, že více rizikovými osobami jsou starší lidé, kteří intenzivně kouří a konzumují alkohol. Přesto je patrný trend snižujícího se věku pacientů a s tím související nárůst případů ztráty hlasu. Přičteme-li již zmíněný psychologický

¹Kvalitou je myšlena věrnost dat dané doméně, dále se mluví o přesnosti ve smyslu bezchybnosti přepisů.

aspekt jeho ztráty, je zřejmé, jak komplikované je získat ke spolupráci i jen jednoho řečníka ochotného podstoupit náročné² nahrávání.

Při libovolné práci s pacienty po TL, dřív nebo později dojde k určité formě spolupráce s oddělením ORL, které má na starosti péči o tyto pacienty. V našem případě nejprve s ORL klinikou při Fakultní nemocnici v Plzni a poté i s ORL klinikou Fakultní nemocnice v Motole. S jejich pomocí jsme získali ke spolupráci jednoho řečníka. Konkrétně se jedná o dámu v důchodovém věku, která podstoupila TL před více než 15 lety. Po překonání ostychu³ se byla schopna naplno vrátit do běžného života a dokonce v určité formě opět přednášet o stomatologii na Lékařské fakultě v Plzni, Univerzity Karlovy.

S její pomocí jsme, v 1. etapě nahrávání, byli schopni pořídit přes 10 hodin promluv, viz tab. 5.1. Získaná data neobsahují žádný nežádoucí ruch, kromě samotného zvuku EL i přesto, že nahrávání neprobíhalo v profesionálním studiu.

Nahrávací aparatura sestávala z miniaturního profesionálního mikrofonu (DPA d:screet 4061-FM), zesilovače (DPA MMA6000), externí zvukové karty a běžného notebooku. Mikrofon byl pomocí bezpolštářkové náplasti přilepen poblíž pravého koutku úst, aby zaznamenaná řeč měla co možná nejvyšší kvalitu.

Celé nahrávání bylo v 1. etapě rozděleno do 14 samostatných sezení a probíhalo od prosince roku 2010 do května roku 2011. Každé sezení trvalo přibližně dvě hodiny, během kterých se podařilo získat necelou hodinu akustických dat. Samotné nahrávání se sestávalo z 10 - 20 minutového úseku nahrávání a přibližně 10 minut dlouhého odpočinku. Ten byl nezbytný hlavně z důvodu únavy řečníka.

Před samotným nahráváním byly, z databáze obsahující stovky tisíc vět, pečlivě vybrány a vytvořeny 2 sady vět:

1. sada obsahující všechny možné české fonémy - 40 vět.

²I pro zdravého člověka je někdy několikahodinové nahrávání vysilující. Pro jedince po TL to je z mnoha důvodů ještě řádově náročnější.

³Podle jejích vlastních slov nebyla schopna několik let po operaci ani zvednout nečekaný telefonní hovor, natož mluvit na veřejnosti.

2. sada obsahující věty s reálnou četností fonémů - 5000 vět [29].

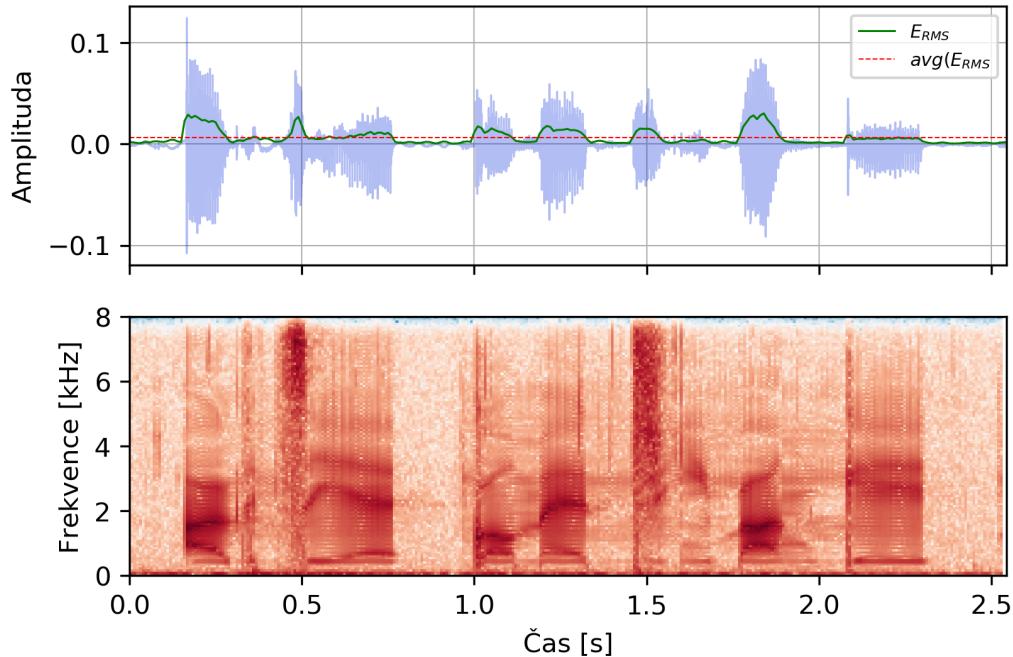
Pořízené nahrávky vždy odpovídají 10 - 20 minutovému úseku nepřerušovaného na hrávání. Výsledné soubory vždy obsahují několik vět. Ty jsou od sebe odděleny minimálně 5 sekundovým úsekem ticha. Nahrávky dále mohou obsahovat opakování chybně vyslovené věty, přeřeknutí, kýchnutí a další neřečové události. Z tohoto důvodu bylo nezbytné pořízené nahrávky anotovat, přestože byly pořízené na základě připravené sady vět.

Ještě před samotným anotováním byly nahrávky, podle úseků s tichem, rozsekány na menší části. V tomto případě se jako nejfektivnější ukázala metoda voice activity detection (angl. zkratka VAD) založená na principu energie. Pro každou nahrávku obsahující více vět se pomocí vzorce

$$E_{RMS}(n) = \sqrt{\frac{1}{N} \sum_{n=1}^N |x(n)|^2}, \quad (5.1)$$

kde N představuje počet vzorků v nahrávce a $x(n)$ představuje pravoúhlé okénko vzorku n . Pro tento případ se ukázalo jako vhodnější volit root-mean-square energy (E_{RMS}) a empiricky se ukázalo, že vhodná délka okénka je v rozmezí 10 – 100 ms. Na obr. 5.1 je zobrazena podoba audio signálu a spektrogram promluvy „Akcie Komercní banky“. Zároveň jsou zde vyneseny vypočtené hodnoty energie a celková průměrná energie. Tyto hodnoty slouží pro určení míst kde začíná a končí věta. Na začátku a konci každého úseku je vhodné mít minimálně 0.5 s ticha, aby bylo zajištěna správná funkce výsledného ASR systému, viz 4.1. Tím pádem, pokud energie nějakého úseku x je $E_{RMS}(x) < avg(E_{RMS})$ a zároveň délka tohoto úseku $dur(x) \geq 1 [s]$, tak je možné nahrávku v tomto úseku rozdělit.

Pokud řečník v průběhu věty z libovolného důvodu udělal pauzu větší než 1 s, tak v důsledku výše popsaného postupu byla věta rozdělena na dvě části. Nejedná se však o významný problém, protože výsledné kratší úseky jsou anotovány. Při vytváření ASR



Obrázek 5.1: Průběh a spektrogram promluvy a vyznačenou energií EL promluvy.

systému není podstatné zda promluva představuje celou větu, ale spíše to, jestli je tento úsek správně přepsán. Fakt, že některé věty jsou rozděleny, je důvodem proč v tab. 5.1 je více souborů než vět.

K anotaci posloužil interní anotační nástroj a podíleli se na ní celkem 3 anotátoři z řad studentů. Přepis jednoho anotátora, byl vždy zkontovalován jiným anotátorem. Ačkoliv bylo potřeba přepsat relativně malé množství dat (cca 10 hodin audio záznamu), tak anotace všech promluv zabrala přibližně 2 měsíce. Hlavním důvodem byla relativně dlouhá doba, po kterou se anotátoři adaptovali na specifika EL řeči. Hlavně ze začátku nebyli schopni porozumět obsahu promluvy, a tím pádem jej správně přepsat. To významně prodloužilo dobu potřebnou k anotaci celého řečového korpusu.

Pokud je k produkci řeči použit elektrolarynx, tak vedlejším produktem je nezanebatelný ruch způsobený samotným zařízením, viz část 3.2.1. Přeci jen jeho jedinou funkcí je vybudit vzduch v dutině ústní, a tím umožnit produkci slyšitelné řeči. Z

tohoto důvodu byly v průběhu anotace ignorovány v podstatě všechny skupiny neřečových událostí, protože většina nahrávek by byla anotována jako, že obsahují šum.

Výsledný řečový korpus představuje 5040 unikátních vět rozdělených do 6385 souborů (viz tab. 5.1), které v průměru obsahují 7 slov o průměrné délce 5 znaků. Tento korpus slouží jako základ pro všechny budoucí experimenty.

Nahrávání	Délka /HH:MM:SS/	Počet vět	Počet souborů
2010.12 - 2011.05	11:42:42	5040	6385

Tabulka 5.1: Informace o korpusu nahrávek z 1. etapy nahrávání.

5.2 Analýza akustického signálu a jeho parametrizace

Rozpoznávání řeči se věnuje nemalé úsilí již od 50. let 20. století a v současné době nikoho nepřekvapí téměř bezchybně fungující obecný rozpoznávač souvislé řeči v mobilních zařízeních. Pro obecné systémy dokonce existují korpusy s desítkami, stovkami i více hodin promluv, které je možné využít při vytváření těchto systémů.

Tyto korpusy však obsahují, ve většině případů, pouze „standardní“⁴ řeč. Pokud je snaha vytvořit nebo ověřit funkčnost systému za specifických podmínek (ať už se jedná o rušné prostředí či speciální typy promluv), tak je nezbytné získat potřebná data, viz 5.1.

⁴Slovním spojením „standardní řeč“ je myšlena řeč neobsahující výrazné řečové vady, případně jiné formy produkce a často v nepříliš akusticky náročném prostředí.

5.2.1 Analýza získaných dat

Získaný korpus obsahuje přes 10 hodin akustických záznamů promluv a více či méně přesných přepisů⁵. V momentě, kdy jsou k dispozici data, je možné se podívat na specifika EL řeči a případně porovnat se zdravým řečníkem.

Pro potřeby porovnání byl použit začátek promluvy „*Akcie Komerční banky...*“. Tato promluva je součástí standardní množiny vět používaných při vytváření řečových korpusů na KKY při ZČU. Tím pádem je k dispozici v relativně velkém množství příkladů pro zdravé řečníky. Tato věta je součástí také korpusu EL řeči.

Na obr. 5.2 je zobrazen průběh amplitudy a spektrogram vybrané promluvy pro zdravého (obr. 5.2a) a EL (obr. 5.2b) řečníka. Už na první pohled je možné zaznamenat určité rozdíly i přesto, že obsah obou promluv je identický. Prvním takovým je délka promluvy. V případě zdravého řečníka je v průměru⁶ o celou 1 vteřinu kratší než v případě EL řeči. Tempo řeči je samozřejmě velmi individuální, ale z principu je EL řeč pomalejší. Z průběhu signálu na obr. 5.2b je patrné, že řečník dělá výraznější pauzy mezi jednotlivými slovy promluvy. To je často způsobené potřebou naplnit jícenem vzduchem. Po TL je dýchání realizováno přes tracheu a pokud nebyl voperován shunt (více v 3.2.2), tak je trvale oddělen hrtan a hltan. Přesto, pro produkci některých neznělých fonémů je potřeba exhalovat vzduch z dutiny ústní. Zkušený EL řečník to dělá naprostě automaticky, nicméně „polykání“ vzduchu zabere nějaký čas. Nevyhnutelným důsledkem je pak velmi častý výskyt samovolného říhání v průběhu promluvy⁷.

Svou roli může hrát i snaha správně artikulovat. Při používání EL je nezbytné, aby bylo produkované řeči alespoň trochu rozumět. A pokud se dobře artikuluje,

⁵I přes nemalou snahu a několikastupňovou kontrolu, je téměř jisté, že by nebylo obtížné najít přepis, který obsahuje chybu například ve formě překlepu.

⁶Vypočteno na základě 10 náhodně vybraných promluv ze standardně používaného korpusu na katedře kybernetiky ZČU.

⁷Fakt, že je říhání jako neřečová událost běžnou součástí téměř každé promluvy, vedl k ignorování těchto událostí během anotace.



(a) Zdravý řečník

(b) EL řečník

Obrázek 5.2: Průběh a spektrogram promluvy a vyznačenou energií promluvy zdravého a EL řečníka.

není snadné mluvit rychle. Při nahrávání bylo velmi běžné, že v průběhu promluvy řečník udělal pauzu, aby mohl lépe umístit EL, protože jeho umístění má velký vliv na kvalitu produkované řeči. Nicméně je třeba říci, že tempo není a priory pro ASR systémy problém, protože různá délka fonémů je v relativně snadno modelována, viz část 4.2.

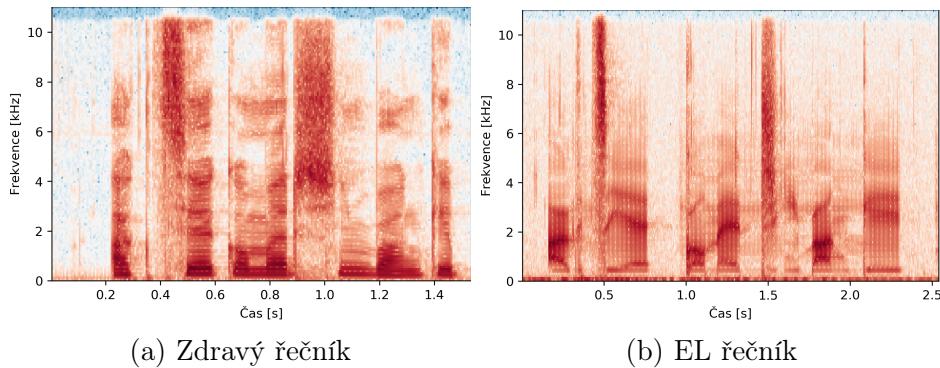
Dalším způsobem jak ukázat rozdíly mezi promluvou zdravého a řečníka s EL, je srovnání ve frekvenční oblasti. Pro větší názornost jsou na obr. 5.3 zobrazena společně spektra ukázkové promluvy zdravého řečníka (5.3a) a toho s EL (5.3b). Obsah obou promluv je identický a přesto jsou obě spektra odlišná.

Prvním markantním rozdílem je mnohem větší zastoupení šumu v úsecích „ticha“, viz obr. 5.3b. To je nepochybně způsobeno samotným EL, který řečník nevypíná mezi jednotlivými slovy. Na obr. 5.1 je zřetelně patrný, zejména na průběhu energie, šum před prvním a druhým slovem promluvy. Zajímavá je přítomnost šumu v celém frekvenčním spektru, přestože EL produkuje konstantní buzení. To je ve spektru (obr. 5.3b) viditelné jako výrazná souvislá linie v nízkých frekvencích. Přítomnost šumu ve vyšších frekvencích je způsobena umístěním mikrofonu, který je nalepen přímo na pokožku, a tím pádem snímá namodulované vibrace, přenášené měkkou tkání. Tento fakt se potvrdil v dalších etapách nahrávání (viz část 6.1), kde je použit studiový

mikrofon vzdálený od úst minimálně 15 cm a tyto vibrace již nezaznamenává. Nicméně z pohledu použitelnosti nějakého budoucího systému je nezbytné počítat i se situací, kdy mikrofon zaznamenává i vibrace přenášené tkání.

Dalším významným rozdílem je absence vyšších frekvencí u většiny produkovaných fonémů. Výjimku tvoří afrikáty /c/ a /č/, u kterých jsou hlasivky (u zdravého jedince) v klidu, a vznikají uvolněním nahromaděného vzduchu v dutině ústní⁸ [20]. U těchto fonémů není, u řečníka po TL, mechanizmus produkce těchto fonémů ovlivněn. Problémem teoreticky může být zdroj vzduchu, jelikož jej z plic není možné dostat do dutiny ústní, ale jak už bylo zmíněno (a spektrogram to potvrzuje), zkušený uživatel EL se dokáže adaptovat.

Absence vyšších frekvencí se dá vysvětlit použitím EL, kde samotný EL má vždy konstantní frekvenci buzení a dále tím, že nedochází k modulaci ve všech dutinách vokálního traktu. Nicméně nejdůležitější složky, zajišťující srozumitelnost, se vyskytují ve frekvenčním pásmu od 1 kHz do 3 kHz. Vyšší frekvence se a priory podílejí na zabarvení hlasu.



Obrázek 5.3: Spektrogram promluvy „Akcie Komerční banky“ dvou řečníků.

Dalším způsobem jak porovnat řeč zdravého a EL řečníka je pomocí analýzy jednotlivých fonémů. Na obr. 5.4, 5.5 a 5.6 jsou zobrazeny průběhy amplitudy v čase⁹ pro fonemy /k/, /g/ a /č/. V případě /k/ a /g/ (obr. 5.4 a 5.5) se jedná o okluzivy, kde

⁸Nahromadění vzduchu je realizováno přitisknutím jazyka k přední/zadní části horního patra.

