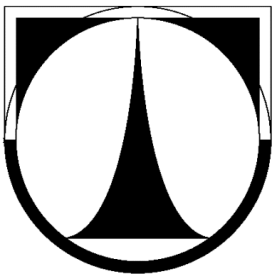
TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových
inženýrských studií



ŘÍZENÁ A NEŘÍZENÁ ADAPTACE NA MLUVČÍHO V SYSTÉMECH ROZPOZNÁVÁNÍ ŘEČI

DISERTAČNÍ PRÁCE

2007 PETR ČERVA

ŘÍZENÁ A NEŘÍZENÁ ADAPTACE NA MLUVČÍHO V SYSTÉMECH ROZPOZNÁVÁNÍ ŘEČI

DISERTAČNÍ PRÁCE

Disertant:	Petr Červa
Studijní program:	2612V Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Tématický okruh:	Počítačové zpracování řeči
Pracoviště:	Ústav informačních technologií a elektroniky Fakulta mechatroniky a mezioborových inženýrských studií Technická univerzita v Liberci
Školitel:	Prof. Ing. Jan Nouza, CSc.

ROZSAH PRÁCE:

Počet stran:	105
Počet obrázků:	7
Počet tabulek:	21
Počet příloh:	1

Prohlášení

Tuto práci jsem vypracoval samostatně s využitím uvedené literatury a na základě konzultací se svým školitelem.

V Liberci dne 17. prosince 2007 Petr Červa

Poděkování

Rád bych poděkoval panu Prof. Ing. Janu Nouzovi, CSc. za jeho pomoc, ochotu a čas věnovaný mi během celého doktorského studia a dále rodině, rodičům a všem kolegům a kolegyním, bez jejichž podpory by tato disertační práce nemohla vzniknout.

Anotace

Disertační práce se zabývá problematikou řízené a neřízené adaptace na mluvčího v systémech rozpoznávání řeči.

Po krátké úvodní části, věnované vysvětlení základních pojmů, je popsán současný stav v řešení problematiky adaptace na mluvčího ve světě a v ČR. Dále jsou vysvětleny motivace pro použití adaptace v systémech vyvíjených na TUL a na základě toho stanoveny cíle práce.

Následně je pozornost věnována základním principům technik používaných pro modelování řeči metodou skrytých Markovových modelů a poté jsou shrnuty základní principy nejčastěji používaných adaptačních metod. Důraz je přitom kladen na ty postupy a metody, které byly využity a dále rozpracovány v rámci této práce.

Obsahem další části jsou pak praktické aspekty adaptace na mluvčího, jehož identita je v době rozpoznávání jeho promluvy známa. Pro tento účel je navrženo a experimentálně ověřeno několik postupů řízené adaptace, které lze prakticky aplikovat v různých systémech vyvinutých pro češtinu na TUL. Jedná se o systémy, které jsou dlouhodobě používány jednou konkrétní osobou (například diktovací program nebo program pro hlasové ovládání PC). Zároveň je vytvořen vlastní adaptační software, který nemá, narozdíl od podobných programů používaných pro tuto úlohu ve většině laboratoří, žádné licenční omezení a je možné ho s uvedenými systémy distribuovat.

Následující kapitola je věnována popisu a experimentálnímu ověření neřízené metody, která je navržena pro účely adaptace v aplikacích, kde není v době rozpoznávání řeči identita mluvčích známa a je velice obtížné ji zjistit automaticky. Jedná se například o úlohu přepisu parlamentních debat či zpravodajských pořadů.

Poslední závěrečná kapitola pak shrnuje všechny dosažené výsledky.

Annotation

The dissertation thesis deals with both supervised and unsupervised speaker adaptation methods in speech recognition systems.

The state of the art in the speaker adaptation task is described in the first part of the work after a short introduction, which explains the basic terms. The following section then summarizes the main motivations for the use of adaptation in systems that are being developed at the Technical University of Liberec (TUL) and after that, the key goals of this work are pointed out.

The second part deals with practical aspects of supervised adaptation on speakers, whose identity is known in the time when their speech is recognized. For this purpose, several practical approaches are proposed and experimentally tested. These can be used in various systems developed for Czech at TUL, which are used by one person on the long-term basis (like dictation system or program allowing voice control of PC). Moreover, an own adaptation software is created for these systems, which does not have, in contrast to systems that are used in most laboratories all around the world, any licence restrictions.

The next part is devoted to the description and experimental verification of an unsupervised method, that is proposed for adaptation in applications (like transcription of parliament debates or broadcast news), where the identity of the speaking person is now known in the time of speech recognition and it is very difficult to determine it automatically.

The last chapter then concludes the work and summarizes all reached results.