⁹Hodnoty času odpovídají časům výskytu v původní promluvě.

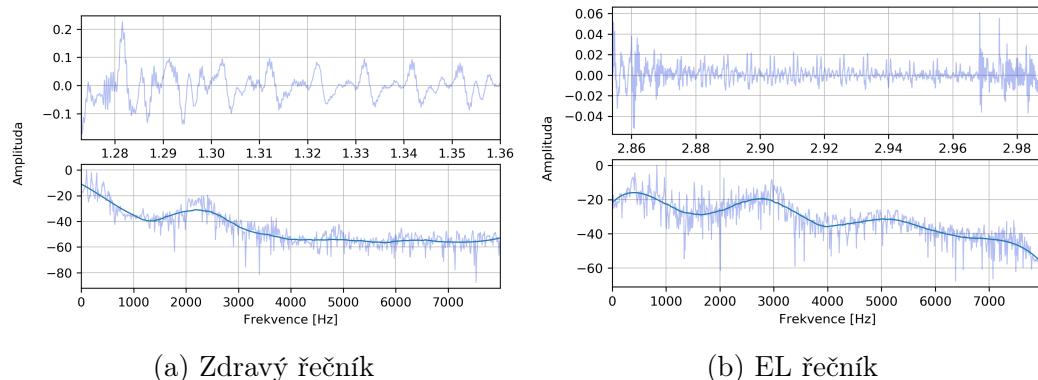
v prvním případě jde o neznělou plozivu a v druhém o znělou plozivu. Tyto fonémy obecně vznikají uzavřením vydechovaného proudu vzduchu pomocí artikulačních orgánů, což se projeví jako krátká pauza (tzv. okluze). Po té následuje náhlé jednorázové uvolnění překážky a únik nahromaděného vzduchu, tzv. exploze [20]. Takto popsáno to samozřejmě funguje u zdravého jedince, ale u EL řečníka jde sice o stejný mechanismus, ale s tím rozdílem, že vzduch nepochází z plic, ale z hltanu. Dalším rozdílem je samozřejmě absence hlasivek.

Foném */k/* je tedy zástupcem neznělých fonémů, ty se vyznačují tím, že do jejich produkce nevstupují hlasivky, které jsou v klidu. Zdrojem buzení je tedy šum, viz část 4. Pokud se podíváme na průběh amplitudy v čase u zdravého řečníka (obr. 5.4a), tak zde není vidět žádný periodický signál. Hlasivky jsou tedy opravdu v klidu. Oproti tomu u EL řečníka (obr. 5.4b) je jasně patrné, že je zde přítomno aktivní buzení vytvořené EL. Ve frekvenční oblasti je zobrazeno tzv. amplitudové spektrum, které znázorňuje vývoj amplitudy signálu ve frekvenci. V případě zdravého řečníka odpovídá vývoj předpokladům, není zde žádná výrazná frekvence a také nedochází k výraznému útlumu. Přestože se v obou případech jedná o stejný foném, tak z časového i frekvenčního průběhu je zřejmé, že parametry signálu se u obou řečníku diametrálně liší.



Obrázek 5.4: Průběh amplitudy */k/* v časové a frekvenční oblasti fonému u normálního a EL řečníka

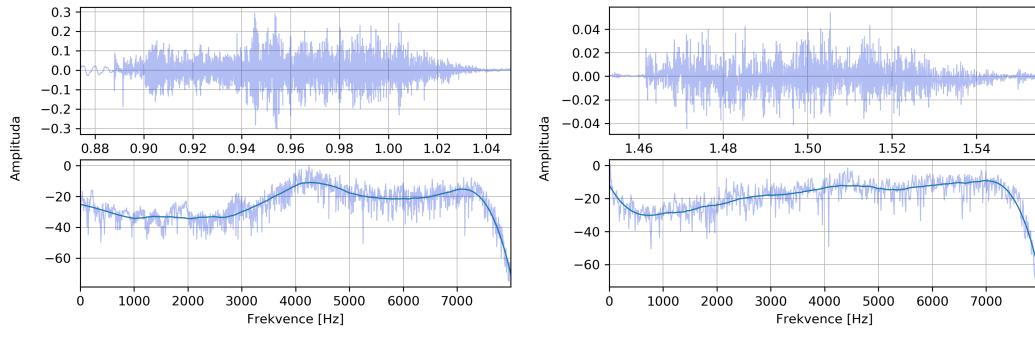
Jako druhý ukázkový foném slouží */g/*. Jedná se o plozivu, ale v tomto případě znělou. U těchto fonémů hrají velký vliv hlasivky, protože jsou zdrojem buzení. Z obr. 5.5a je toto buzení zřetelné ve formě periodického průběhu amplitudy. U EL řečníka (obr. 5.5b) je také vidět periodický signál, ale úplně jiného charakteru. Svým způsobem dost podobný tomu, který je zřetelný u fonému */k/*. Rozdíl je zřetelný i ve frekvenční oblasti, kdy u EL řečníka nedochází k útlumu ve střední oblasti frekvenčního spektra.



Obrázek 5.5: Průběh amplitudy fonému */g/* v časové a frekvenční oblasti fonému u normálního a EL řečníka

Posledním ukázkovým fonémem je již zmínované */č/*. Jedná se o neznělý foném, který vzniká přiložením jazyku k zadní části horního patra. Tím je zadržen vzduch v dutině ústní a vzniká krátká pauza. Uvolněním pak dochází k explozi a vytvoření zvuku. [20] Do produkce se nezapojují hlasivky a produkovaný zvuk by měl být dostatečně intenzivní, aby jej (v případě EL řečníka) také neovlivňoval EL. Tím pádem by měl být průběh signálu, u obou řečníků podobný, a to jak v časové, tak i ve frekvenční oblasti, viz obr. 5.6.

Z doposud provedené analýzy plyne, že EL řeč je v mnoha charakteristikách odlišná od té produkované zdravým řečníkem. Zejména u porovnání ve frekvenční oblasti (obr. 5.4 a 5.5) je to nejvíce patrné.



Obrázek 5.6: Průběh amplitudy fonému /č/ v časové a frekvenční oblasti fonému u normálního a FL řečníka

5.3 Aplikace obecného systému rozpoznávání a dosažené výsledky

Z provedené analýzy plyne, že získaný EL korpus je odlišný od „standardního“ řečového korpusu, který se běžně používá k trénování obecných akustických modelů. Tyto modely jsou nezávislé na řečníkovi a vyznačují se robustností. Je tedy otázka, zda takovýto model nebude schopen pracovat s EL daty.

K ověření byl vytvořen TDNN akustický model (více o těchto modelech v části [4.2.3](#)), který byl natrénován daty z korpusu čítající 1000 hodin promluv od velkého počtu řečníků. Celkový počet HMM stavů je **XXXX**.

Jazykový model je postaven na trigramech a k jeho natrénování posloužil textový korpus čítající velké množství novinových článků, webových reportáží, filmových titulků a dalších textových záznamů. Slovník jazykového modelu čítá více než 1 milion unikátních slov.

Testovacím vstupem vytvořeného ASR systému jsou data z EL korpusu. Celková slovní přesnost, počítaná podle vzorce (4.63), dosáhla hodnoty 18,49 %¹⁰.

¹⁰Dosaženo na state-of-the-art ASR systému v době psaní práce. V době vytvoření EL korpusu (kolem roku 2011) převládaly HMM-GMM akustické modely. Tento systém dosáhl přesnosti na slovech 12,59 %.

Dosažený výsledek zřetelně ilustruje odlišnost EL domény, protože obecný na řečníkovi nezávislý systém s velkým jazykovým modelem není schopen obstojně rozpoznat EL promluvu.

Pokud jsou k natrénování akustického modelu (taktéž využívajícího TDNN síť) použita pouze data¹¹ z EL korpusu, tak výsledná slovní přesnost dosáhla hodnoty 83,33 %, opět počítáno podle vzorce (4.63). Jazykový model je identický jako v případě obecného systému. Dosažený výsledek demonstруje výhodu vytvoření individuálního modelu z EL nahrávek. Zároveň ukazuje schopnost akustického modelu namodelovat specifika EL řeči. Přestože je výsledek individuálního modelu výrazně lepší, než obecného modelu zpracovávající EL promluvy, tak zdaleka nedosahuje hodnot nejlepších ASR systémů, které jsou schopny v ideálních podmírkách dosahovat více než 90 % slovní přesnosti.

5.3.1 Hledání optimálních parametrů baseline modelu

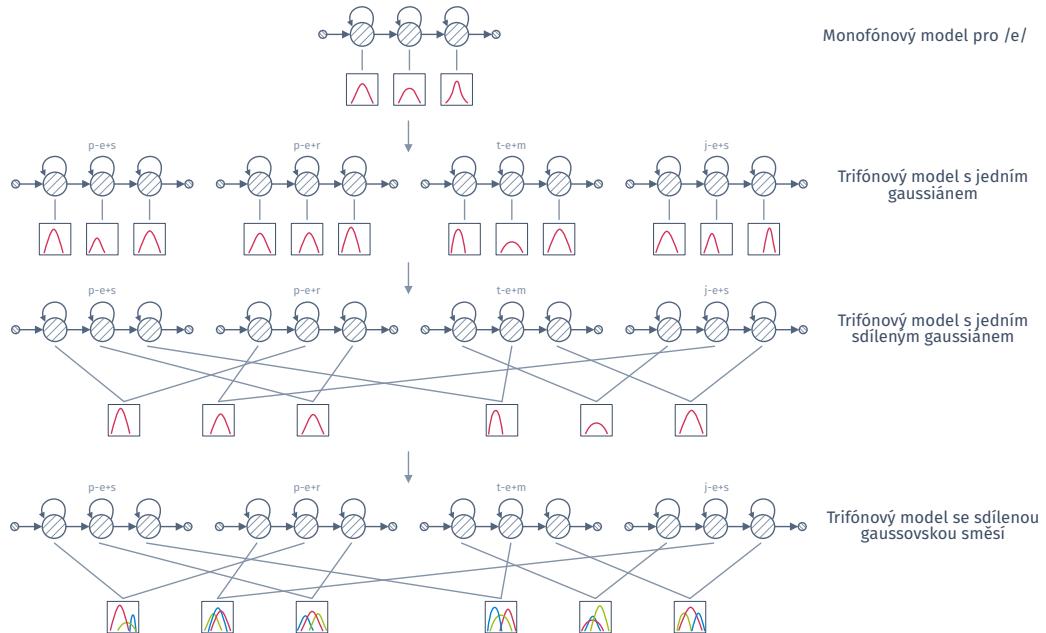
V rámci ověřování funkčnosti individuálního modelu je vhodné zkusit různé varianty k nalezení optimálních parametrů modelu. Hlavními uvažovanými hyperparametry jsou vzorkovací frekvence audio nahrávek a počet *HMM* stavů. Originální pořízené nahrávky mají vzorkovací frekvenci rovnu 44,1 kHz, pro úlohu rozpoznávání EL řeči je to však zbytečně vysoká frekvence, protože nejhodnotnější informace je obsažena u EL řeči ve frekvenčním pásmu do 4 kHz. Vyšší frekvence a priory ovlivňují zabarvení hlasu a další individuální charakteristiky. [20] Samotná EL řeč však obsahuje věci, které běžná řeč neobsahuje a tak je otázka, zde vhodnější vzorkovací frekvence rovna 8 kHz nebo lépe 16 kHz.

Počet stavů modelu ovlivňuje množství modelovaných trifónů, viz 4.2.1. Čím více akustických jednotek je modelováno, tím více musí mít HMM model unikátních stavů. Stinnou stránkou je, že čím více stavů model má, tím více trénovacích dat je potřeba

¹¹Korpus byl náhodně rozdělen na trénovací a testovací sadu v poměru 90 % (10h41m44s) trénovací a 10 % (1h00m58s) testovací sada. Toto rozdělení je použito ve všech experimentech.

k natrénování robustního modelu. Celkem jsou uvažovány modely s 1024, 2048 a 4096 stavy.

Takže při uvažování vzorkovacích frekvencí 8 kHz a 16 kHz bylo natrénováno 6 modelů. K vytvoření akustických modelů je použit HTK-Toolkitu v3.4., který je určen k vytváření *HMM* modelů. Při trénování je nejprve vytvořen monofónový akustický model s jedním gaussiánem pro každý stav. Ten poté slouží jako základ pro trifónové modely. Výsledný trifónový model využívá směs gaussiánů tak, jak je to popsáno v 4.2.2. Celý proces trénování je znázorněn na obr. 5.7.



Obrázek 5.7: Princip trénování HMM-GMM modelu

Snahou je nalezení vhodných parametrů baseline modelu, zejména pak toho akustického. K tomu je potřeba minimalizovat vliv jazykového modelu. Z tohoto důvodu je použit speciální zerogramový jazykový model. Standarně, když se mluví o jazykovém modelu, se předpokládá, že základní jednotkou modelu je slovo. Pro ně jsou počítány četnosti z trénovacích dat a vytvořen model, viz část 4.3. Obecně, ale není nutné, aby základní jednotkou byla slova. V tomto případě je mnohem lepší vytvořit model jehož základní jednotkou je foném, protože ten je výstupem akustického

modelu. Výstupem ASR systému tak bude sekvence fonémů, která z praktického pohledu není úplně užitečná, ale pro testování vlastností akustického modelu se hodí dokonale. V rámci této práce je takovýto model nazýván jako fonémový zerogramový jazykový model. Velikost slovníku tohoto modelu odpovídá velikosti fonémové sady a tady pravděpodobnost libovolného fonému je rovna $P(w_n) = 1/N^{12}$, kde $N = 40$. Výsledná přesnost je tedy a priory závislá pouze na akustickém modelu. Akustická data byla parametrizována pomocí MFCC s 26 filtry, 12 kepstrálními koeficienty a energií. Příznakový vektor obsahuje první i druhou derivaci těchto koeficientů. Více o parametrizaci založené na procesu slyšení v části 4.1.2.

Tab. 5.2 znázorňuje dosažené výsledky HMM-GMM modelů. Opět se potvrdilo, že individuální ASR systém s EL daty může fungovat. Pokud dosažené výsledky porovnáme s výsledky obecného modelu ($Acc_{word} = 18,49\%$), tak i zde je vidět rapidní nárůst přesnosti (78,63 % u nejhoršího individuálního GMM modelu). Ze získaných výsledků je zřejmé, že použití vzorkovací frekvence 16 kHz je vhodnější. Oproti 8 kHz je dosaženo zlepšení přesnosti o 1,41 % absolutně, tedy téměř 7 % relativně. Dodatečné experimenty ukázaly, že použití vyšší vzorkovací frekvence než 16 kHz přinese jen zanedbatelné zlepšení.

Dále se ukázalo, že počet stavů nehraje tak zásadní roli při posuzování kvality akustického modelu, jako vzorkovací frekvence. Z testované množiny počtu stavů dosáhl nejlepšího výsledku model, který měl maximálně 4096 stavů, nicméně oproti modelu s 1024 stavů je nárůst přesnosti pouze 0,4 % absolutně v případě 16 kHz modelů, což není tak významné. Logicky se nabízí otázka, proč nezkusit ještě více stavů? Odpoověď se skrývá ve skutečném počtu stavů modelu s maximálním počtem 4096 stavů. Slovíčko „maximálním“ je zde podstatné. Algoritmus trénování akustického modelu se snaží rozdistribuovat všechny možné akustické jednotky (v tomto případě trifóny)

¹²Označení w_n může evokovat použití slov u jazykového modelu. Změna písmena by však mohla vést ke změně čtenáře, protože by byla použita nestandardní notace. Z tohoto důvodu je i pro speciální model použito označení fonémů w_n .

do maximálního počtu stavů. Pokud chceme méně stavů než jednotek, tak dochází k určité formě shlukování (za pomocí fonetického rozhodovacího stromu) [18]. Pokud je dostatek dat k natrénování konkrétního shluku, je tento shluk akceptován, pokud není dostatečné množství, je tento shluk spojen s jiným, který je svými parametry nejblíže. V případě, že je k dispozici dostatek dat k natrénování maximálního počtu stavů má model tento počet stavů. Pokud není dostatek dat, může mít model méně stavů. U modelu s maximálním počtem 4096 stavů je skutečný počet stavů 3257, tzn. i kdyby se trénoval model s 8192 stavů, tak by se tato hodnota změnila jen velmi málo.

Počet HMM stavů	Acc_p [%]	
	8 kHz	16 kHz
1024	78,63	80,80
2048	79,55	81,09
4096 (3527)	79,79	81,20

Tabulka 5.2: Vliv frekvence na kvalitu modelu.

Hledání optimálních parametrů baseline modelu bylo realizováno na přelomu let 2013 a 2014. V tuto dobu byly stále dominantní GMM modely. Z tohoto důvodu byl později tento experiment zopakován s HMM-DNN akustickým modelem. Vstupem modelu byla stejná MFCC parametrizace s 26 filtry, 12 kepstrálními koeficienty plus energie, delta a delta-delta příznaky. Tato parametrizace je provedena na mikrosegmentu t a jeho okolí $t - 5$ a $t + 5$. Každý mikrosegment má délku 10 ms. Samotná síť se skládá z 6 vrstev, každá s 4096 neurony, výstupní vrstva je typu softmax s dimenzí rovnou počtu HMM stavů. Dosažené výsledky jsou v tab. 5.3. Z nich je patrné, že nalezené optimální hyperparametry jsou shodné i při použití DNN. Nicméně je zde i zřejmý důvod následné dominance HMM-DNN modelů. Fungují totiž výrazně

lépe. Pouhou náhradou GMM za DNN bylo dosaženo zlepšení o 4 % absolutně oproti nejlepšímu GMM výsledku.