Obsah

Seznam obrázků

Seznam tabulek

Seznam zkratek

SI ..... Speaker Independent
.. .. nezávislý na mluvčím
SD ..... Speaker Dependent
.. .. závislý na mluvčím
SA ..... Speaker Adapted
.. .. adaptovaný na mluvčího
GD ..... Gender Dependent
.. .. závislý na pohlaví
IWSR ..... Isolated-Word Speech Recognition
.. .. rozpoznávání izolovaných slov
CSR ..... Continuous Speech Recognition
.. .. rozpoznávání plynulé řeči
BNT ..... Broadcast News Transcription
.. .. přepis zpravodajských pořadů
MLE ..... Maximum Likelihood Estimation
.. .. maximálně věrohodný odhad
SVD ..... Singular Value Decomposition
.. .. metoda singulárního rozkladu
DTW ..... Dynamic Time Warping
.. .. dynamické borcení času
GMM ..... Gaussian Mixture Model
.. .. gaussovský mixturový model
HMM ..... Hidden Markov Model
.. .. skrytý Markovův model
MFCC ..... Mel-Frequency Cepstral Coefficients
.. .. melovské frekvenční kepstrální koeficienty
MAP ..... Maximum A Posteriori
.. .. maximální aposteriorní pravděpodobnost
MLLR ..... Maximum Likelihood Linear Regression
.. .. maximálně věrohodná lineární regrese
FSA ..... Feature Space Adaptation
.. .. adaptace v prostoru příznaků

RMP ..... Regression based Model Prediction
.. .. predikce modelů založená na regresi
WNR ..... Weighted Neighbor Regression
.. .. regrese s vážením sousedů
SMAP ..... Structural MAP
.. .. strukturální MAP
MAPLR ..... Maximum A Posteriori Linear Regression
.. .. maximálně aposteriorní lineární regrese
EV ..... EigenVoices
.. .. vlastní hlasy
VTLN ..... Vocal Tract Length Normalization
.. .. normalizace délky řečového traktu
SST ..... Speaker Selection Training
.. .. trénování s výběrem řečníků
SAT ..... Speaker Adaptive Training
.. .. trénování s adaptací na řečníky
CAT ..... Cluster Adaptive Training
.. .. trénování s adaptací na skupiny řečníků

KAPITOLA 1

ÚVOD

Přepis různých typů mluvených záznamů do textové podoby je jednou z nejaktuálnějších úloh současného výzkumu v oblasti počítačového zpracování řeči. Intenzivní rozvoj této vědní disciplíny v několika posledních letech souvisí se stále rostoucí potřebou naší společnosti mít přístup k co největšímu množství informací, které jsou velmi často uchovávány právě ve formě zvukových záznamů, neboť nejpřirozenější formou lidské komunikace je řeč.

Kromě ve světě již poměrně rozšířených systémů pro hlasové diktování do počítače nebo přepis záznamů z diktafonu, jsou tak stále častěji vyvíjeny také systémy mnohem komplexnější, které umožňují převádět do textové podoby rozsáhlé databáze zvukových dat nebo přepisovat televizní a rozhlasové pořady. Jejich textový výstup pak umožňuje snadné vyhledávání a třídění informací či detekci klíčových slov. V současnosti jsou proto vyvíjeny pro většinu světových jazyků, například angličtinu [NGU05], němčinu [McTait05], francouzštinu [Boulianne06] či čínštinu [Diany05].

Všechny výše zmíněné typy systémů obsahují celou řadu modulů, které postupně zpracovávají vstupní zvukový záznam na různých úrovních, počínaje parametrizací signálu a konče finální úpravou rozpoznaného textu do požadovaného formátu, přičemž klíčovým modulem je vždy rozpoznávač řeči. Moderní rozpoznávače řeči jsou přitom založeny na principu statistického modelovaní akustického signálu a daného jazyka.

V rámci akustického modelování se v naprosté většině případů využívají skryté Markovovy modely. Jejich parametry jsou optimalizovány v době trénování systému tak, aby statisticky co nejlépe vystihovaly charakteristiku promluv obsažených v trénovací databázi. Protože řečové charakteristiky různých mluvčích jsou více či méně odlišné, v závislosti na jejich pohlaví, věku, dialektu či řečnickém stylu, dosahuje každý rozpoznávací systém nejlepších výsledků pouze pro mluvčí a na datech, jejichž charakteristika odpovídá použité trénovací množině. Rozpoznávání navíc komplikuje i skutečnost, že různé promluvy jednoho konkrétního mluvčího se liší i vzájemně zejména různou úrovní šumů a hluků na pozadí (typicky například v úloze přepisu televizních zpráv), která je způsobena prostředím, v němž mluvčí promluvu pronáší.

Aby se předešlo horším výsledkům rozpoznávání pro některé mluvčí a zvýšila se robustnost systému, je akustický model obvykle natrénován jako na mluvčím nezávislý (speaker independent - SI). Pro jeho trénování je použito velké množ-

ství různorodých promluv s velkou variabilitou mluvčích. Právě tato skutečnost ovšem zároveň komplikuje praktické nasazení každého systému, neboť limituje jeho úspěšnost díky tomu, že obecný akustický model garantuje pro každého mluvčího pouze průměrné výsledky.

První logicky se nabízející možností, jak zlepšit výsledky rozpoznávání pro jednoho konkrétního mluvčího, je natrénovat systém jako závislý na mluvčím (speaker dependent - SD) pouze použitím promluv od tohoto mluvčího. Velkou výhodou uvedeného řešení je skutečnost, že takto vytvořený systém dává při rozpoznávání pro mluvčího, jemuž je určen, teoreticky nejlepší možné výsledky. Rozhodující nevýhodou při tvorbě SD systému je ovšem nutnost získat od daného mluvčího pro trénování velké množství promluv (typicky několik hodin), které navíc musí splňovat řadu speciálních požadavků, a z tohoto důvodu je obtížné v praxi SD systém vytvořit. Stejný postup se stejnou zásadní nevýhodou lze aplikovat i při nutnosti vytvořit systém co nejlépe fungující pro jednu konkrétní úlohu, například přepis jednoho konkrétního typu televizního pořadu.