Počet HMM stavů	Acc_p [%]	
	8 kHz	16 kHz
1024	77,54	83,98
2048	79,79	84,59
4096	80,42	85,23

Tabulka 5.3: Vliv frekvence na kvalitu modelu využívajícího *DNN*

Hodnoty přesnosti baseline modelu jsou tedy pro GMM $Acc_p^{GMM} = 81,20\%$ a pro DNN $Acc_p^{DNN} = 85,23\%$.

5.3.2 Redukce fonetické sady

Při mluvení je elektrolarynx permanentně zapnutý, a to i v případě neznělých fonémů. Jejich rozdílný průběh je patrný i z analýzy provedené v 5.2. Nabízí se tak předpoklad, že všechny neznělé fonémy mají podobu znělých párových fonémů, a tím pádem je možné redukovat fonetickou sadu. Teoreticky, pokud jsou všechny neznělé fonémy produkovány jako znělé, je redukována fonetická sada a dojde tak ke snížení perplexity modelu. Rozhodnutí, zda se jedná o variantu slova obsahující znělý nebo neznělý foném, je pak přenecháno jazykovému modelu.

K ověření tohoto předpokladu je potřeba experimentálního ověření. Myšlenka experimentu je jednoduchá. Je potřeba natrénovat několik modelů lišících se pouze tím, jaký fonetický pár (viz tab. 5.4) je použit pro redukci fonetické sady. V rámci experimentu jsou uvažovány tyto případy:

- *Baseline* - standardní model s plnou fonetickou sadou.
- $/f/ \rightarrow /v/$ - foném $/f/$ je nahrazen fonémem $/v/$.

- $/k/ \rightarrow /g/$ - foném $/k/$ je nahrazen fonémem $/g/$.
- $/s/ + /š/ \rightarrow /z/ + /ž/$ - foném $/s/ (/š/)$ je nahrazen fonémem $/z/ (/ž/)$.
- $/t/ + /ť/ \rightarrow /d/ + /d'/$ - foném $/t/ (/ť/)$ je nahrazen fonémem $/d/ (/d')$.
- *Náhrada všech* - všechny uvažované neznělé fonémy jsou nahrazeny znělým ekvivalentem.

Neznělé fonémy	Znělé fonémy
$/f/$	$/v/$
$/k/$	$/g/$
$/s/$	$/z/$
$/š/$	$/ž/$
$/t/$	$/d/$
$/ť/$	$/d'/$

Tabulka 5.4: Korespondující páry fonémů.

Pro porovnání jsou stejné modely vytvořeny i pro zdravého řečníka. U něj by při libovolné redukci fonetické sady, mělo dojít ke zhoršení oproti *baseline* modelu.

K natrénování akustických modelů byly použity korpusy čítající 5000 vět¹³, což představuje více než 10 hodin řeči pro každého řečníka. Akustická data byla parametrizována pomocí MFCC s 26 filtry a 12 kepstrálními koeficienty a energií. Dále vektor parametrů obsahuje delta a delta-delta příznaky. To dohromady dává vektor 36 příznaků pro každých 10 ms nahrávky [30].

V rámci experimentu byly otestovány dva přístupy vzájemně se lišící řečovou jednotkou. V prvním případě se jednalo o monofónový akustický model a v druhém trifónový. U obou přístupů je řečová jednotka reprezentována pětistavovým HMM-GMM modelem se spojitou výstupní pravděpodobnostní funkcí pro každý stav, viz

¹³Pro oba řečníky jsou použity stejné věty pocházející z databáze popsané v [29].

4.2.1. Pro určení optimálních parametrů modelu pro EL byly použity znalosti z části **5.3.1**. Pro zdravého řečníka je pro každou část experimentu vytvořeno několik modelů lišících se počtem stavů a gaussovských směsí. Všechny akustické modely jsou natrénovány pomocí HTK-Toolkitu v3.4. Celkem bylo vytvořeno 24 akustických modelů, 12 pro EL řečníka (6 monofónových a 6 trifónových) a 12 pro zdravého řečníka.

Pro otestování modelů byla vytvořena testovací sada čítající 500 vět náhodně vybraných z původních korpusů (pro oba řečníky stejná). Testovací sada tak představuje přibližně 1 hodinu řeči pro každého řečníka. V rámci tohoto experimentu jsou uvažovány dva jazykové modely

1. *zero gramový jazykový model* - v tomto případě mají všechna slova v modelu stejnou pravděpodobnost $P_r(w_n|w_1, \dots, w_{n-1}) = \frac{1}{N}$, kde N je počet slov ve slovníku. Konkrétně $N = 2885$, jinými slovy perplexita modelu je 2885. Testovací slovník je vytvořen z testovací sady, model tedy neobsahuje OOV¹⁴.
2. *trigramový jazykový model* - u tohoto modelu odpovídá pravděpodobnost následujícího slova $P_r(w_n|w_1, \dots, w_{n-1}) = p(w_n|w_{n-2}, w_{n-1})$. K získání $p(w_n|w_{n-2}, w_{n-1})$ posloužil SRILM Toolkit s Kneser-Ney vyhlazováním¹⁵ [31], které se podle [32] ukázalo jako optimální pro tyto typy modelů. Jako trénovací data byly použity texty z novinových článků, webových stránek a přepisů televizních pořadů. Celkem model obsahuje 360K nejvíce frekventovaných slov. OOV bylo 3,8% a perplexita 3380.

V kombinaci s vytvořenými akustickými modely to představuje 4 dílčí experimenty. Jen pro doplnění je nutné poznamenat, že přesnost modelů je vyhodnocována na slovech, oproti fonémovému baseline modelu v **5.3.1**.

¹⁴Out-of-vocabulary (OOV) - slova, která nejsou obsažena ve slovníku jazykového modelu.

¹⁵Vyhazování slouží k vyřešení problému s OOV, kdy trénovací data neobsahovala tato OOV, a proto není k dispozici $p(w_n|w_{n-2}, w_{n-1})$.

Tab. 5.5 a 5.6 zobrazují výsledky pro monofónový akustický model a zero-gramový jazykový model, resp. trigramový jazykový model. V obou případech je vidět očekávané chování přesnosti modelu u zdravého řečníka. Redukcí fonetické sady je omezena komplexita modelu, a tím pádem dochází ke zhoršení přesnosti. Překvapující mohou být horší výsledky u zdravého řečníka v tab. 5.6. Toto chování může být vysvětleno vyšší perplexitou trigramového jazykového modelu v kombinaci s relativně jednoduchým monofónovým akustickým modelem.

U EL řečníka je vidět dílčí zlepšení u 2 modelů (tab. 5.5), resp. 1 modelu v případě trigramového modelu (tab. 5.6). Ve většině případů však redukce fonetické sady vedla ke zhoršení přesnosti. Při použití trigramového jazykového modelu došlo obecně ke zlepšení výkonu systému i přesto, že tento model má vyšší perplexitu. To nasvědčuje tomu, že monofónový akustický model je příliš jednoduchý pro tuto úlohu.

Model	$Acc_w [\%]$	
	Zdravý	EL
Baseline	91,35	83,05
/f/ → /v/	89,96	83,05
/k/ → /g/	90,68	83,10
/s/+/š/ → /z/+/ž/	88,77	83,71
/t/+/ť/ → /d/+/ď/	90,05	82,47
Náhrada všech	86,58	82,78

Tabulka 5.5: Vliv redukce fonetické sady na přesnost ASR systému s monofoním akustickým a zero-gramovým jazykovým modelem ($N = 2885$) pro zdravého a EL řečníka.

V tab. 5.7 a 5.8 jsou pak vypsány výsledky pro trifónový akustický model se zero-gramovým resp. trigramovým jazykovým modelem. Stejně jako u předchozích dvou experimentů, tak i zde je vidět, že redukce fonetické sady vede u zdravého řečníka vždy ke zhoršení přesnosti modelu. Také je tu možné vydedukovat, že trifónový

Model	Acc_w [%]	
	Zdravý	EL
Baseline	87,47	84,92
/f/ → /v/	87,42	84,51
/k/ → /g/	86,36	85,50
/s/+/š/ → /z/+/ž/	84,81	84,75
/t/+/ť/ → /d/+/ď/	86,38	84,38
Náhrada všech	83,77	84,34

Tabulka 5.6: Vliv redukce fonetické sady na přesnost ASR systému s monofoním akustickým a trigramovým jazykovým modelem obsahujícím 360k slov pro zdravého a EL řečníka.

akustický model dosahuje výrazně lepších výsledků než monofoní model. Zhoršení u EL řečníka v tab. 5.7 je s největší pravděpodobností způsobeno fonetickými stromy, protože není dostatek dat pro všechny možné varianty trifónů. Tím pádem model pro určité trifóny vrací špatné sekvence znaků. Zerogramový jazykový model pak nedokáže pomoci, protože všechna slova mají stejnou pravděpodobnost $P_r(w_n|w_1, \dots, w_{n-1}) = \frac{1}{2885}$. Tím pádem dochází k rozpoznávání špatného slova a nižší celkové přesnosti. Tuto domněnkou potvrzuje rapidní zlepšení v případě trigramového jazykového modelu (tab. 5.8), kde již jazykový model významně přispívá k přesnosti modelu.

U obou experimentů s trifónovým akustickým modelem došlo ke zlepšení u dvou modelů (tab. 5.7 a 5.8), ale stejně jako v případě monofonového modelu vedla ve většině případů redukce fonetické sady ke zhoršení.

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 pokud se nahradí např. dvojice /s/ a /š/ tak, že za každých okolností to povede k lepším výsledkům. Při hlubší analýze se ukázalo, že velmi záleží na kontextu daného

Model	Acc_w [%]	
	Zdravý	EL
Baseline	92,66	82,60
/f/ → /v/	92,41	82,23
/k/ → /g/	92,57	83,30
/s/+/š/ → /z/+/ž/	92,28	83,28
/t/+/ť/ → /d/+/ď/	92,28	82,13
Náhrada všech	91,03	82,18

Tabulka 5.7: Vliv redukce fonetické sady na přesnost ASR systému s trifónovým akustickým a zerogramovým jazykovým modelem pro zdravého a EL řečníka.

Model	Acc_w [%]	
	Zdravý	EL
Baseline	95,80	87,65
/f/ → /v/	95,46	87,51
/k/ → /g/	95,55	88,38
/s/+/š/ → /z/+/ž/	95,07	88,31
/t/+/ť/ → /d/+/ď/	95,39	87,60
Náhrada všech	94,53	86,97

Tabulka 5.8: Vliv redukce fonetické sady na přesnost ASR systému s trifónovým akustickým a trigramovým jazykovým modelem s 360k slov pro zdravého a EL řečníka.

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 se u nahradby /s/ za /z/ vynechal trifón *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. Přestože se jedná o marginální zlepšení, tak ho bylo docíleno jedním trifónem. Bohužel určení toho, jaké trifóny vynechat z nahrazování není vůbec triviální úloha.

Zajímavý je také rozdíl mezi přesností modelu pro zdravého a EL řečníka. Přestože se v obou případech jedná o individuální modely šité „na míru“ řečníkovi, tak průměrný rozdíl je 6,24 % absolutně a 40,38 % relativně. To značí, že je potřeba se zabývat myšlenkou úpravy akustického modelu, aby dosahoval lepších výsledků a v ideálním případě podobných výkonů jako modely pro zdravé řečníky.

Naopak očekávaným výsledkem bylo zhoršená přesnosti u zdravého řečníka ve všech případech redukce fonetické sady. Dále se potvrdilo, že komplexnější trifónový model dosahuje ve většině případů lepších výsledků. To je nepochybně způsobeno faktem, že každý foném je modelován pomocí více HMM stavů, protože se bere v potaz i jeho okolí, kdežto u monofónového modelu nikoli.

Kapitola 6

Návrh a realizace úprav ASR

Experimenty provedené v části 5.3 jasně ukázaly, že individuální ASR modely relativně obstojně zvládají rozpoznávat EL řeč. Individuální a obecné modely pro zdravého řečníka však dosahují významně lepších výsledků. Z experimentů s redukcí fonetické sady (viz část 5.3.2) vyvstala potřeba rozšířit řečový korpus o příklady promluv obsahující slova mající rozdílný význam, ale lišící se pouze ve znělosti jednoho fonému.

Toto rozšíření totiž umožní lepší porozumění problematice znělosti EL řeči a může s návrhem úprav akustického modelu tak, aby výsledná přesnost se co možná nejvíce maximalizovala.

6.1 Doplnění řečového korpusu o specifická data - vliv nových dat na kvalitu AM

Před samotným pořízením nahrávek promluv je nezbytné vybrat co možná nejvíce dvojic slov, které se liší významem a znělostí právě jednoho fonému. Příkladem takovýchto slov může být dvojice slov *kosa + koza* nebo *přibít + připít*. Pro tento účel je použit algoritmus výběru slov, který je následující:

1. načtení dat (slovník a hledané párové fonémy);

2. shluknutí všech slov vedoucích ke stejné fonetické transkripcii;
3. vytvoření všech možných kombinací dvojcí slovních transkripcí;
4. nalezení dvojcí transkripcí, které se liší právě ve znělosti jednoho fonému¹;
5. výběr dvojcí slov na základě vybraných fonetických transkripcí;

Vstupem algoritmu je „slovník“ obsahující slova a jejich fonetický přepis, dále pak dvojice fonémů (znělý + neznělý). Jako slovník posloužil seznam slov s fonetickými přepisy, které pocházejí z jazykového modelu obsahujícího 1,2 milionu slov. Pomocí výše zmíněného algoritmu se podařilo nalézt 160 párů slov lišících se znělostí právě jednoho fonému, celkem tedy 320 slov. Ke každému nalezenému slovu se následně vybrala minimálně jedna věta obsahující toto slovo (ale nikoli druhé slovo z dvojice). Těchto vět je pak 418. Příklad vybraných vět je uveden níže:

Zkoušel jsem to několikrát, ale pokaždé padla kosa na kámen.

Do basy nemusí, vlk žere, koza žije.

Vybraná slova a věty se staly základem pro 2. etapu nahrávání. Ta se uskutečnila během dvou sezení v červenci roku 2016. Jde tedy o relativně velký časový odstup od 1. etapy. Nahrávání se zhostil stejný řečník jako v případě 1. etapy (viz část 5.1). Jednotlivá nahrávací sezení měla mezi sebou týdenní rozestup. Oproti 1. etapě probíhalo nahrávání v odhlučněné nahrávací komoře a za pomocí profesionálního nahrávacího zařízení. Mikrofon byl od úst řečníka vzdálen přibližně 15 cm, protože byl použit studiový mikrofon. K nahrávání byl použit speciální software, který kontroloval, zda každá nahrávka splňuje určité parametry. Každá akceptovaná nahrávka musí mít na svém začátku a konci minimálně 0,5 s ticha a zároveň celá nahrávka nesmí být příliš tichá a zároveň přebuzená (kontrolováno pomocí energie). Pokud nahrávka nesplňuje definované parametry, je zamítnuta a řečník musí promluvu zopakovat.

¹Konkrétně algoritmus vzájemně porovná obě slova a najde rozdílné fonemy. Pokud tyto rozdíly odpovídají některé z dvojcí párových fonémů, tak je dvojice přijata.

V části 5.1 je zmíněno, že je nezbytné provádět anotaci nahrávek, aby mohl být korpus kompletní. Samotná anotace je relativně zdlouhavý proces, a proto je dobré pořídit přesné promluvy vybraných slov a vět již v průběhu nahrávání. K tomu slouží další z funkcí nahrávacího softwaru, který řečníkovi vždy ukáže text, který je potřeba vyslovit. Společně s audio záznamem je pak uložen i tento text. K dispozici je tedy nahrávka a její „přepis“. Nicméně samotný řečník často může udělat chybu aniž by si toho všiml (např. záměnou podobných slov apod.). Software nijak nekontroluje co je ve skutečnosti vysloveno. Z tohoto důvodu je nahrávání přítomen operátor, který poslouchá co bylo řečeno a v případě potřeby zamítne nahrávku. Řečník následně musí promluvu opakovat, dokud nahrávka neodpovídá požadovaným parametry a zároveň je její obsah správný.

Na obr. 6.1 a 6.2 jsou ukázky audio záznamu slova „kosa“ a věty „Zkoušel jsem to několikrát, ale pokaždé padla kosa na kámen.“. Pokud se nahrávky porovnají s daty získanými v 1. etapě (obr. 5.1), tak hlavním rozdílem je vyšší kvalita nahrávek, zejména vyšší amplituda. Ze zobrazených spektrogramů je zřejmé, že šum je přítomen v menším množství a intenzitě než v předchozích nahrávkách. Hlavní vliv na tom má nový mikrofon, který není přilepen ke tváři řečníka a nezaznamenává, tak vibrace přenášené měkkou tkání. Další rozdíl je vidět v nižších frekvencích spektrogramu, ty jsou výraznější. Přestože se jedná o stejného řečníka, tak zaznamenaná řeč nemá úplně identické parametry. Jedním z důvodů bude nepochybně změna nahrávací aparatury a procesu nahrávání. Nezanedbatelný vliv má i relativní nestálost parametrů EL řeči, zvlášť v delším časovém období. Ty jsou totiž velmi závislé na typu a pozici elektrolarynxu, ten se v době mezi nahrávánimi navíc změnil, což v konečném důsledku představuje asi hlavní důvod diferenčních parametrů.

Tab. 6.1 přibližuje souhrnné parametry nahrávek pořízených v 2. etapě nahrávání. Celkem se podařilo získat přibližně 2 hodiny řeči (každá nahrávka obsahuje 0,5 s ticha na začátku a konci). Z toho přibližně jen 10 % představují vybraná izolovaná slova.



Obrázek 6.1: Průběh a spektrogram slova „kosa“ s společně s vyznačenou energií EL promluvy.



Obrázek 6.2: Průběh a spektrogram promluvy obsahující slovo „kosa“ s vyznačenou energií EL promluvy.

Dohromady s novými daty obsahuje korpus téměř 14 hodin audio záznamů a k nim příslušných přepisů.

Nahrávání	Délka /HH:MM:SS/	Počet slov	Počet vět	Počet souborů
2016.07 - 2016.07	2:13:56	320	420	740

Tabulka 6.1: Informace o korpusu nahrávek z 2. etapy nahrávání.

6.1.1 Vliv nových dat na kvalitu modelů

Rozšíření korpusu umožňuje vytvoření nových modelů, které ověří konzistenci a vliv nových dat na přesnost. Oproti baseline modelu v části 5.3.1 jsou všechny následující modely vytvořeny ve frameworku Kaldi. Ten se po roce 2015 stal standardem pro vytváření akustických modelů, protože je velmi flexibilní a umožňuje snadné přidávání nových typů akustických modelů. [33] Tento framework je uvolněn jako open-source.

Již při vytváření baseline modelu (viz část 5.3.1) se ukázala lepší funkce DNN modelů. Přestože vývoj výpočetních GPU postupuje závratnou rychlostí, tak natréno-vání HMM-DNN modelu je časově náročnější než vytvoření HMM-GMM akustického modelu. Navíc, jak bylo popsáno v 4.2.3, k natrénování DNN modelu je potřeba za-rovnání získané pomocí HMM-GMM modelu. Proto je vhodné prvotní validaci nových dat provést na jednodušším modelu.

Proces vytvoření akustického modelu vycházejí z předpřipravených Kaldi trénova-cích skriptů pro vytvoření modelu pomocí Wall Street Journal korpusu. Tyto skripty jsou jen drobně upraveny, aby výsledný model mohl být natrénován z EL korpusu. Data jsou parametrizována pomocí Perceptual Linear Prediction (PLP) s 12 kepstrál-

ními, delta a delta-delta koeficienty². Nejprve je vytvořen monofónový model, který slouží jako iniciační model pro trifónové modely, viz obr. 5.7.

V části 5.3 bylo popsáno rozdelení korpusu na trénovací a testovací sadu. Po rozšíření korpusu je rozdelení dat z 1. etapy ponecháno a nová data k jednotlivým sadám přidána. Všechny nahrané věty v 2. etapě jsou přidány do trénovací a všechna slova naopak do testovací sady. Toto rozdelení vychází z impulzu pro rozšíření korpusu o specifická slova. Ta tedy a priory mají sloužit k otestování nových modelů, a tím lépe porozumět problematice znělosti u EL řeči.

Jazykový model je opět fonémový zerogramový a na kompletní testovací sadě byla dosažena přesnost $Acc_p^{GMM} = 54,96\%$ ³. V případě, že testovací sada obsahuje pouze nově nahraná slova, tak dokonce jen $Acc_p^{GMM} = 42,97\%$ ⁴. Což je významné zhoršení oproti výsledkům dosažených u baseline modelu v 5.3.1 ($Acc_p^{GMM} = 81,20\%$). Výpočet přesnosti je opět realizován vztahem (4.63).

Jelikož došlo ke změně ASR frameworku je potřeba ověřit, že nevznikla chyba při vytváření akustického modelu. K ověření je použit křížový test, kdy pomocí stejného procesu jsou natrénovány modely z původních (1. etapa) a nových (2. etapa) dat a křížově otestovány na kompletní, původní a jen nové části testovací sady. Trénovaný akustický model má stejně parametry jako v předchozím případě. Vstupem jsou PLP data s 12 kepstrálními, delta a delta-delta koeficienty. Výsledný model může mít až 4096 stavů. Výsledky testu jsou v tab. 6.2. Z té je jasné patrné, že trénovací proces je „správný“. Problém je tedy v datech.

²V rámci úprav Kaldi skriptů se PLP parametrizace ukázala jako vhodnější pro EL řeč. Ověření proběhlo experimentálně, kdy byly vytvořeny dva identické HMM-GMM modely s 4096 stavů, ale každý byl natrénován na jinak parametrizovaných datech (MFCC a PLP). PLP model dosáhl o 1,31 % absolutně vyšší přesnost.

³Celková délka nahrávek v testovací sadě složené z vět a slov činila 1h16m42s.

⁴Celková délka nahrávek v testovací sadě složené pouze slov činila 15m44s.

Model	$Acc_p [\%]$	
	1. etapa	2. etapa
1. etapa	76,64	19,24
2. etapa	15,63	82,97

Tabulka 6.2: Křížový test modelů natrénovaných a otestovaných na datech z 1. a 2. etapy.

6.1.2 Eliminace vlivu kanálu

Z prezentovaných výsledků plyne, že nová data jsou příliš odlišná od původních a v parametrickém prostoru příliš vzdálena těm původním. Zároveň je těchto dat relativně malé množství, aby se mohly modely plně adaptovat. Na zmíněný rozdíl v datech je možné nahlížet jako na změnu kanálu, která je příčinou těchto změn. Řečník je totiž stejný. V předchozí části 6.1.1 bylo zmíněno, že v rámci 2. etapy došlo ke změně nahrávací procedury a elektrolarynxu. Tím byl pozměněn kanál a logicky výsledná zaznamenaná řeč má jiné parametry než ta původní z 1. etapy. Mezi další prvky, které mohou způsobit změnu kanálu je prostředí, tedy zda je řeč produkována uvnitř nějaké místnosti, či venku, jestli je na pozadí přítomen šum atd.

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, viz dále.

Zaznamenaný signál je možné popsat jako konvoluci promluvy a vlivu kanálu, matematicky zapsáno jako

$$y[n] = x[n] \otimes h[n], \quad (6.1)$$

kde $x[n]$ představuje vstupní signál, tedy řeč, a $h[n]$ odezvu kanálu na jednotkový impulz. Zaznamenaný signál je jejich již zmíněnou lineární konvolucí. Ve frekvenční oblasti je pak rovnice (6.1) zapsaná následovně

$$Y[f] = X[f] \cdot H[f]. \quad (6.2)$$

K převodu slouží FFT. Ve frekvenční oblasti se z konvoluce stává násobení. Dalším krokem je převedení hodnot do kepstrální oblasti. To je realizováno pomocí logaritmu spektra, stejně jako v případě MFCC parametrizace, viz 4.1.2. V kepstrální oblasti má vzorec (6.1) následující podobu

$$Y[q] = \log(Y[f]) = \log(X[f] \cdot H[f]) = X[q] + H[q], \quad (6.3)$$

kde q představuje kepstrální koeficient. V kepstrální oblasti je vliv kanálu aditivní složkou výsledného záznamu. Problémem však je, že konkrétní hodnota vlivu kanálu je neznáma. K dispozici je pouze výsledný ovlivněný signál. Předpokládejme však, že vliv kanálu je stacionární⁵, tak poté je možné každý frame nahrávky i zapsat jako

$$Y_i[q] = H[q] + X_i[q], \quad (6.4)$$

kde $Y_i[q]$ představuje i frame kepstra q nahrávky a $X_i[q]$ představuje i frame kepstra q neovlivněné řeči. Z této rovnice je pak možné vypočítat střední hodnotu

$$\frac{1}{N} \sum_i Y_i[q] = H[q] + \frac{1}{N} \sum_i X_i[q]. \quad (6.5)$$

Vliv kanálu je následně možné eliminovat odečtením této střední hodnoty kepstra q od aktuální hodnoty kepstra $Y_i[q]$

⁵Jedná se sice o silný, ale logický předpoklad. Pokud se vztáhne k pořízenému řečovému korpusu, tak v rámci jedné etapy nahrávání, je proces nahrávání neměnný, tzn. je použita stejná aparatura a k nahrávání dochází vždy ve stejné místnosti.

$$\begin{aligned}
R_i[q] &= Y_i[q] - \frac{1}{N} \sum_j Y_j[q] \\
&= H[q] + X_i[q] - \left(H[q] + \frac{1}{N} \sum_j X_j[q] \right) \\
&= X_i[q] - \frac{1}{N} \sum_j X_j[q]
\end{aligned} \tag{6.6}$$

S pomocí rovnice (6.6) je možné odfiltrovat vliv kanálu a teoreticky tak získat nezkreslený signál. Otázkou je, přes jaký úsek počítat střední hodnotu. Je možné ji počítat přes posuvné okénko fixní délky, přes jednotlivé věty/nahrávky, nebo dokonce přes všechny nahrávky konkrétní etapy. Výběr optimálního úseku je realizován na základě experimentu.

Určení délky úseky pro výpočet CMN

K určení vhodné délky úseku pro výpočet CMN je použita stejná trénovací procedura jako v předešlých experimentech s novými daty. Je tedy trénován HMM-GMM model s maximálně 4096 stavy. Vstupní data jsou parametrizována pomocí PLP. Na ně je aplikováno CMN počítáno z různé délky úseku. Celkem jsou uvažovány dva experimenty, a to

- *CMN* počítáno pro každou nahrávku,
- *CMN* počítáno pro celou etapu.

V tab. 6.3 jsou výsledky experimentu s *CMN* počítaném přes jednotlivé nahrávky. Z dosažených výsledků je zřejmé, že je oproti výsledkům v tab. 6.2 dosaženo určitého zlepšení, zvláště pokud je model natrénován na datech z 1. etapy a otestován na datech z 2. etapy. Výsledky však nejsou zdaleka tak dobré, jako v případě trénování a testování modelu na datech ze stejné sady. Vliv tu hráje fakt, že zejména nahrávky

izolovaných slov jsou relativně krátké a vypočtené střední hodnoty, tak nabývají odlišných hodnot.

Model	Acc_p [%]	
	1. etapa	2. etapa
1. etapa	76,63	46,50
2. etapa	27,43	82,72

Tabulka 6.3: Křížový test modelů natrénovaných a otestovaných na datech z 1. a 2. etapy s CMN přes jednotlivé věty.

Další experiment je s daty, kde bylo aplikováno CMN vypočtené ze všech nahrávek konkrétní etapy. V tab. 6.4 je vidět markantní zlepšení výsledků. Pokud je model natrénován na datech z 1. etapy a otestován daty z libovolné etapy, jsou dosažené výsledky velmi podobné. Nejhoršího výsledku je dosaženo, pokud je model natrénován na datech z 2. etapy a otestován na těch z 1. V tomto případě hraje velký vliv, relativně malé množství dat (pouhé 2 hodiny). Pokud je tedy CMN počítáno přes všechny nahrávky v dané etapě, je dosaženo významného zlepšení a vliv kanálu je v podstatě eliminován. Pro doplnění je dobré zmínit, že pokud je model natrénován ze všech trénovacích dat (1. a 2. etapa) a otestován pomocí kompletní testovací sady, tak je dosažená přesnost s fonémovým zerogramovým jazykovým modelem rovna $Acc_p = 77,69\%$.

Model	Acc_p [%]	
	1. etapa	2. etapa
1. etapa	77,84	75,92
2. etapa	60,39	82,64

Tabulka 6.4: Křížový test modelů natrénovaných a otestovaných na datech z 1. a 2. etapy s CMN přes všechny nahrávky v etapě.

Z výsledků v tab. 6.3 je možné odvodit, že pokud by se *CMN* počítalo přes posuvné okénko fixní délky, tak dosažené výsledky by nebyly zrovna dobré. Což se i experimentálně potvrdilo, když výsledná přesnost dosáhla hodnoty $Acc_p = 56,51\%$ na kompletní trénovací a testovací sadě. Samotný framework Kaldi umožňuje aplikování CMVN, což je Cepstral mean and variance normalization. Jedná se o variaci rovnice (6.6), kde je kromě střední hodnoty počítána i variance. Kaldi CMVN je počítáno přes okénko fixní délky a výsledná přesnost HMM-GMM modelu s CMVN dosáhla hodnoty $Acc_p = 76,15\%$ na kompletní trénovací a testovací sadě. Tento výsledek je srovnatelný s modelem využívajícím *CMN* přes všechny nahrávky v dané etapě.

Výsledky modelů po eliminaci vlivu kanálu

Aplikací *CMN* dosáhl HMM-GMM model srovnatelných výsledků jako v části 5.3.1. Dalším krokem je tedy natrénování HMM-DNN modelu. Trénovaná neuronová síť FF síť má 5 skrytých vrstev, výstupní vrstva je typu softmax s dimenzí rovnou počtu *HMM* stavů. Postupně je natrénována síť s 1024, 2048 a 4096 neurony v každé skryté vrstvě. Vstupní data jsou parametrizována pomocí PLP s 12 kepstrálními, delta a delta-delta koeficienty a CMN počítané ze všech nahrávek dané etapy. Jazykový model je fonémový zerogramový tak, aby byl co nejvíce amplifikován vliv akustického modelu. Tab. 6.5 zobrazuje dosažené výsledky všech natrénovaných variant. Nejvyšší přesnosti dosahuje model s 4096 neurony v každé vrstvě, ale rozdíl od ostatních variant není velký. 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 6.5: Dosažená přesnost neuronové sítě s monofónovým zerogramovým jazykovým modelem.

6.2 Poslechový test a porovnání výsledků člověka a stroje

V předchozím textu byly prezentovány různé dosažené výsledky, ale ty zatím nedokázali odpovědět na zásadní otázku: „Dokáže se stroj⁶ vyrovnat človéku?“. Přestože je EL řeč na první poslech obtížně srozumitelná, tak již po krátké době je člověk schopen obstojně rozumět. S přibývajícím časem se do určité míry porozumění ještě zlepšuje. Jak je na tom tedy stroj v porovnání s člověkem?

Ještě než je vůbec možné na tuto otázku odpovědět, tak je dobré si odpovědět na otázku: „Jakým způsobem porovnat schopnosti člověka a stroje?“. K tomu může posloužit poslechový test, ve kterém mají posluchači za úkol vybrat, z předem definovaných možností, co je obsahem promluv. Otestování schopností stroje pak probíhá standardně pomocí experimentu. Vstupem ASR systému jsou stejné promluvy, která jsou součástí poslechového testu. Výstupem je přepis. Metrika experimentu je počítána na základě správně/špatně určeného obsahu promluv v přepisu⁷. Prostým porovnáním počtu správných odpovědí člověka a stroje je možné odpovědět na první „položenou“ otázku.

Při přípravě experimentu vykristalizovaly tyto varianty poslechového testu:

⁶Stroj je zde reprezentován systémem automatického rozpoznávání řeči.

⁷Výstup ASR systému je považován za správný i v případě, že se liší např. i/y. Z akustického pohledu jsou totiž oba fonémy identické.

- test na izolovaných slovech,
- test na slovních bigramech.

Tím, že promluvy obsahují pouze izolovaná slova (v druhém případě dvojici slov) je do značné míry eliminován vliv kontextu. Ten v mnoha případech pomáhá se správným určením významu i přesto, že nebylo dobře rozumět. Pokud se bude experiment skládat z množiny promluv, které obsahují pouze slova popsaná v části 6.1, tak bude možné určit, jak „dobře“ dokáže člověk (stroj) určit význam těchto slov a případně je od sebe odlišit.

6.2.1 Izolovaná slova

Rozpoznání slova, které bylo vysloveno v klidném prostředí se jeví jako velice jednoduchý úkol, ale pokud jej vyslovil řečník používající EL, tak už to tak snadné být nemusí. Zvlášť pokud se jedná o slova popsaná v 6.1. 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.

Výstupem poslechového testu je tabulka s procentuálním zastoupením jednotlivých odpovědí pro každou nahrávku. V tab. 6.6 je zobrazen výňatek získaných vý-

sledků. Správné odpovědi jsou zvýrazněny tučně. Výsledky slov *borce* a *porce* reprezentují situaci, kdy účastník nebyl jednoznačně schopen určit význam slova. Druhý příklad (*kosa* + *koza*) ukazuje situaci, kdy všichni účastníci vybrali z komplementárních slov vždy pouze jediné, a to nehledě na to, které jim bylo ve skutečnosti puštěno. V tomto konkrétním případě tedy posluchači vždy „slyšeli“ slovo „koza“. Dalo by se tedy usuzovat, že slova „kosa“ je akusticky identické se slovem „koza“. Poslední případ reprezentuje situaci, kdy většina účastníků byla schopna určit správný význam slova. Celková přesnost dosáhla hodnoty $Acc_w^{human} = 70,47\%$ a byla vypočtena vzorcem

$$Acc_w^{human} = \frac{1}{N} \sum_{i=1}^N f_i * 100, \quad (6.7)$$

kde $N = 320$ a f_i se rovná relativní četnosti správných odpovědí na otázku i v poslechovém testu s izolovanými slovy.