Daleko lepší možností jak zvýšit úspěšnost rozpoznávání pro jednoho konkrétního mluvčího, ať už například uživatele diktovacího systému či osobu často se vyskytující v daném televizním pořadu, je adaptovat (přizpůsobit) některé parametry SI systému na daného mluvčího a vytvořit tak systém na něj adaptovaný (speaker adapted - SA). Právě problematikou adaptace na konkrétního mluvčího se zabývá tato disertační práce, neboť klíčovou výhodou adaptace je skutečnost, že systém s adaptovanými parametry může konvergovat k přesnosti SD systému při použití výrazně menšího množství trénovacích promluv. Úspěšnost rozpoznávání může být adaptací v závislosti na použité metodě významně zvýšena už při použití několika promluv - v extrémním případě pouze jedné. Při adaptaci SI systému na mluvčího se navíc parametry modelů zároveň adaptují i na konkrétní použitý mikrofon, zvukovou kartu počítače a také na šum prostředí, v kterém mluvčí v danou chvíli hovoří. V současné době se proto bez nějaké formy adaptace neobejde žádný komerční systém pro rozpoznávání řeči.

Tato disertační práce je strukturována následujícím způsobem: V kapitole 2 je uveden současný stav v řešení problematiky adaptace na mluvčího ve světě a v České republice, jsou popsány hlavní motivace pro použití metod adaptace v rámci Laboratoře počítačového zpracování řeči na TUL a na základě toho stanoveny cíle této disertační práce. Následující kapitola 3 se zabývá problematikou modelování řeči metodou skrytých Markovových modelů a kapitola 4 potom teoretickým členěním a rozborem nejčastěji používaných adaptačních metod. Pátá kapitola následně krátce popisuje míry používané pro hodnocení úspěšnosti automatického rozpoznávání řeči a adaptace na mluvčího. V pořadí šestá kapitola je pak věnována návrhům praktického řešení pro úlohu adaptace na mluvčího, jehož identita je v době rozpoznávání jeho promluvy známa. Náplní kapitoly 7 je podrobný popis a experimentální vyhodnocení vlastní metody neřízené adaptace, která byla navržena pro systémy umožňující přepisovat zvukové záznamy (například zpravodajské pořady) namluvené mluvčími, jejichž identita je v době zpracování jejich promluvy neznámá.

KAPITOLA 2

SOUČASNÝ STAV PROBLEMATIKY, MOTIVACE A CÍLE DISERTAČNÍ PRÁCE

Před popsáním současného stavu problematiky ve světě a v České republice je třeba nejprve uvést základní členění adaptačních metod z hlediska této disertační práce, a to dle znalosti (správného) textového přepisu promluvy určené pro adaptaci. Podle tohoto kritéria rozlišujeme dva základní typy adaptace na mluvčího:

- *Řízená adaptace, též adaptace s učitelem (supervised adaptation)*
  K dispozici je fonetický přepis promluvy, který je vytvořený nejčastěji člověkem a tudíž v principu správný.

- *Neřízená adaptace, či adaptace bez učitele (unsupervised adaptation)*
  Fonetický přepis promluvy k dispozici není, ale lze ho vytvořit automaticky pomocí rozpoznávače řeči. Následkem toho může ovšem obsahovat více chyb.

2.1 Současný stav problematiky ve světě

2.1.1 Úloha řízené adaptace

Prvně jmenovaná úloha řízené adaptace je přirozeně jednodušší a v literatuře lze nalézt řadu různých metod, které se v současné době ve světě pro tento typ adaptace používají. Tyto metody jsou podrobně popsány v kapitole 4 a nachází své uplatnění zejména v systémech, které jsou dlouhodobě užívány jedním uživatelem a kde lze od tohoto uživatele získat promluvy, jejichž textový přepis je znám či předem připraven. Typicky se jedná o diktovací systémy nebo systémy pro přepis záznamů z diktafonu či počítače. Jednotlivé metody se přitom od sebe liší kromě svého principu zejména podle množství potřebných adaptačních dat.

Za základní a klasickou adaptační techniku lze dnes zřejmě považovat metodu MAP (Maximum A Posteriori - maximální aposteriorní pravděpodobnosti) [Gauvain04]. Její výhodou je konvergence k teoreticky nejpřesnějšímu SD

modelu, nevýhodou naopak nízká účinnost při menším množství adaptačních dat, kdy zůstávají některé parametry SA modelů nedotrénované.

Druhou třídu technik tvoří metody založené na lineární regresi, které se snaží transformovat parametry původních modelů tak, aby nové adaptované modely více odpovídaly charakteristikám daného mluvčího. Jejich typickým představitelem je metoda MLLR (Maximum Likelihood Linear Regression - maximálně věrohodné lineární regrese) [Leggetter95], [Matsoukas97]. Její největší výhoda spočívá ve zvýšení rychlosti adaptace, neboť jedna transformace může být v principu použita najednou pro několik akusticky blízkých Gaussových komponent různých stavů různých modelů, které tvoří jednu regresní třídu.