Slovo	Relativní četnost odpovědí [%]		
	<i>a)</i>	<i>b)</i>	<i>Nevím</i>
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 6.6: Ukázka výsledku poslechového testu na izolovaných slovech.

6.2.2 Slovní bigramy

V druhém poslechovém testu mají posluchači za úkol vyslechnout 333 nahrávek slovních bigramů⁸ a vybrat jednu z předem definovaných odpovědí. Ty mají vždy tento formát

- 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. Rozšířený řečový korpus, tak jak je popsáný v části 6.1, ale neobsahuje tento typ nahrávek. Tím pádem je potřeba je vytvořit „uměle“. Což není velký problém, každá nahrávka izolovaného slova obsahuje minimálně 0,5 s ticha na svém začátku a konci. Pokud jsou tyto nahrávky spojeny⁹, 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 slovem¹⁰. 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 u testu s izolovanými slovy je výstupem testu tabulka obsahující procentuální zastoupení jednotlivých odpovědí na každou otázku. Tab. 6.7 obsahuje

⁸Nahrávky obsahují dvě po sobě vyslovená slova.

⁹Ke spojení je možné použít nástroj *ffmpeg* nebo *sox*.

¹⁰Ve 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*.

ukázku těchto výsledků. Stejně jako v předchozím případě, správné odpovědi jsou zvýrazněny tučně. Ačkoli jsou nyní v testu dvojice slov, tak dosažené výsledky do značné míry korespondují s výsledky z testu s izolovanými slovy (viz tab. 6.6). V prvním případě (*borci + porci*) nebyli účastníci schopni jednoznačně určit význam slov v nahrávce. V druhém případě (*kosa + koza*) všichni posluchači až na jednoho vybrali možnost *d*), tedy *koza + koza*. Správný výběr jedním účastníkem lze považovat spíše za náhodu, protože u opačného pořadí slov již nikdo správnou odpověď nevybral. Je dobré zmínit, že účastníci v žádném poslechovém testu nebyli omezeni v počtu opětovného přehrání promluvy. Tím pádem je velmi pravděpodobné, že tento konkrétní participant opakovaně poslouchal danou nahrávku a hledal rozdíl až nějaký drobný zaznamenal. Otázkou však je, jestli to spíše nebyla sugesce a to již zmíněné štěstí.

Poslední prezentovaný příklad zastupuje množinu odpovědí, kdy účastníci naprostě správně určili význam slov. Průměrná dosažená přesnost člověka, počítána pomocí rovnice (6.7), dosáhla hodnoty $Acc_w^{human} = 66,24\%$.

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

Tabulka 6.7: Ukázka výsledku poslechového testu na dvojicích slov.

6.2.3 Výsledky porovnání

Výsledky poslechového testu ukázaly, jak je na tom člověk. Nyní je potřeba zjistit, jak je na tom stroj zastoupený ASR systémem. K tomu je nezbytné natrénovat akustický model a použít vhodný jazykový model. Jako akustický model je použit HMM-DNN model. Konkrétně jde o DNN síť s 6 vrstvami (5 skrytých vrstev, každá s 4096 neuronů), výstupní vrstva je typu softmax s dimenzí rovnou počtu HMM stavů. Jako parametrizace je použito PLP (12 kepstrálních koeficientů + delta + delta-delta parametry) a pro eliminaci vlivu kanálu *CMN* počítané přes všechny nahrávky v rámci etapy. Výsledný příznakový vektor má dimenzi 36. Vstupem neuronové sítě je pak parametrisované okénko mající kontext přes 11 mikrosegmentů, tedy $t - 5$ a $t + 5$. Vstupní dimenze neuronové sítě je tedy 396. Oproti dosavadním experimentům je v tomto případě použit vlastní real-time dekodér (více o dekódování v části 4.4). Tento LVCSR systém je optimalizován pro co nejnižší latenci a je schopný pracovat s velmi velkými slovníky čítající i miliony položek. Tento dekodér byl vyvinut na katedře kybernetiky Fakulty aplikovaných věd.

Z poslechových testů jsou k dispozici dva výsledky. První reprezentuje schopnost určit význam izolovaného slova. Druhý schopnost rozeznat dvě velmi podobná slova. Pro potřeby porovnání, s těmi dosaženými ASR systémem, jsou vytvořeny celkem 3 experimenty využívající výše popsány akustický model.

První experiment odpovídá poslechovému testu s izolovanými slovy a jeho základem je zero-gramový LM obsahující více než 1 milion slov. Většina předchozích experimentů využívala fonémový zero-gramový model, aby bylo možné eliminovat vliv LM na výsledné přesnosti. U těchto experimentů je tento LM nevhodující, protože cílem je správně určit celé slovo, a proto je využit slovní jazykový model. U zero-gramového modelu mají 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 ti-

tulků a přepisů televizních pořadů. Využití takto velkého LM vychází z představy, že i člověk má velkou slovní zásobu a dopředu neví co bude obsahem konkrétní promluvy v rámci testu. Tento test je pojmenován jako „one-mil“.

Ve skutečnosti však, v rámci poslechového testu, účastníci znají seznam slov zahrnutých v testu a mají tak určitou výhodu oproti „one-mil“ nastavení. Ke kompenzaci tohoto faktu, je vytvořen druhý experiment, který má redukovaný LM. Obsahuje pouze slova, která se opravdu vyskytla v rámci poslechového testu ($N = 320$). Tento experiment je nazván jako „reduced“. Výsledky obou těchto experimentů jsou porovnány s poslechovým testem na izolovaných slovech. Další možností by mohlo být vytvoření speciálních LM pro každou promluvu. Ten by obsahoval pouze kompletní slova. Problémem tohoto postupu je 3 možnost, kterou obsahuje poslechový test (tzv. „nemohu rozhodnout“). V případě, že by LM obsahoval pouze dvě slova, tak by ASR experiment, svými parametry, neodpovídal poslechovému testu.

Poslední experiment odpovídá druhému poslechovému testu. K získání srovnatelných výsledků, je ke každé nahrávce se slovním bigramem vygenerován speciální zerogramový LM. Ten obsahuje vždy pouze všechny 4 kombinace slov. Tím pádem odpovídá dostupným možnostem v rámci poslechového testu. Tento experiment je nazván jako „bigrams“. Jeho výsledky jsou porovnány s druhým poslechovým testem.

Výsledkem rozpoznávače je nejlepší hypotéza (případně N nejlepších hypotéz), tudíž slovo. To však není porovnatelné s výsledkem poslechového testu. Z tohoto důvodu jsou všechny výsledky ohodnoceny 1, pokud bylo výstupem správné slovo, a v opačném případě 0. Následně byl z tohoto ohodnocení vypočten průměr. Pro upřesnění je nutné zmínit, že i/y na výsledném ohodnocení nehraje roli. Dosažené výsledky jsou pak v tab. 6.8. Ty ukazují, že požadovaný úkol je výzvou jak pro člověka, natož pro stroj. V případě experimentu „one-mil“ je výkon ASR systému významně horší než výkon člověka. To je zejména způsobeno enormní perplexitou jazykového modelu. Ta je přímo rovna velikosti slovníku. Zmenšením slovníku se

podařilo získat výsledky srovnatelné s člověkem. Je dobré zdůraznit, že i v případě „reduced“ experimentu hrají karty ve prospěch člověka, protože čelí pouze perplexitě 3, protože se kdykoli může podívat na nabízené možnosti. Řešením by bylo nechat účastníky přepsat obsah promluvy a porovnat ho se skutečným obsahem. Nicméně toto by významně zvýšilo náročnost (zejména časovou) poslechového testu a bylo by velmi komplikané získat kompletní výsledky od relevantního množství účastníků. Už jen podíl odpadlíků mezi prvním a druhým poslechovým testem činí závratných 30 %.

Velmi zajímavé jsou výsledky u „bigrams“ experimentu. Na první pohled se může jevit jako snazší, protože úkolem je vybrat z jasně definovaných kombinací slov. Avšak slova jsou si akusticky velmi podobná a v mnoha případech je velmi náročné je od sebe rozeknat. Jak člověk, tak stroj, dosáhli v tomto testu nejhorších výsledků. Při analýze se ukázalo, že rozdíly mezi hypotézami ASR systému jsou velmi malé, což naznačuje velkou podobnost mezi inkriminovanými modely fonémů. Zároveň tyto výsledky s velmi podobnými hypotézami korelují s výsledky poslechového testu, kde posluchači nebyli schopni jednoznačně rozhodnout o významu jednotlivých slov.

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

Tabulka 6.8: Porovnání dosažených výsledků člověka a stroje.

Z dosažených výsledků je zřejmé, že stroj nedosahuje schopností člověka. Pokud se vezme v úvahu, že byl člověk oproti stroji vždy v malé výhodě, tak dosažené výsledky jsou relativně optimistické. Minimálně v jednom případě se stroj téměř vyrovnal člověku. Samotnou kapitolou je vliv slovního kontextu. Ještě před samotnými ASR experimenty byl ověřen výkon ASR systému na „kontinuální“ řeči, zde reprezentované větami z testovací sady. Jazykovým modelem je v tomto případě trigramový slovní

model obsahující 1,2 milionu unikátních slov. Přesnost na slovech (počítaná pomocí rovnice (4.63)) dosáhla hodnoty 86,10 %. Při porovnání s výsledky z tab. 6.8 jasné plyne, že pokud je k dispozici dostatečný kontext, tak je ASR schopen správně určit variantu slova. Přeci jen slovo „kosa“ se většinou vyskytuje v trochu jiném slovním kontextu, než slovo „koza“ a toto platí u většiny dvojic.

6.3 Augmentace dat

Poslechový test jasně ukázal, že správné rozpoznání pronesené EL promluvy není lehký úkol ani pro člověka. Naprosto markantní význam hraje kontext. Ten pomáhá, pokud některé části promluvy nebylo dobře rozumět. Navíc, ze zkušeností získaných při pořizování řečového korpusu (části 5.1 a 6.1), plyne, že EL řečník má tendenci mluvit ve spíše kratších dávkách slov, mezi kterými dělá drobné pauzy. Pro člověka není problém udržet v povědomí kontext, ale stroji to může někdy způsobovat problémy. Otázkou tedy je, jak „vylepšit“ stroj tak, aby poskytoval lepší výsledky?

Ať se řečník snaží sebevíc, tak se současnými metodami rehabilitace hlasu (viz 3.2), se při ztrátě hlasivek část informace z produkované řeči ztrácí. V poslední době bylo prezentováno několik přístupů jak ztracenou informaci obnovit. Souhrn těch nejperspektivnějších je v [34]. Ve valné většině případů se využívá obohacení akustického modelu o artikulační data, nebo dokonce využití jen těchto artikulačních dat. [35] Problém ale spočívá v tom, že ne všechny akustické nuance mezi podobnými fonémy jsou artikulací ovlivněny. Navíc její záZNAM s sebou často nese používaní dalšího zařízení (kamery, ultrazvuku, atp. [36]), nebo dokonce nutnost podstoupení dalšího operačního zákroku (magnety [37]). Samozřejmě je férové říct, že většina těchto vyvíjených systémů si klade za cíl kompletně nahradit současné metody rehabilitace. Na druhou stranu faktem je, že ani po dlouholetém vývoji se většina těchto systémů nedostala z

raně vývojové fáze. Nepochybně hraje roli, že je tato problematika přeci jen na okraji 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 informací konkrétní cílenou změnou produkované řeči. Řečník by ideálně neměl být touto změnou ovlivněn. Jako optimální se jeví změna produkované řeči, která je zohledněna modelem. Samozřejmě takovýto přístup nezbaví řečníka EL, ale může mu pomoci v situacích, které jsou pro něj stresující a v konečném důsledku mu velmi komplikují život.

Jako nejjednodušší možnost augmentace se jeví protažení určitých fonémů. Člověk je naprosto bez problémů schopen měnit tempo promluvy. Dokonce velmi často se děje mimoděk, protože tempo řeči velmi významně závisí na emočním a fyzickém stavu jedince. Pokud by se řečník naučil automaticky protahovat určité fonemy, teoreticky by to mohlo pomoci při rozpoznávání. U HMM modelů se délka fonému modeluje pomocí přechodu ze stavu s_x do stejného stavu s_x , viz 4.2.1. Z výsledků „bigrams“ experimentu (část 6.2.2) se dá usuzovat, že modely fonémových párů lišících se znělostí jsou si velmi podobné. Protažení jednoho fonému z inkriminovaného páru může vést k lepšímu odlišení těchto modelů a tím pádem vyšší přesnosti rozpoznání.

K ověření jsou potřeba data. Bohužel získání reálných dat je zdlouhavý proces (viz 5.1 a 6.1), navíc není zřejmé jestli se vůbec vyplatí je pořizovat, protože se jedná o hypotézu. Mnohem prozaičtěji se jeví možnost uměle data protáhnout v místech výskytu zájmových fonémů. Toto protažení je teoreticky možné realizovat dvěma způsoby:

1. protažení na příznacích
2. protažení na zvuku

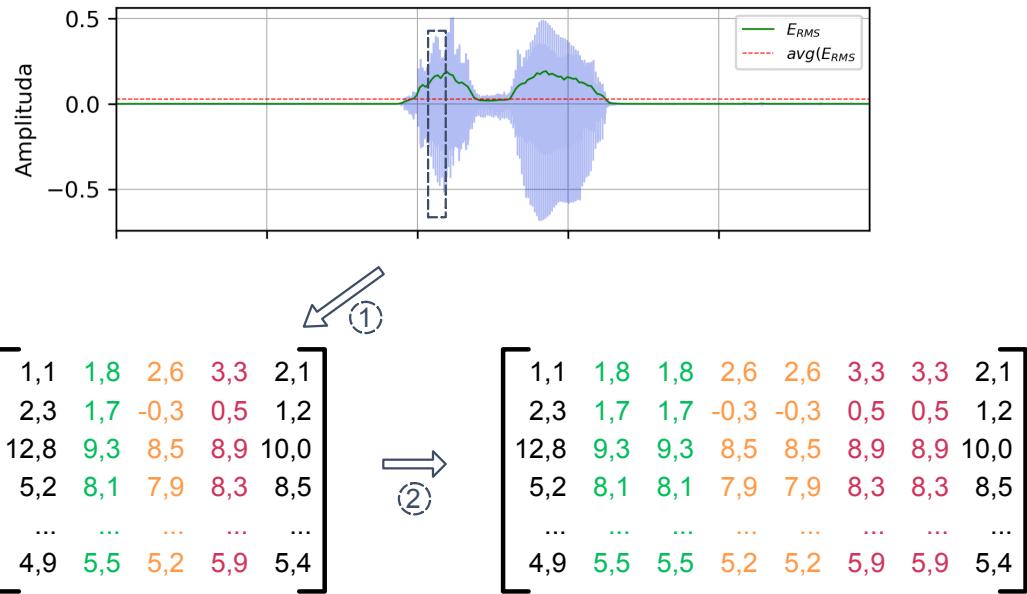
V obou případech je nezbytné získat co možná nejpřesnější fonetické zarovnání. Pokud bude obsahovat chyby, tak mohou být protahovány úplně jiné úseky řeči. K natrénování HMM-DNN modelu na protažených datech je zapotřebí zarovnání získané HMM-GMM nebo HMM-DNN modelem, viz 4.2.3. U obou variant protažení 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 výsledky porovnány s těmi dosavadními. Tyto experimenty navíc pomohou určit vhodné parametry pro případné skutečné protažení dat.

6.3.1 Protažení na příznacích

Protažení na příznacích vychází z představy, že při protažení (např. fonému) a následné parametrizaci, budou v inkriminovaných mikrosegmentech po sobě následovat velmi podobné příznakové vektory. Teoreticky, v krajiném případě, by mohlo dojít i k tomu, že část těchto příznakových vektorů je identických. Pokud je tedy cílem zjistit, zda protažení může pomoci při rozpoznávání EL řeči, tak je možné (teoreticky) docílit protažení zkopirováním určitých příznakových vektorů. Lépe je tato myšlenka ilustrována obr. 6.3. Nejprve je nahrávka standardně parametrizována. Barvně jsou vyznačeny případné vektory odpovídající zájmovému úseku (tedy fonému). Konkrétní hranice jsou získány ze zarovnání. Tyto vektory jsou pak zduplicovány, a tím je „dosaženo“ dvojnásobného protažení.



Obrázek 6.3: Princip protažení na příznacích.