Třetí významnou a v současné době asi nejmodernější skupinu představují techniky vyvinuté pro práci s extrémně malým množstvím adaptačních dat, které jsou založené na *shlukování* respektive *klastrování* (modelů) *mluvčích* (z anglického *speaker clustering*). Jejich typickým představitelem je metoda označovaná jako EV (EigenVoices - vlastní hlasy) [Kuhn96] či metoda SST (Speaker Selection Training - trénování s výběrem řečníka) [Padmanabhan98].

Poslední čtvrtou skupinu tvoří techniky tzv. *„normalizace dle mluvčího“* (z anglického *speaker normalization*). Na rozdíl od předchozích postupů, které měnily parametry akustického modelu, pracují tyto metody většinou přímo s příznakovými vektory signálu. Typickým představitelem je metoda VTLN (Vocal Tract Length Normalization - normalizace délky řečového traktu) [Zhan97] využívající skutečnost, že rozdíly v hlasových charakteristikách jednotlivých mluvčích jsou kromě jiného způsobeny i odlišnou délkou jejich hlasového traktu.

Podrobný popis a rozbor všech výše uvedených typů metod je obsahem kapitoly 4.

2.1.2 Úloha neřízené adaptace

V úloze neřízené adaptace lze z výše uvedených metod obecně použít přístupy založené na lineární transformaci (MLLR) či normalizaci mluvčího (např. VTLN), které umožňují dosáhnout zajímavého zlepšení rozpoznávacího skóre při použití menšího množství adaptačních dat. Fonetický přepis promluvy musí být ovšem v tomto případě většinou nejprve vytvořen rozpoznávačem řeči a proces rozpoznávání je tím pádem víceprůchodový.

V případě, že je k dispozici pouze extrémně malé množství adaptačních dat (např. pouze jedna promluva) je výhodné použít některou z metod založených na shlukování mluvčích, například metodu STT [Padmanabhan98]. Tato metoda je založena na použití množiny SD modelů, které jsou připraveny předem ve fázi trénovaní systému pro skupinu *referenčních* mluvčích. Pro každého neznámého mluvčího, na nějž je prováděna adaptace, je pak nalezena podmnožina *N* referenčních mluvčích, kteří mají podobné řečové charakteristiky jako neznámý mluvčí, a adaptovaný model je vytvořen kombinací těchto vybraných modelů. Jednotlivé modely referenčních mluvčích přitom bývají z důvodu nedostatku dat často vytvořeny některou z klasických metod pro řízenou adaptaci.

2.2 Současný stav problematiky v ČR

Pro češtinu bylo zatím, kromě vlastních prací autora této práce, publikováno jen několik málo článků (například [Hajek96]), které se zabývaly adaptací na mluvčího a dále jedna disertační práce [Železný01], spolu s několika dalšími souvisejícími články, která se zabývala metodami adaptace systémů pro rozpoznávání spojité řeči. V rámci ní byla pomocí existujícího softwaru [Young00] realizovaná adaptace rozpoznávače spojité řeči metodou MAP a nad ní pak navržena a implementována nadstavbová metoda svazování parametrů.

2.3 Motivace a cíle disertační práce

V rámci Laboratoře počítačového zpracování řeči na TUL je vyvíjeno několik systémů, v nichž najdou metody adaptace své uplatnění. Jedná se například o systém MyVoice pro hlasové ovládání počítače [Nouza05-1], kde je adaptace potřebná z toho důvodu, že motoricky hendikepovaní lidé jsou často postiženi i vadou řeči. Další aplikace zahrnují systém hlasového diktátu do počítače [Nouza05], systém pro přepis nahrávek z diktafonů a komplexní systém pro přepis televizních a rozhlasových pořadů [Nouza06]. S ohledem na výše uvedené skutečnosti byly stanoveny následující cíle disertační práce:

- Prozkoumat a uceleným způsobem popsat principy nejčastěji používaných adaptačních metod.
- Modifikovat již existující metody popřípadě najít vhodný praktický postup, který by umožňoval provádět efektivní řízenou adaptaci v systémech dlouhodobě používaných jedním konkrétním uživatelem. Jedná se například o diktovací systémy či systém hlasového ovládání PC.
- Vytvořit pro tento účel prakticky použitelný software, který by mohl být distribuován spolu s cílovými aplikacemi.
- Navrhnout vlastní metodu, která by umožňovala provádět efektivní neřízenou adaptaci v systému pro přepis televizních a rozhlasových pořadů a tuto metodu implementovat.
- Experimentálně vyhodnotit úspěšnost všech použitých a navržených postupů na různých typech dat a v různých úlohách a systémech.

KAPITOLA 3

ZÁKLADNÍ PRINCIPY MODELOVÁNÍ ŘEČI METODOU HMM

Cílem této kapitoly je popsat základní principy modelování řeči metodou skrytých Markovových modelů. Pozornost je přitom zaměřena zejména na ty postupy a metody, které se často používají v úloze adaptace akustického modelu na konkrétního mluvčího a které byly využity a rozpracovány v rámci této disertační práce. Detailní vysvětlení uvedených i dalších aspektů problematiky akustického modelování řeči lze najít v například v [Huang01], [Huang90] nebo v [Psutka06].

3.1 Reprezentace řečových jednotek

Při analýze a rozpoznávání řeči je akustický signál nejprve rozdělen do krátkých časových úseků, které budou dále označovány jako *rámce* (z anglického *frame*), kde se jeho parametry mění jen málo a kde ho lze považovat za stacionární. Pro každý rámec je následně vypočítána sada parametrů, z nichž je sestaven příznakový vektor. Sekvence příznakových vektorů je pak porovnávána s modely popisujícími akustickou, lexikální a jazykovou složku řeči. V této práci je hlavní pozornost zaměřena na akustické modelování řeči, pro které se dnes nejčastěji používají takzvané skryté Markovovy modely (Hidden Markov Models - HMMs).

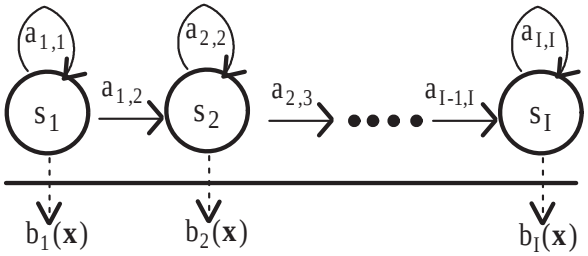
Skryté Markovovy modely představují speciální případ obecnějších Markovových modelů, které patří do kategorie pravděpodobnostních stavových modelů (konečných automatů) a které jsou široce používané pro modelování procesů majících takzvanou Markovovu vlastnost. Tu lze slovně vyjádřit podmínkou, že současný stav modelu daného procesu, respektive jeho pravděpodobnostní rozložení, závisí pouze na n stavech předchozích. Podle hodnoty n se pak rozlišují Markovovy modely n-tého řádu. Každý Markovův model je přitom charakterizován pouze pravděpodobnostmi přechodů mezi jednotlivými stavy uvnitř modelu a pravděpodobnosti přechodů jsou jeho jedinými parametry. Výstupem z Markovova modelu je tak přímo posloupnost stavů.

Skryté Markovovy modely jsou naproti tomu specifické tím, že z vygenerované

výstupní posloupnosti symbolů (dat) nelze zpětně určit, kterými stavy proces prošel. Posloupnost stavů tedy zůstává skryta, neboť každý *stav modelu je charakterizován pravděpodobnostním rozložením nad množinou všech možných výstupních hodnot*. Lze tak například pouze vypočítat, s jakou věrohodností byla výstupní data vygenerována konkrétní posloupností stavů.

Struktura skrytých Markovových modelů prvního řádu, která je nejčastěji používaná pro rozpoznávání řeči, je znázorněna na obr. 3.1. Další používané struktury lze najít například v [Psutka06]. Každý model je tvořen posloupností celkem *I* stavů, které reprezentují buď stacionární úseky řečového signálu představujícího jedno celé slovo, pak mluvíme o celoslovních modelech, nebo menší řečovou jednotku - nejčastěji konkrétní foném daného jazyka. Celoslovní Markovovy modely přitom mívají nejčastěji kolem šesti až dvanácti stavů, při modelování fonémů vystačíme s menším počtem stavů, typicky se třemi.

Přecházet mezi stavy uvnitř modelu lze jen zleva doprava (levopravý model) a nejčastěji jen mezi dvěma stavy sousedními, což dobře vystihuje skutečnost, že řeč plyne postupně s rostoucím časem. Možnost setrvání v daném stavu je na obr. 3.1 znázorněna smyčkou.



Obrázek 3.1: Typická struktura skrytého Markovova modelu používaná pro modelování řečových jednotek.

V současné době se téměř výhradně používají modely se spojitou výstupní pravděpodobnostní hustotou(Continuous Density HMM - CDHMM). Tato funkce je většinou dána vícerozměrným Gaussovým rozložením, podle počtu příznaků počítaných ze signálu řeči, a dále bude označována symbolem b_i. Pro *P*-rozměrný příznakový vektor a vícemodální Gaussovo rozložení s celkem *M* komponentami má tedy funkce b_i tvar

$$b_i(\mathbf{x}) = \sum_{m=1}^{M} c_{im} b_{im}(\mathbf{x}), \tag{3.1}$$

kde

$$b_{im}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^P |\mathbf{\Sigma}_{im}|}} exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{im})' \mathbf{\Sigma}_{im}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{im}) \Big) \tag{3.2}$$

je pravděpodobnostní hustota a c_{im} váhový koeficient m-té komponenty stavu i (přičemž $\sum_{m=1}^{M} c_{im} = 1$). Vztah 3.1 vyjadřuje míru pravděpodobnosti, že příznakový vektor $\mathbf{x}$, jeden konkrétní rámec promluvy, byl vygenerován právě stavem i.

První rámec řečového signálu přitom vždy musí být přiřazen prvnímu stavu modelu a poslední rámec signálu poslednímu stavu I. Symbol $\mathbf{\Sigma}_{im}$ značí kovariační matici a $\boldsymbol{\mu}_{im}$ je vektor středních hodnot m-té komponenty stavu i.

Označíme-li dále symbolem $\mathbf{\Phi}$ množinu všech parametrů daného Markovova modelu a je-li celá modelovaná promluva $\mathbf{X}$ reprezentována časovou posloupností celkem T vektorů příznaků $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$, pak výraz $p(\mathbf{X}|\mathbf{\Phi})$ vyjadřuje míru pravděpodobnosti respektive věrohodnost, že promluva X byla vygenerována modelem s parametry $\mathbf{\Phi}$. Výpočet této věrohodnosti je popsán v kapitole 3.2 a její využití pro estimaci parametrů HMM v kapitole 3.3.