Toto protažení je však spíše hypotetická možnost. V reálné situaci by řečník mluvil jako doposud a k protažení by docházelo až při zpracování. Což je velmi netriviální úkol. Teoreticky by se musel algoritmus parametrizace doplnit o mechanismus, který by určité příznakové vektory definovaným způsobem duplikoval. Problém je, že by ke kopírování docházelo v momentě předzpracování, kdy jsou o daných příznakových vektorech k dispozici minimální informace. Prosté zkopírování navíc poruší dynamický charakter řeči. V rámci jednoho zpracovávaného mikrosegmentu jsou parametry považovány za statické, ale jak se okénko v rámci zpracování posouvá, tak už nelze hovořit o stacionaritě parametrů. Tento problém by se musel řešit nějakým druhem interpolace mezi dvěma konsekutivními vektory. Mechanismus by zároveň vyřešil omezení, kdy je kopírováním možné získat pouze protažení odpovídající celočíselnému násobku původní délky. Proč tedy vůbec zkoušet tento typ protažení? Odpověď je jednoduchá, nehráje zde takovou roli přesnost zarovnání. V průběhu zpracování je využíváno posuvného okénka a překryvu. Díky tomu dojde k určitému „rozmažání“ hranic. Pro prvotní experimenty je to pak relativně vhodné zjednodušení úlohy.

Dosažené výsledky

Prvním bodem výše zmíněného algoritmu je získání standardního modelu, který je použit k zarovnání dat. K tomu je možné použít již natrénovaný model z experimentů popsaných v části 6.1.2. Konkrétně se jedná o HMM-DNN model s 5 skrytými vrstvami, každá s 4096 neurony. Výstupní vrstva je pak typu softmax dimenze rovné počtu HMM stavů. Tento model dosáhl s fonémovým zerogramovým LM přesnosti 84,66 %. S jeho pomocí je získáno zarovnání trénovací i testovací sady.

Jako prvotní ověřovací experiment je zvoleno dvojnásobné protažení fonému /s/. Jinými slovy, všechny vektory odpovídající /s/ jsou zduplikovány. Následně je standardním způsobem natrénován HMM-DNN model. Neuronová síť má 5 skrytých vrstev s 4096. Výstupní vrstva je typu softmax. Otestování je jako v předchozích případech realizováno na testovací sadě s fonémovým zerogramovým jazykovým modelem, aby byl minimalizován vliv LM. Tento nový model dosáhl přesnosti 85,11 %, což je malé zlepšení oproti baseline HMM-DNN modelu.

Protažení /s/ posloužilo k ověření procesu vytváření modelu. Další experiment je realizován na protažených fonémech /k/, /p/, /s/, /t/ a /v/, což představuje většinu neznělých zájmových fonémů. Zarovnání je identické jako u předchozího experimentu. Opět je uvažováno dvojnásobné protažení. Všechny vektory inkriminovaných fonémů jsou zduplikovány. Znovu je natrénován HMM-DNN model se stejnými parametry. Model je otestován s fonémovým zerogramovým jazykovým modelem. Přesnost na testovací sadě dosáhla hodnoty 87,50 %, což lze považovat za významné zlepšení.

Doposud se uvažovalo pouze dvojnásobné protažení, v další fázi je tedy potřeba ověřit jestli jiné hodnoty nemohou poskytnout lepší výsledek. Celkem je uvažované 3x, 4x a 5x protažení. Protaženy jsou fonémy /k/, /p/, /s/, /t/ a /v/. Proces natrénování a otestování modelu je stejný jako v předchozích případech. Dosažené výsledky jsou pak v tab. 6.9. Pro úplnost je tabulka doplněna o baseline model s 1x protažením a již prezentované 2x protažení. Z výsledků je patrný jasný trend, větší než 2x protažení

vede ke zhoršení přesnosti. Optimální hodnota protažení tak teoreticky leží někde v intervalu $(1, 3)x$. Bohužel s protahováním pomocí kopírování příznakových vektorů není možné přesné určení hodnoty.

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

Tabulka 6.9: Vliv míry protažení na přesnost modelu.

Zhoršení přesnosti u vyšších hodnot protažení dozajista souvisí s faktem, že výsledná augmentovaná data neodpovídají realitě. Čím vícekrát je vektor zkopírován, tím více je vnášena chyba způsobená ignorováním dynamické povahy signálu. Nicméně jako proof-of-concept myšlenky posloužil tento experiment velmi dobře.

6.3.2 Protažení na zvuku

Protažení na příznacích vedlo k významnému zlepšení, ale tento přístup není bohužel reálně použitelný. Tím může být až model pracující s fonemy protaženými přímo v audiu signálu. Tato 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. Na základě je možné určitý úsek protáhnout například pomocí:

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

Asi nejednodušší je převzorkování dat, stačí načíst všechny vzorky odpovídající vybranému fonému a změnit vzorkovací frekvenci. Pokud je cílem úsek protáhnout, je

výsledná nová vzorkovací frekvence menší než originální. Hlavním problémem této metody je tonální posun¹¹. Cílem je protáhnout konkrétní foném, který z celkové délky nahrávky zabírá jen malou část. Proto by se dal tento nepříznivý jev ignorovat. Snaha je však vygenerovat co možná nejreálnější protažené nahrávky, a proto není protažení pomocí převzorkování nevhodnější metodou.

Zbylé dva uvažované přístupy umožňují sofistikovanější úpravy signálu. Snahou je upravit časové vlastnosti signálu aniž by byl nepříznivě ovlivněn tón. Obě metody využívají *analýzy* signálu, následované *zpracováním* a zakončené *syntézou*. Rozdíl je hlavně ve způsobu. Metody z rodiny *PSOLA* pracují s hlasíkovými pulsy, které jsou nejprve v analytické části nalezeny¹², aby pak v části zpracování došlo k jejich transformaci na základě požadavků na výslednou řeč. V posledním kroku dochází k syntéze signálu na základě upravených analytických krátkodobých signálů, tedy hlasíkových pulsů. Více detailněji se o této metodě hovoří v [20].

Fázový vokodér pracuje na podobném principu, s tím rozdílem, že v analytické části dochází k převodu signálu do frekvenční oblasti pomocí FFT. Ve fázi zpracování je signál upraven, aby ve fázi syntézy byl opět převeden do časové oblasti pomocí inverzní FFT.

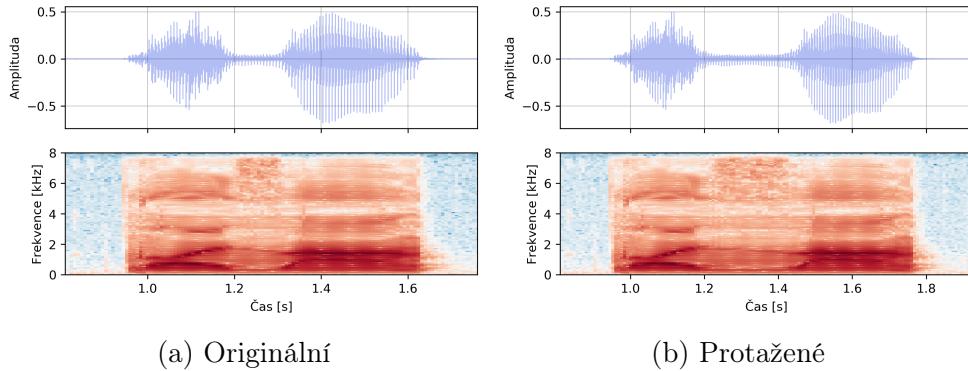
Pomocí těchto dvou zmíněných metod je možné upravit nejen délku, ale i F_0 signálu. Stejně jako u převzorkování mají vliv na signál, ale ten není tak markantní jako v případě převzorkování. U *TD-PSOLA* mohou například vznikat artefakty způsobené nespojitostmi mezi sousedními upravenými úseky řeči. U fázového vokodéru nevznikají artefakty vlivem nespojitostí, ale vlivem fázového posunu.

Obě metody však poskytují velmi dobré výsledky protažení na jednotlivých fonémech. Výsledné protažení je téměř identické. Interně vyvinutý nástroj umožňující ovlivnění délky řeči (a priory používaný při syntéze řeči) poskytuje obě zmíněné me-

¹¹Mění se fundamentální frekvence F_0 . Pokud dojde ke zrychlení, frekvence se zvýší. Při zpomalení naopak sníží.

¹²Výsledkem analýzy jsou periodicky se opakující značky, angl. pitch marks. Úpravou jejich parametrů dochází ke změnám parametrů řeči.

tody. Pro výsledné protažení je použita metoda *TD-PSOLA*. Ukázka původního a protaženého slova „kosa“ je na obr. 6.4. Protahován byl foném /s/, který je v signálu vidět jako šum mezi dvěma výraznými částmi signálu. Inkriminovaný foném byl protažen na dvojnásobek. Na obr. 6.4b je pak zřetelně vidět protažení úseku odpovídající /s/. V signálu a ve spektru není vidět žádný významný artefakt.



Obrázek 6.4: Amplituda a spektrogram původního (protaženého) slova „kosa“.

Dosažené výsledky s DNN

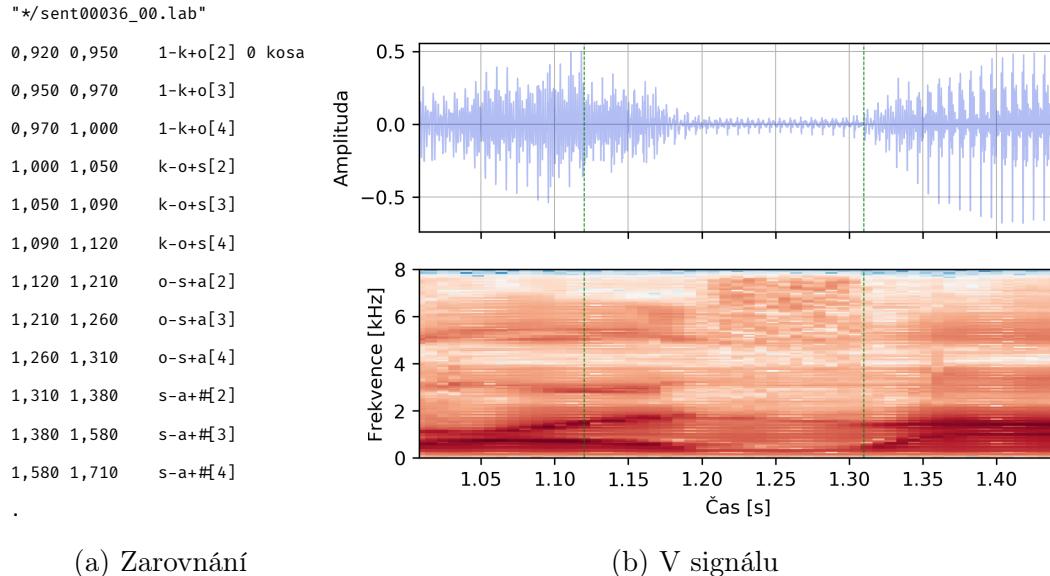
K ověření schopností modelu pracovat s uměle protaženými daty je použit stejný HMM-DNN model jako v předchozích případech. Neuronová síť má 5 skrytých vrstev, každá s 4096 neurony. Výstupní vrstva je pak typu softmax dimenze rovné počtu HMM stavů. Vstupní data jsou parametrizována pomocí PLP (12 statických kepspektrálních koeficientů společně s delta a delta-delta parametry). V datech jsou protaženy všechny výskytu 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 vypsány v tab. 6.10. 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 6.10: Vliv míry protažení fonému na přesnost *DNN* modelu.

Upravené zarovnání a time delay neural network

Při analýze výsledků se ukázalo, že zarovnání v mnoha případech není zrovna nej-přesnější a to zvláště u inkriminovaných neznělých fonémů. Na obr. 6.5 je získané zarovnání slova „kosa“ společně s vyznačenými hranicemi v audiu signálu a spektru. Z obr. 6.5b je zřejmé, že počáteční hranice /s/ zasahuje do předchozího fonému /o/. Tím pádem dochází k protažení nevhodné části signálu a model se tak učí na špatných datech. Pokud by všechny fonémy /s/ následovaly po /o/, tak by se nejednalo o závažný problém, ale toto samozřejmě neplatí.



Obrázek 6.5: Špatně zarovnaný foném /s/ ve slově „kosa“.

V době experimentů s protažením konkrétních fonémů se začaly stále více prosazovat time-delay neural networks (*TDNN*). Přestože patří do rodiny feed-forward

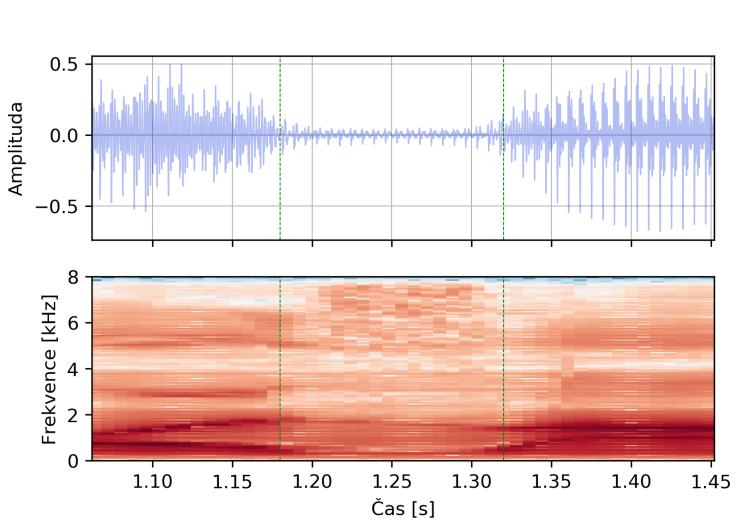
sítí jako DNN, tak oproti nim se snaží vzít v potaz i dynamickou složku řeči, více o rozdílu mezi DNN a TDNN v části 4.2.3.

Stejně jako u DNN modelu je na počátku trénování nutné mít k dispozici zarovnání. To, ale nemusí být naprostě přesné, protože je vstupní vektor zpracováván jiným způsobem než u DNN. Díky FIR filtraci a více množin vah je brán v potaz i dynamický charakter řeči [28]. Model založený na TDNN by tak měl generovat přesnější zarovnání, a tím zlepšit výsledky modelu pracujícího s uměle protaženými daty.

Jako startovní bod trénování je použito DNN zarovnání z předchozího experimentu. Topologie sítě vychází z hodnot prezentovaných v [28], tedy síť má 4 skryté vrstvy. Každá vrstva má 650 neuronů. První vrstva pracuje s kontextem $t - 2$ a $t + 2$, druhá vrstva s $t - 1$ a $t + 2$, třetí vrstva s $t - 3$ a $t + 4$ a čtvrtá s $t - 7$ a $t + 2$. Výpočet výstupu pak bere v potaz kontext $t - 13$ a $t + 9$ mikrosegmentů, viz [28].

Na obr. 6.6 je zobrazeno získané zarovnání slova „kosa“ TDNN modelem. Z vyznacných hranic fonému */s/* (obr. 6.6b) je patrné v podstatě přesné zarovnání. Přesnost TDNN modelu s fonémovým zerogramovým jazykovým modelem dosáhla hodnoty $Acc_p = 85,41\%$.

```
"/sent00036_00.lab"
0,920 0,950    #-k+o[2] 0 kosa
0,950 0,970    #-k+o[3]
0,970 1,010    #-k+o[4]
1,010 1,080    k-o+s[2]
1,080 1,140    k-o+s[3]
1,140 1,180    k-o+s[4]
1,180 1,210    o-s+a[2]
1,210 1,270    o-s+a[3]
1,270 1,320    o-s+a[4]
1,320 1,380    s-a+#[2]
1,380 1,550    s-a+#[3]
1,550 1,680    s-a+#[4]
```



(a) Zarovnání

(b) V signálu

Obrázek 6.6: Správně zarovnaný foném */s/* ve slově „kosa“.

S relativně přesným zarovnáním je možné přistoupit k protažení fonémů $/k/, /p/, /s/, /t/, /v/$ a vytvoření nového modelu pracujícího s těmito daty. Nejprve je natrénovat HMM-GMM model, který poskytne prvotní zarovnání. V předchozích experimentech následovalo trénování DNN modelu, ale TDNN zarovnání ukázalo lepší vlastnosti tohoto modelu. 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 vypsány v tab. 6.11. Oproti těm v tab. 6.10 je vidět výrazné zlepšení přesnosti oproti baseline 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 podobných výsledků. To poskytuje relativně široký pracovní interval pro případné skutečně protažená data 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 6.11: Vliv míry protažení fonému na přesnost TDNN modelu.

Podstatnou otázkou je robustnost duration modelu, tedy jak moc se výsledky změní pokud vstupem modelu, natrénovaného na datech s určitým protažením, jsou data s jinou mírou protažení. Tab. 6.12 ukazuje, že pokud jsou vstupem $2,5x$ modelu data od $2,0x$ do $3,0x$, tak zhoršení přesnosti dosahuje maximálně $1,66\%$ absolutně. Dá se tedy říci, že duration model je v rámci možností robustní v celém širokém rozsahu protažení. Očekávaným výsledkem je nejmarkantnější pokles pokud jsou vstupem nepotažená data (konkrétně $9,29\%$). Zároveň většina chyb byla v inkriminovaných protažených fonémech. Tento výsledek je předpokladem pro funkci trenérku prezentovaného v části 6.5.

	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 6.12: 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. Svůj díl na tom má i použití TDNN modelu. Model pracující s duplikovanými příznaky naznačoval, že optimální hodnota protažení bude v intervalu $(1, 3) x$, což druhý typ modelu potvrdil.

6.3.3 Aktualizace výsledků porovnání

V části 6.2 je prezentováno srovnání schopností člověka a stroje. Posloužily k tomu dva poslechové testy a celkem 3 ASR experimenty „one-mil“, „reduced“ a „bigrams“. S novým modelem je možné aktualizovat hypotetické výsledky stroje. Hypotetické z toho důvodu, že použitá data jsou uměle protažena. Nicméně to nebrání provedení tohoto experimentu. Získané hodnoty mohou být brány jako jakási teoretická maxima ASR systému. Uměle upravená data budou přeci jen relativně přesně a konzistentně protažena. U reálných dat toto nelze a priory očekávat.

K aktualizaci výsledků je použit nejlepší model z části 6.3.2, tedy ten s $2,5x$ protaženými daty. Parametry experimentů jsou totožné s těmi v části 6.2. V případě „one-mil“ je použit zerogramový jazykový model s 1 milionem slov, „reduced“ pak pouze zerogramový LM se slovy obsaženými v poslechovém testu ($N = 320$). Speciální LM, obsahující 4 kombinace slov, je vygenerován pro každou položku „bigrams“ experimentu.

Dosažené výsledky jsou v 6.13. Ve všech třech experimentech došlo k významnému zlepšení. U „one-mil“ to je 23 % absolutně, u „reduced“ pak 24 %. K nejmarkantnějšímu zlepšení došlo u experimentu „bigrams“, dokonce 44 %. Výsledky člověka jsou stejné, protože realizace poslechového testu je zdlouhavý proces a je obtížné získat dostatečný počet respondentů. Nicméně se dá očekávat, že i člověk by dosáhl zlepšení. Pokud by měl znalost o fonémech, které jsou protaženy. V opačném případě by ke zlepšení nutně nemuselo dojít, protože kromě protažení nebyl zvuk nijak pozměněn.

Velmi zajímavé je porovnání zvýšení přesnosti TDNN modelu s fonémovým zero-gramovým LM (2,49 % absolutně mezi baseline a $2,5x$ modelem, viz tab. 6.11) a výsledky dosaženými prezentovanými v tab. 6.13. Oproti nim je zlepšení o 2,49 % absolutně poměrně zanedbatelné, přesto velmi významné zlepšení. Jasně to ukazuje ideu experimentů s fonémovým zero-gramovým jazykovým modelem. I drobné zlepšení u akustického modelu může vést k rapidnímu zlepšení sofistikovanějšího systému.

	$Acc_p [\%]$		
	one-mil	reduced	bigrams
<i>člověk</i>	70,47	70,47	66,24
<i>stroj (baseline)</i>	61,24	69,91	54,82
<i>stroj (augmented)</i>	84,95	94,36	98,80

Tabulka 6.13: Aktualizované porovnání dosažených výsledků člověka a stroje.

6.3.4 Reálně protažená data

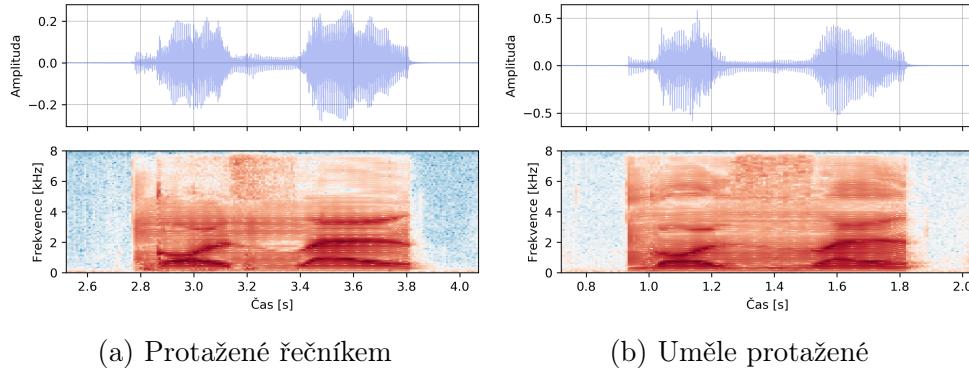
Získané výsledky s uměle protaženými daty potvrdily hypotézu, že model pracující s těmito daty může dosahovat lepších výsledků. Dalším krokem je získání reálně protažených dat. Nahrávání je relativně zdlouhavý proces, jak ukázala 1. a 2. etapa nahrávání (části 5.1 a 6.1), proto je nezbytné dobře vybrat promluvy. Problémem je navíc samotné protažení. Pokud by bylo cílem získat protažené celé slovo, lze řeč-

níka instruovat, aby mluvil pomaleji. To, ale cílem není. Výsledné promluvy mají mít protažené pouze určité fonémy, a to přibližně na dvojnásobek. Jako nejjednodušší, a svým způsobem i elegantní, se ukázal zápis se zdvojenými písmeny, které mají být protaženým fonémem, např. „kossa“. Řečník je obeznámen, že pokud slovo v promluvě obsahuje tento dvojitý zápis, měl by se pokusit toto slovo patřičným způsobem protáhnout. Tento zápis navíc řečníka podvědomě „nutí“ vyslovit slovo jinak než v případě normálního zápisu.

Nahrávání se zhstil 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. K nahrávání byla použita stejná nahrávací místo a aparatura jako v případě 2. etapy nahrávání (viz 6.1). Nahrávací protokol byl taktéž stejný, tzn. nahrávání byl vždy přítomen operátor, který kontroloval, zda se řečník nevědomky nedopustil chyby či přečeknutí. Stejně jako v předchozí etapě obsahuje každá nahrávka minimálně 0,5 s pauzu na začátku a konci. Tato opatření významně zkracují potřebnou dobu k získání kvalitního přepisu. Celkem se takto v rámci 3. etapy nahrávání podařilo získat dohromady 998 promluv obsahující věty a slova (protažená i neprotažená), kterých je 267. Celkový čas promluv v 3. etapě dosáhl 2 hodin a 28 minut.

Na obr. 6.7 je zobrazena amplituda a spektrogram slova „kose“ protaženého řečníkem (6.7a) a 2x uměle (6.7b). Hlavním viditelným rozdílem je slabší zastoupení frekvencí kolem 2 kHz ve spektrogramu mezi časem 3,2 s a 3,4 s. Ač to tak na první pohled nevypadá, tak obě protažení mají téměř identickou délku 0,258 s (reálné) vs. 0,261 s (umělé). Vizuální rozdíl je způsoben vyšší celkovou délkou reálně protaženého slova. Foném /s/ je zástupcem neznělých fonémů, zdrojem je tedy a priory šum

a nikoli periodický signál produkováný hlasivkami, tím pádem je velikost amplitudy také velmi podobná. Další možný vizuální rozdíl (v místě protažení fonému /s/) mezi obr. 6.7a a 6.7b je způsoben vyšší maximální amplitudou u uměle protažené nahrávky (6.7b), která pochází z 2. etapy nahrávání.



Obrázek 6.7: Amplituda a spektrogram slova „kose“ protažené řečníkem/uměle.

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.

Porovnáním neprotažených nahrávek z 2. etapy a protažných z 3. etapy se jako ideální jeví použití 2x protaženého modelu z části 6.3.2 (viz tab. 6.11). Jedná se tedy o TDNN model, který je natrénován z dat obsahujících 2x protažené fonémy /k/, /p/, /s/, /t/ a /v/. Na testovací sadě dosáhl přesnosti $Acc_p = 87,71\%$. K otestování výkonu na reálně protažených datech jsou použity všechny věty a slova obsahující protažení zmíněných fonémů¹³. S touto testovací sadou dosáhl zmíněný model s fonémovým zero gramovým jazykovým modelem $Acc_p = 84,51\%$.

¹³Stejně jako u 1. a 2. etapy bylo i zde aplikováno CMN napočítané přes všechny nahrávky v rámci etapy.

Dosažená přesnost je sice horší než původní přesnost na uměle protažených datech, ale zároveň nedošlo k dramatickému propadu jako v případě křížového testu mezi 1. a 2. etapou (viz tab. 6.4¹⁴). Dosažený výsledek tak potvrzuje podobnost uměle a reálně protažených dat.

V případě, že se k trénovací sadě přidaly věty z 3. etapy, a to té části, která neobsahuje protažené fonemy, tak výsledná přesnost TDNN modelu s fonémovým zero gramovým jazykovým modelem dosáhla hodnoty $Acc_p = 85,88\%$. Což je lepší hodnota než baseline TDNN model ($Acc_p = 85,41\%$, viz tab. 6.11). Tyto experimenty podporují ideu trenažéra, prezentovaného v části 6.5.

6.4 Model akcentující protažení dat

Augmentace dat (viz část 6.3) ukázala, že ztracenou informaci EL řeči je částečné možné nahradit protažením inkriminovaných fonémů. Experimenty s reálně protaženými daty (viz část 6.3.4) navíc prokázaly schopnost člověla toto protažení realizovat. Dalším krokem je tedy úprava modelu tak, aby tuto změnu co možná nejvíce reflektoval.

Z principu fungování a nejpoužívanější topologie HMM modelu (viz 4.2.1) je délka fonému modelována pomocí přechodových pravděpodobností. Ty zase vedou na funkce geometrické distribuce pravděpodobnosti [38]. Bohužel skutečná podoba těchto distribucí odpovídá spíše gamma nebo logaritmicko-normálnímu rozdělení [39].

Správné modelování délky může být realizováno úpravou přechodových hustotních funkcí v HMM nebo změnou topologie modelu. Druhou možností je vytvoření speciálního modelu pracujícího s délkou jednotlivých fonémů (duration model) a re-skórováním výstupních N-best hypotéz či celé rozpoznávací mřížky.

¹⁴V tabulce jsou prezentovány hodnoty odpovídající GMM modelu. Srovnání na první pohled není úplně korektní, nicméně validní, protože GMM model trénováný na uměle protažených datech a otestovaný na reálně protažených datech dosáhl $Acc_p = 81,29\%$. Což je srovnatelná hodnota s ostatními GMM modely.

Přestože se první uvažovaný způsob jeví jako vhodnější, tak všechny dosavadní publikované výsledky ukazují významné zvýšení výpočetní náročnosti dekódování a komplexity modelu [38] [40] [41]. Tento přístup je však často používán u HMM syntézy řeči, viz [42].

6.4.1 Princip explicitních duration modelů

Druhou možností je vytvoření explicitního modelu pracujícího s délkou fonémů. S tímto modelem je často problém rozpoznávání přeformulován do úlohy nalezení nejlepší sekvence slov W^* a odpovídajících délek D^* , na základě akustického modelu [39]. Za předpokladu, že je dána sekvence slov W a vektory pozorování O lze považovat za nezávislé na délkách D , je možné rovnici 4.1 upravit jako

$$\begin{aligned} W^*, D^* &= \operatorname{argmax}_{W,D} P(W, D|O) \\ &= \operatorname{argmax}_{W,D} P(O, D|W) P(W) \\ &= \operatorname{argmax}_{W,D} P(O|W) P(D|W) P(W). \end{aligned} \quad (6.8)$$

Úkolem duration modelu je tedy odhad pravděpodobnosti $P(D|W)$. Délku D je možné dekomponovat na m délek jednotlivých fonémů d_i

$$P(D|W) = P(d_1, \dots, d_m|W). \quad (6.9)$$

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

$$\begin{aligned}
P(d_1, \dots, d_m | W) &= \prod_{i=1}^m P(d_i | d_1, \dots, d_{i-1}, W) \\
&\approx \prod_{i=1}^m P(d_i | d_{i-n+1}, \dots, d_{i-1}, W).
\end{aligned} \tag{6.10}$$

Model tedy odhaduje $P(D|W)$ na základě n předchozích délek fonémů a předpokládaného slova W . Některé duration modely navíc ještě pracují i s tempem řeči [40], nicméně tento efekt je u modelu využívající rovnici 6.10 zakomponován v délkách n předchozích fonémů.

Ve skutečnosti je vhodné vytvořit model, který bere v potaz nejen předchozí délky, ale i příznakové vektory těchto fonémů [39]. Tedy $P(d_i|x_i)$, 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. [39] [43]

Na odhad $P(d_i|x_i)$ je možné nahlížet ze dvou pohledů. V prvním případě je cílem modelu odhadnout parametry pravděpodobnostní distribuce pomocí conditional density estimation network (CDEN). [39] V tomto případě je předpokládáno, že délky fonémů odpovídají určitému pravděpodobnostnímu rozdělení, nejčastěji logaritmicko-normálnímu. Konkrétní hodnota pravděpodobnosti je pak vypočtena dosazením do příslušného vzorce 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 [43] ukazují, že tento přístup je vhodnější¹⁵.

¹⁵Při 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.

6.4.2 Duration model se softmax vrstvou

Neuronová síť mající na svém výstupu softmax vrstvu (viz rovnice 4.53) určuje diskrétní pseudo-pravděpodobnosti m tříd. V případě duration modelu, se jako vhodné jeví reprezentovat jednotlivé třídy jako počet mikrosegmentů ($d = 1, 2, 3, \dots$), které odpovídají danému fonému. Čistě teoreticky může být těchto mikrosegmentů nekonečné množství, síť však na svém výstupu potřebuje konečný počet tříd (počet neuronů ve výstupní vrstvě). Jako vhodné řešení tohoto problému se ukázalo zvolení maximální délky fonému D . Pro všechny s délkou $d \geq D$ platí, že $p(d) = p(D)$. [43] Volba D závisí na konkrétní doméně a je vhodné ji určit experimentem.

Cílem modelu je predikovat sekvenci délek na základě sekvence fonémů. To implikuje možnost použití levého L i pravého R kontextu fonému i . Do vstupního vektoru sítě, ale mohou přijít délky pouze fonémů L nebo R kontextu. Pokud by totiž byly použity oba kontexty, tak by délka fonému i závisela na délce fonému $i + 1$. Zároveň by, ale délka fonému $i + 1$ závisela na délce fonému i . Tím pádem by došlo ke kruhové závislosti, kterou není možné vyřešit. Standardně se volí L kontext pro délky. Příznakový vektor tedy 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 5.3.2). 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 $d = 1, 2, \dots, D$ pomocí sigmoid funkce

$$d' = \frac{2}{1 + e^{-0,01d}} - 1, \quad (6.11)$$

která transformuje hodnoty do požadovaného intervalu $(0, 1)$ [39]. Pokud není kontext k dispozici (krajní případy), tak $d = 0$.

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

Samotné reskórování výstupní mřížky je realizováno přidáním $\log p(d_i|x_i)$, kde x_i je vstupní příznakový vektor duration modelu, k hodnotám získaným z akustického a jazykového modelu. Mřížka je mezivýsledek, ze kterého je následně vydekódován výstup ASR systému. Samotné duration skóre je navíc přenásobeno konstantou získanou z development sady v průběhu trénování modelu tak, aby jeho řád odpovídal hodnotám z ostatních modelů. [43] Stejně jako v případě jazykového modelu je i zde tzv. váha duration modelu, která umožňuje měnit vliv tohoto modelu.

6.4.3 Dosažené výsledky

Stejně jako v případě augmentace dat (viz 6.3.2) je potřeba k natrénování duration modelu kvalitní zarovnání. Jedním 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 datech¹⁶, viz tab. 6.11.

Samotný duration model (popsaný v předchozí části 6.4.2) je typu feedforward. Počet skrytých vrstev sítě se odvíjí od konkrétní řešené domény, ale standardně se uvažují 2 případně 3, viz [43]. Velikost těchto skrytých vrstev je volena jako násobek

¹⁶Natrénování TDNN modelu pomocí reálně protažených dat nebylo vhodné, protože nebylo k dispozici dostatečné množství reálně protažených dat.

dimenze příznakového vektoru, v tomto případě byl zvolena hodnota $3I$. Aktivační funkce je typu RELU. Velikost výstupní vrstvy odpovídá maximálnímu počtu mikrosegmentů D , v [43] 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 datech¹⁷. 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 stejně trénovací a testovací sady jako modely v části 6.3.2. 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 ($Acc_p = 87,71\%$). 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 nepartažená 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 [43], kde se však pracovalo se standardní neprotoženou řečí. Tab. 6.14 ukazuje vliv maximální délky na přesnost modelu. Speciální je hodnota $D = 189$, která je určena automaticky na základě zarovnání před samot-

¹⁷Hodnota $2x$ je zvolena, protože se nejvíce blíží reálně protaženým datům.

ným trénováním. Model s $D = 189$ zároveň dosáhl nejvyšší přesnosti $Acc_p = 88,58\%$, což představuje drobné zlepšení oproti původnímu modelu s $D = 50$.

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

Tabulka 6.14: Vliv maximální délky na přesnost modelu.

Dalším hyperparametrem, který může ovlivnit kvalitu modelu je počet vrstev neuronové sítě. V tab. 6.15 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 [43]. Z dosažených výsledků je patrné, že velikost sítě není úplně zásadním parametrem. Rozdíl mezi přesností sítě s 2 a 3 vrstvami je minimální. Přínos bottleneck vrstvy, oproti výsledkům prezentovaným v [43], 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 6.15: Vliv počtu skrytých vrstev na přesnost modelu ($D = 189$ ¹⁸).

Posledním hyperparametrem, který může mít vliv na přesnost modelu je velikost L a R kontextu. Z tab. 6.16 a tab. 6.17 vyplývá, že nejlepších výsledků dosahuje modely, které má délku kontextu $L + R = 6$. Úplně nejlepšího výsledku pak dosáhl

¹⁸V průběhu určování nejlepších kombinací hyperparametrů byly otestovány všechny kombinace velikosti sítě a maximální délky D . Nejlepších výsledků dosahovaly modely s $D = 189$.

model mající symetrický kontext, ale oproti modelům s asymetrickým kontextem je rozdíl spíše zanedbatelný.