Kromě parametrů funkce b_i se dá ve fázi trénování (kapitola 3.4) také statisticky vyhodnotit, kolik rámců signálu řeči daný stav představuje a na základě toho určit pravděpodobnosti setrvání a přechodů mezi jednotlivými stavy modelu. Pravděpodobnost přechodu ze stavu i do následujícího stavu $i + 1$ je dána hodnotou $a_{i,i+1}$, pravděpodobnost setrvání v daném stavu hodnotou $a_{i,i}$. Protože oba jevy jsou komplementární, platí $a_{i,i} + a_{i,i+1} = 1$.

3.2 Metody výpočtu věrohodnosti vygenerování dat

Zcela intuitivní možnost jak vypočítat věrohodnost toho, že daná sekvence dat $\mathbf{X}$ byla vygenerována modelem s parametry $\mathbf{\Phi}$, je určit všechny možné posloupnosti stavů $\mathbf{S}$ o délce T, kterými mohl model při generování dat projít, a následně sečíst jednotlivé věrohodnosti odpovídající tomu, že právě konkrétní sekvence vygenerovala uvažovaná data. Označíme-li množinu všech možných posloupností stavů o délce T symbolem Ψ, lze $p(\mathbf{X}|\mathbf{\Phi})$ vypočítat dle rovnice

$$p(\mathbf{X}|\mathbf{\Phi}) = \sum_{\mathbf{S} \in \mathbf{\Psi}} p(\mathbf{X}, \mathbf{S}|\mathbf{\Phi}). \tag{3.3}$$

Za předpokladu, že uvažovaný model má Markovovu vlastnost a že vektory dat jsou statisticky nezávislé, lze pravděpodobnostní hustotu $p(\mathbf{X}, \mathbf{S}|\mathbf{\Phi})$ vyjádřit jako

$$p(\mathbf{X}, \mathbf{S}|\mathbf{\Phi}) = b_{s_1}(\mathbf{x}_1) a_{s_1,s_2} b_{s_2}(\mathbf{x}_2) a_{s_2,s_3} ... a_{s_{T-1},s_T} b_{s_T}(\mathbf{x}_T). \tag{3.4}$$

Výpočet vztahu 3.4 má exponenciální složitost, kterou lze snížit tak, že jsou v jeho průběhu ukládány mezivýsledky, které jsou poté používány pro všechny posloupnosti stavů se stejným počátečním pořadím. Tento postup pak bývá označován jako Forward algoritmus a je vysvětlen v následující kapitole.

3.2.1 Forward algoritmus

Forward algoritmus je ve své podstatě algoritmus rekurzivní, který využívá Markovovu vlastnost, že výpočet součinu $p(s_t|s_{t-1}, \mathbf{\Phi})p(\mathbf{x}_t|s_t = i, \mathbf{\Phi})$ závisí pouze na stavu s_{t-1}, stavu s_t a hodnotě $\mathbf{x}_t$. Pro jeho vysvětlení je nejprve třeba definovat takzvanou *dopřednou* proměnnou $\alpha_t(i)$ (z anglického *forward*), která vyjadřuje míru pravděpodobnosti, že se daný model s parametry $\mathbf{\Phi}$ o celkem I stavech nachází

v čase t ve stavu i a při cestě do tohoto stavu vygeneroval částečnou posloupnost vektorů příznaků $\mathbf{X}_1^t = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t\}$. $\alpha_t(i)$ je definována rovnicí

$$\alpha_t(i) = p(\mathbf{X}_1^t, s_t = i|\mathbf{\Phi}). \tag{3.5}$$

Z definice dopředné proměnné a skutečnosti, že poslední rámec signálu musí být přiřazen k poslednímu stavu modelu, vyplývá že $p(\mathbf{X}|\mathbf{\Phi}) = \alpha_T(I)$. Právě postup výpočtu dopředné proměnné pak bývá označován jako Forward algoritmus, který lze pro uvažované levopravé modely (viz obr. 3.1) přepsat do následujícího tvaru:

Forward algoritmus

krok1: *inicializace*

$\alpha_1(1) = b_1(\mathbf{x}_1)$

$\alpha_1(i) = 0 \qquad i = 2, ..., I$

krok2: *indukce*

$\alpha_t(1) = \alpha_{t-1}(1)a_{1,1}b_1(\mathbf{x}_t) \qquad t = 2, ..., T$

$\alpha_t(i) = \Big(\sum_{j=i-1}^{i} \alpha_{t-1}(j)a_{i,j}\Big)b_i(\mathbf{x}_t) \qquad t = 2, ..., T;\ i = 2, ..., I$

krok3: *ukončení*

$p(\mathbf{X}|\mathbf{\Phi}) = \alpha_T(I)$

3.2.2 Backward algoritmus

Při výpočtu věrohodnosti vygenerování dat modelem je možné použít i obrácený postup výpočtu směrem od posledního k prvnímu stavu. Podobně jako $\alpha_t(i)$ lze definovat i takzvanou *zpětnou* proměnnou $\beta_t(i)$ (z anglického *backward*) vyjadřující míru pravděpodobnosti, že daný model s parametry $\mathbf{\Phi}$ nacházející se v čase t ve stavu i, přejde v čase od $t+1$ do T postupně až do koncového stavu I a vygeneruje přitom částečnou posloupnost příznakových vektorů $\mathbf{X}_{t+1}^T = \{\mathbf{x}_{t+1}, \mathbf{x}_{t+2}, ..., \mathbf{x}_T\}$. Tuto proměnnou lze definovat jako

$$\beta_t(i) = p(\mathbf{X}_{t+1}^T, s_t = i|\mathbf{\Phi}) \tag{3.6}$$

a její výpočet může být proveden pomocí Backward algoritmu:

Backward algoritmus

krok1: *inicializace*

$\beta_T(I) = 1$

$\beta_T(i) = 0 \qquad i = 1, ..., I-1$

krok2: *indukce*

$\beta_t(I) = a_{I,I}b_I(\mathbf{x}_{t+1})\beta_{t+1}(I) \qquad t = T-1, ..., 1$

$\beta_t(i) = \sum_{j=i-1}^{i} a_{i,j}b_j(\mathbf{x}_{t+1})\beta_{t+1}(j) \qquad t = T-1, ..., 1;\ i = 1, ..., I-1$

krok3: *ukončení*

$p(\mathbf{X}|\mathbf{\Phi}) = \beta_1(1)$

3.2.3 Viterbiho algoritmus

V předchozích dvou algoritmech byla vždy vypočítána celková věrohodnost přes všechny možné posloupnosti stavů, že daný model vygeneroval data $\mathbf{X}$. V některých aplikacích, například při klasickém rozpoznávání řeči, ovšem stačí tuto věrohodnost nahradit maximální hodnotou vypočítanou přes všechny přípustné posloupnosti. Tento výpočet je totiž efektivnější. Někdy, například při trénování metodou Viterbiho přiřazení 3.4.1, může být navíc užitečné znát posloupnost stavů, při které je dosaženo maximální věrohodnosti. Tuto maximální věrohodnost a jí odpovídající posloupnost stavu je možné získat pomocí Viterbiho algoritmu [Viterbi67]. Při jeho implementaci je možné zavedením kumulovaného součinu $V(t,i)$ využít metody dynamického programování. Kumulovaný součin je definován jako:

$$V(t,i) = b_i(\mathbf{x}_t) Max\Big(a_{i,i} V(t-1,i), a_{i-1,i} V(t-1,i-1)\Big). \qquad (3.7)$$

Dle [Nouza97] je pak možné Viterbiho algoritmus rozepsat v následujícím tvaru:

Viterbiho algoritmus

krok1: *inicializace*
$V(1,1) = b_1(\mathbf{x}_1), B(1,1)$
$V(1,i) = -\infty \qquad 2 \leq i \leq I$
krok2: *rekurze*
for $t = 2, ..., T$
for $i = 1, ..., I$
pomocná proměnná $P = a_{i,i} V(t-1,i)$
pole zpětných ukazatelů $B(t,i) = i$
if $(I > 1)$
if $(P < a_{i-1,i} V(t-1,i-1))$
$P = a_{i-1,i} V(t-1,i-1)$
$B(t,i) = i - 1$
$V(t,i) = b_i(\mathbf{x}_t) P$
krok3: *ukončení*
$p(\mathbf{X}|\mathbf{\Phi}) = V(T,I)$
krok4: *určení posloupnosti stavů* $\mathbf{S}$
$\mathbf{S}(T) = I$
for $t = T-1, ..., 1$
$\mathbf{S}(t) = B(t+1, \mathbf{S}(t+1))$

3.3 Vybrané metody estimace parametrů

Estimace nebo-li odhad parametrů modelů je jedním z klíčových prostředků adaptace a také trénování. Většina adaptačních technik zaměřených na adaptaci akustického modelu v sobě zahrnuje i některou z níže popsaných estimačních metod.

3.3.1 Estimace metodou ML

Estimace parametrů Markovových modelů metodou maximální věrohodnosti (maximum likelihood estimation - MLE) je pro svoji efektivitu jednou z nejčastěji používaných estimačních metod. Je založena na předpokladu, že hledané optimální parametry modelu $\mathbf{\Phi}$ jsou pevné, respektive jejich pravděpodobnostní rozložení je rovnoměrné, a neznámé hodnoty a snaží se je najít tak, aby byla maximalizována věrohodnost, že daná sekvence dat byla vygenerována právě uvažovaným modelem

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}}(p(\mathbf{X}|\mathbf{\Phi})). \tag{3.8}$$

Protože jednotlivé vektory příznaků jsou nezávislé, lze $p(\mathbf{X}|\mathbf{\Phi})$ vyjádřit jako

$$p(\mathbf{X}|\mathbf{\Phi}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{\Phi})). \tag{3.9}$$

V praxi je pak jednodušší hledat maximum logaritmu věrohodnosti dle rovnice

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}} \sum_{t=1}^{T} log\ p(\mathbf{x}_t|\mathbf{\Phi}). \tag{3.10}$$

Pro modely typu CDHMM s vícemodalní pravděpodobnostní hustotou, které jsou uvažovány v rámci této disertační práce, zatím nebyla nalezena žádná metoda, která by umožnila dosažení globálního maxima této věrohodnosti. Metoda MLE totiž nemůže být pro tyto modely aplikována přímo díky tomu, že nelze konkrétně určit, která sekvence stavů a které komponenty jednotlivých stavů vygenerovaly dané vektory příznaků. Tato informace zůstává skryta a z hlediska teorie estimace jsou tak data $\mathbf{X}$ nekompletní. Pro odhad parametrů se proto používá postup založený na algoritmu EM (expectation-maximization - očekávání-maximalizace) [Dempster77], který uvedenou věrohodnost maximalizuje alespoň lokálně v závislosti na prvotním odhadu parametrů. Tento postup bývá označován jako Baum-Welchův (Forward-Backward) algoritmus.