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 6.16: 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 6.17: Vliv levého a pravého kontextu v případě, že celková délka $L + R = 6$.

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 vyvzakovat závěr, že pokud by byl model natrénován z dostatečného množství reálně protažených dat, tak by se jeho výsledky blížily výsledkům modelu na uměle protažených datech.

6.4.4 Aktualizace výsledků porovnání

V části 6.2 a 6.3.3 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. 6.18 jsou pak předchozí výsledky doplněny o hodnoty dosažené duration modelem jehož parametry odpovídají nejlepšímu modelu z předchozí části 6.4.2.

Stejně jako v předchozích experimentech je u „one-mil“ experimentu použit zero-gramový 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.

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

Tabulka 6.18: Aktualizované porovnání dosažených výsledků člověka a stroje.

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. Hlavním nedostatkem aktuální implementace reskóringu pomocí duration modelu spočívá v tom, že se jedná o offline přístup. Promluvy jsou nejprve zpracována 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.

6.5 Trenažér

V předchozích částech (6.3 a 6.4) byla rozvíjena a ověřována myšlenka doplnění chybějící informace, způsobenou ztrátou hlasivek, pomocí protahování určitých vybraných fonémů. Zejména pak /k/, /p/, /s/, /š/, /t/, /t'/ a /v/ reprezentující neznělé fonémy. Prezentované výsledky (viz tab. 6.11 a 6.14) prokazují, že daný přístup poskytuje výrazné zlepšení přesnosti ASR systému, zejména u promluv s minimálním kontextem.

Hlavním problém tohoto přístupu spočívá v protažení 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 6.4.2 demonstrují, že model natrénovaný na uměle protažených datech je schopen lépe rozpoznávat i reálně protažené fonémy. V případě rozpoznávání nepotažených dat, pak přesnost modelu významně klesá. Tento výsledek je fundamentálním předpokladem pro myšlenku trenažéra. Jeho hlavní funkcí je pomoci řečníkovi naučit se automaticky protahovat inkriminované fonémy tak, aby přesnost rozpoznávání byla maximální. Zároveň je možné pomocí trenažéra 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 trenažéra je pomoc řečníkovi 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 6.3.4). Ten se ukázal jako velmi názorný a podvědomně 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 rádech 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 modelu¹⁹.

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. [44]

Jednou z prerekvizit trenažéru je možnost individuálního použití doma. Zejména, protože odpadá nutnost použití specializovaného HW, či zvukové komory. Nezanebatelným 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 [45] a [46].

¹⁹V případě dostatečného množství reálně protažených dat, pak natrénování nového modelu na reálných datech.

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 priory se však jedná spíše o implementační detailly, 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 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. 6.12, kde je jasné 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.

Kapitola 7

Závěr

Přestože nádorovitá onemocnění hrtanu nepatří mezi nejčastější onemocnění, je jim věnována značná pozornost. Případné následky totiž mohou výrazně zhoršit kvalitu života pacienta. Klasické rehabilitační techniky (zmíněné v části 3.2) dokážou navrátit schopnost mluvit, ale kvalita produkované řeči nemusí být rozhodně optimální. 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, popsané v části 3.2, ale ty jsou zatím používány spíše autorštími 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 je také podstatný psychologický efekt způsobující obtíže při mluvení na veřejnosti.

Pomoc může poskytnou rozvoj technologií zpracovávajících přirozenou řeč (ASR). Hlavním nedostatkem obecných ASR systému je jejich nekompatibilita s TL řečí (viz část 5.3). I v případě natrénování individuálního ASR systému nedosahují jejich výkony (viz tab. 5.3) těch obecných ASR systémů se zdravými řečníky. Hlavním problémem je přílišná odlišnost TL řeči od té normální. Jako problematické se pak ukazují zejména neznělé fonémy, které se už z podstaty produkce TL řeči liší od neznělých zdravého řečníka (viz část 5.2). A je jedno jestli se jedná o jícnový hlas, tracheoezofageální fistuli nebo elektrolarynx.

Problém TL řeči se dá vztáhnout ke ztrátě určitého množství informace z promluvy. V případě, že je k dispozici dostatečný slovní kontext, tak velmi pozitivně výsledek ovlivňuje jazykový model. Přece jen vetšina slov mající jiný význam a akusticky se lišících pouze znělostí jednoho fonému se vyskytuje v jiném slovním kontextu. Problém by se tak mohl jevit jako marginální. Bohužel TL řečníci se v mnoha případech snaží mluvit spíše v kratších úsecích, což problém s chybějící informací podtrhuje.

Doplnění chybějící informace je možné v případě využití různých druhů multimodálních systémů, které se snaží různými způsoby, zejména na základě snímání artikulace. Tyto systémy jsou však ještě musí urazit dlouhou cestu k dosažení produkčního nasazení. Hlavním problémem jsou zatím výkony některých systému a také jimi kladené nároky na mluvčího.

Tato práce se zaměřila na možnost doplnění informace pomocí drobné cílené změny produkované řeči a úpravy ASR systému tak, aby tato změna byla co možná nejvíce akcentována. V části 6.3 je představena a následně otestována možnost protahování vybraných fonémů k minimalizaci problémů ASR systémů s krátkými promluvami. K ověření funkčnosti konceptu byla data nejprve uměle protažena. Dosažené výsledky (viz tab. 6.9 a tab. 6.11) potvrzují nezanedbatelné zlepšení pouze drobnou úpravou

pronášených promluv. Pokud se navíc upraví ASR systém tak, aby toto prodloužení příslušně využil, dojde ještě k dalšímu zlepšení (viz tab. 6.15, 6.16 a 6.17).

K reálně použitelnému systému jsou samozřejmě zapotřebí reálně protažená data, ale díky těm umělým bylo možné určit vhodnou míru protažení. V části 6.3.4 je pak představen jednoduchý a pro řečníka intuitivní způsob získání reálně protažených dat. Tento způsob společně s uměle protaženými daty jsou pak stavebními kameny představeného trenažéra (viz část 6.5), který slouží ke snadnému trénování řečníka v protahování. Sekundární funkcí je pak získání reálných dat, které pak mohou posloužit k vytvoření co možná nejlepšího ASR systému.

Hlavním nedostatkem představených ASR systému postavených na duration modelech (viz část 6.4) je v současné chvíli jejich offline funkcionality. V současné podobě je není možné využít jejich použití v systémech zpracovávajících promluvy v reálném čase. Principiálně tomu však nic nebrání.

V ideálním případě pak může ASR systém s duration modelem ve spojení s TTS systémem posloužit v určitých situacích ke snadnější komunikaci TL řečníků s ostatními lidmi. Při diskuzích s pacienty po TL laryngektomii se jako velmi lukrativní jeví využití takového systému například při telefonování, protože to drasticky redukuje míru stresu a strachu z reakce druhé strany. Ač se to na první pohled nemusí zdát, může to vést k významnému zlepšení kvality života a snížení psychologické zátěže TL řečníků. Zejména pak v kritické době následující po operaci.

Seznam použité literatury

- [1] Slavíček, Aleš. *Operace hrtanu*. Praha: Nakladatelství TRITON, s.r.o., 2000, s. 53. ISBN: 80-7254-130-7.
- [2] Škvárová, Jana. „Úloha ošetřovatelské péče při zvládání psychických a sociálních obtíží u nemocných po tracheostomii po totální laryngektomii“. Disertační práce. Masarykova univerzita, Lékařská fakulta, 2010.
- [3] Gussenbauer, Carl a Billroth, Theodor. *Über die erste durch Th. Billroth am Menschen ausgeführte Kehlkopf Extirpation und die Anwendung des künstlichen Kehlkopfes*. Sittenfeld, 1874.
- [4] Kramp, Burkhard a Dommerich, Steffen. „Tracheostomy cannulas and voice prosthesis.“ In: *GMS current topics in otorhinolaryngology, head and neck surgery* 8 (led. 2009), Doc05. ISSN: 1865-1011. DOI: [10.3205/cto000057](https://doi.org/10.3205/cto000057).
- [5] Šebová-Šedenková, Irina. „Možnosti rehabilitácie hlasu po laryngektómii (Historický prehľad a súčasné trendy)“. In: *Choroby hlavy a krku (Head and Neck Diseases)* 1 (2006), s. 44–50. ISSN: 1210-0447.
- [6] Seeman, M. „Speech and voice without larynx“. In: *Cas Lek Cas* 41 (1922), s. 369–72.
- [7] Brown, Dale H. et al. „Postlaryngectomy Voice Rehabilitation: State of the Art at the Millennium“. In: *World Journal of Surgery* 27.7 (čvc 2003), s. 824–831. ISSN: 0364-2313. DOI: [10.1007/s00268-003-7107-4](https://doi.org/10.1007/s00268-003-7107-4).

- [8] Hradecká, Zuzana. „Fyziologie lidského hlasového ústrojí“. Diplomová práce. Masarykova univerzita, Lékařská fakulta, 2007, s. 1–105.
- [9] Kikuchi, Yoshinobu a Kasuya, Hideki. „Development and evaluation of pitch adjustable electrolarynx“. In: *Speech Prosody 2004, International ...* (2004), s. 761–764.
- [10] Uemi, Norihiro et al. „Design of a new electrolarynx having a pitch control function“. In: *Proceedings of 1994 3rd IEEE International Workshop on Robot and Human Communication*. IEEE, 1994, s. 198–203. ISBN: 0-7803-2002-6. DOI: [10.1109/ROMAN.1994.365931](https://doi.org/10.1109/ROMAN.1994.365931).
- [11] Goldstein, Ehab a et al. „Design and implementation of a hands-free electro-larynx device controlled by neck strap muscle electromyographic activity.“ In: *IEEE transactions on bio-medical engineering* 51.2 (ún. 2004), s. 325–32. ISSN: 0018-9294. DOI: [10.1109/TBME.2003.820373](https://doi.org/10.1109/TBME.2003.820373).
- [12] Liu, Hanjun a Ng, Manwa L. „Electrolarynx in voice rehabilitation.“ In: *Auris, nasus, larynx* 34.3 (zář. 2007), s. 327–32. ISSN: 0385-8146. DOI: [10.1016/j.anl.2006.11.010](https://doi.org/10.1016/j.anl.2006.11.010).
- [13] Leunisse, C et al. „Biofilm formation and design features of indwelling silicone rubber tracheoesophageal voice prostheses—an electron microscopical study.“ In: *Journal of biomedical materials research* 58.5 (led. 2001), s. 556–63. ISSN: 0021-9304. DOI: [10.1002/jbm.1054](https://doi.org/10.1002/jbm.1054).
- [14] Slavíček, Aleš. „Možnosti rehabilitace pacientů po onkologických výkonech v oblasti hlavy a krku : Hlavní téma: Onkologie v otorinolaryngologii“. In: *Postgraduální medicína : odborný časopis pro lékaře* 4.9 (2002), s. 1029–1035. ISSN: 1212-4184.
- [15] Saito, Hitoshi et al. „Tracheoesophageal shunt method with omohyoid muscle loop for voice restoration.“ In: *Archives of otolaryngology–head & neck surgery*

129.3 (břez. 2003), s. 321–3. ISSN: 0886-4470. DOI: [10.1001/archtol.129.3.321](https://doi.org/10.1001/archtol.129.3.321).

- [16] Narula, Tony et al. *Laryngeal transplantation: working party final report*. London: The Royal College of Surgeons of England, 2011, s. 15.
- [17] Strome, M et al. „Laryngeal transplantation and 40-month follow-up.“ In: *The New England journal of medicine* 344.22 (květ. 2001), s. 1676–9. ISSN: 0028-4793. DOI: [10.1056/NEJM200105313442204](https://doi.org/10.1056/NEJM200105313442204).
- [18] Holmes, Wendy. *Speech synthesis and recognition*. CRC press, 2001.
- [19] Benesty, Jacob, Sondhi, M Mohan a Huang, Yiteng. *Springer handbook of speech processing*. Springer, 2007.
- [20] Psutka, Josef et al. *Mluvíme s počítačem česky*. Prague: Academia, 2006, s. 752. ISBN: 80-200-1309-1.
- [21] Van Der Malsburg, C. „Frank Rosenblatt: Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms“. In: *Brain Theory*. Ed. Palm, Günther a Aertsen, Ad. Berlin, Heidelberg: Springer Berlin Heidelberg, 1986, s. 245–248. ISBN: 978-3-642-70911-1.
- [22] Amodei, Dario et al. „Deep speech 2: End-to-end speech recognition in english and mandarin“. In: *International conference on machine learning*. 2016, s. 173–182.
- [23] Hinton, Geoffrey et al. „Deep neural networks for acoustic modeling in speech recognition“. In: *IEEE Signal processing magazine* 29 (2012).
- [24] Veselý, Karel et al. „Sequence-discriminative training of deep neural networks.“ In: *Interspeech*. Sv. 2013. 2013, s. 2345–2349.
- [25] Hannun, Awni Y. et al. „Deep Speech: Scaling up end-to-end speech recognition“. In: *CoRR* abs/1412.5567 (2014). arXiv: [1412.5567](https://arxiv.org/abs/1412.5567).

- [26] Waibel, A. et al. „Phoneme recognition using time-delay neural networks“. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.3 (břez. 1989), s. 328–339. ISSN: 0096-3518. DOI: [10.1109/29.21701](https://doi.org/10.1109/29.21701).
- [27] Jin, Craig, Schenkel, Markus a Carlile, Simon. „Neural system identification model of human sound localization“. In: *The Journal of the Acoustical Society of America* 108 (říj. 2000), s. 1215–35. DOI: [10.1121/1.1288411](https://doi.org/10.1121/1.1288411).
- [28] Peddinti, Vijayaditya, Povey, Daniel a Khudanpur, Sanjeev. „A time delay neural network architecture for efficient modeling of long temporal contexts“. In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.
- [29] Radová, Vlasta a Psutka, Josef. „UWB_S01 corpus - a czech read-speech corpus“. In: led. 2000, s. 732–735.
- [30] V., Psutka Josef, Šmíd, L. a Pražák, A. „Searching for a robust MFCC-based parameterization for ASR application“. In: Lisabon: INSTICC PRESS, 2007, s. 196–199. ISBN: 978-989-8111-13-5.
- [31] Stolcke, Andreas. „SRILM-an extensible language modeling toolkit“. In: *Seventh international conference on spoken language processing*. 2002.
- [32] Pražák, Aleš et al. „Efficient combination of N-gram language models and recognition grammars in real-time LVCSR decoder“. In: *2008 9th International Conference on Signal Processing*. IEEE. 2008, s. 587–591.
- [33] Povey, Daniel et al. „The Kaldi Speech Recognition Toolkit“. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Catalog No.: CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, pros. 2011.

- [34] Denby, Bruce 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](https://doi.org/10.1016/j.specom.2009.08.002).
- [35] Hofe, Robin 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](https://doi.org/10.1016/j.specom.2012.02.001).
- [36] Hueber, Thomas 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](https://doi.org/10.1016/j.specom.2009.11.004).
- [37] Hofe, Robin et al. „Speech Synthesis Parameter Generation for the Assistive Silent Speech Interface MVOCA.“ In: *INTERSPEECH*. August. 2011, s. 3009–3012.
- [38] Rabiner, Lawrence R. „A tutorial on hidden Markov models and selected applications in speech recognition“. In: *Proceedings of the IEEE* 77.2 (1989), s. 257–286.
- [39] Alumäe, Tanel. „Neural network phone duration model for speech recognition“. In: *Interspeech 2014*. Singapore, 2014.
- [40] Pylkkonen, Janne a Kurimo, Mikko. „Duration modeling techniques for continuous speech recognition“. In: *Eighth International Conference on Spoken Language Processing*. 2004.
- [41] Russell, Martin a Moore, Roger. „Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition“. In: *ICASSP'85. IEEE International Conference on Acoustics, Speech, and Signal Processing*. Sv. 10. IEEE. 1985, s. 5–8.

- [42] Yoshimura, Takayoshi et al. „Duration modeling for HMM-based speech synthesis“. In: *Fifth International Conference on Spoken Language Processing*. 1998.
- [43] Hadian, Hossein et al. „Phone Duration Modeling for LVCSR Using Neural Networks.“ In: *INTERSPEECH*. 2017, s. 518–522.
- [44] Mertl, Jiří, Žáčková, Eva a Řepová, Barbora. „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](https://doi.org/10.1080/17483107.2017.1319428). eprint: <https://doi.org/10.1080/17483107.2017.1319428>.
- [45] Jůzová, Markéta, Romportl, Jan a Tihelka, Daniel. „Speech Corpus Preparation for Voice Banking of Laryngectomised Patients“. In: *Text, Speech, and Dialogue*. Ed. Král, Pavel a Matoušek, Václav. Cham: Springer International Publishing, 2015, s. 282–290. ISBN: 978-3-319-24033-6.
- [46] Jůzová, Markéta et al. „Voice Conservation and TTS System for People Facing Total Laryngectomy“. In: *Proc. Interspeech 2017*. 2017, s. 3425–3426.