V rámci tohoto algoritmu je nejprve proveden prvotní odhad parametrů $\mathbf{\Phi}$. Na jeho základě je poté vypočítána věrohodnost, že všechny možné sekvence stavů a jejich jednotlivé komponenty vygenerovaly uvažovaná data. Tímto způsobem jsou vlastně použitá data doplněna o onu chybějící informaci a může být určen maximálně věrohodný odhad nových parametrů $\hat{\mathbf{\Phi}}$. Hodnoty nových parametrů jsou tím ovšem závislé na prvotním odhadu a pokud uvedený postup několikrát iteračně opakujeme, je zajištěna konvergence pouze k lokálnímu maximu věrohodnosti (odvození konvergence viz [Huang01]). Forward-Backward algoritmus lze popsat následovně:

Forward-Backward (EM) algoritmus

krok1: *inicializace*
Jsou vypočteny prvotní odhady parametrů $\mathbf{\Phi}$.

krok2: *výpočet očekávané hodnoty (E-step))*
Pomocí $\mathbf{\Phi}$ je vypočtena pomocná funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$.

krok3: *maximalizace (M-step))*
Výpočet $\hat{\mathbf{\Phi}}$ aby byla maximalizována pomocná funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$.

krok4: *iterační výpočet*
Hodnotě $\mathbf{\Phi}$ je přiřazena $\hat{\mathbf{\Phi}}$ a kroky 2-3 jsou opakovány, dokud algoritmus konverguje.

Pro definování funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$ je nejprve třeba rozšířit vztah 3.4 pro výpočet $p(\mathbf{X}, \mathbf{S}|\mathbf{\Phi})$ s ohledem na skutečnost, že chceme estimovat parametry jednotlivých Gaussových komponent systému

$$p(\mathbf{X}, \mathbf{S}|\mathbf{\Phi}) = \prod_{t=1}^{T} a_{s_{t-1},s_t} b_{s_t}(\mathbf{x}_t) =$$

$$= \prod_{t=1}^{T} a_{s_{t-1},s_t} \Big[\sum_{k=1}^{M} c_{s_t k} b_{s_t k}(\mathbf{x}_t) \Big] = \tag{3.11}$$

$$= \sum_{k_1=1}^{M} \sum_{k_2=1}^{M} ... \sum_{k_T=1}^{M} \Big[\prod_{t=1}^{T} a_{s_{t-1},s_t} c_{s_t k_t} b_{s_t k_t}(\mathbf{x}_t) \Big]. \tag{3.12}$$

Definujeme-li nyní sdruženou pravděpodobnostní hustotu $p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi})$ jako

$$p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi}) = \prod_{t=1}^{T} a_{s_{t-1},s_t} c_{s_t k_t} b_{s_t k_t}(\mathbf{x}_t), \tag{3.13}$$

lze $p(\mathbf{X}|\mathbf{\Phi})$ vypočítat dle vztahu

$$p(\mathbf{X}|\mathbf{\Phi}) = \sum_{\mathbf{S} \in \mathbf{\Psi}} \sum_{\mathbf{K} \in \mathbf{\Omega}^T} p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi}). \tag{3.14}$$

kde $\mathbf{K}$ je produkt T-rozměrného kartézského součinu množiny $\mathbf{\Omega} = \{1, 2, ..., M\}$ v prostoru $\mathbf{\Omega}^T$. Součet přes členy $\mathbf{K}$ a $\mathbf{S}$ vyjadřuje skutečnost, že posloupnost vektorů příznaků o délce T mohla být vygenerována různými posloupnostmi stavů o stejné délce, přičemž každý konkrétní vektor příznaků $\mathbf{x}_t$ mohl být zároveň vygenerován kteroukoli komponentou stavu s_t, v němž se model v čase t nacházel.

Funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$ splňující podmínku konvergence k lokálnímu maximu má pro uvažované modely tvar

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = \sum_{\mathbf{S}\in\mathbf{\Psi}} \sum_{\mathbf{K}\in\mathbf{\Omega}^T} \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi})}{p(\mathbf{X}|\mathbf{\Phi})} log\, p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\hat{\mathbf{\Phi}}), \tag{3.15}$$

kde $log\, p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\hat{\mathbf{\Phi}})$ lze s využitím 3.13 vyjádřit jako

$$log\, p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\hat{\mathbf{\Phi}}) = \sum_{t=1}^{T} log\, \hat{a}_{s_{t-1},s_t} + \sum_{t=1}^{T} log\, \hat{b}_{s_t k_t}(\mathbf{x}_t) + \sum_{t=1}^{T} log\, \hat{c}_{s_t k_t}. \tag{3.16}$$

Dosazením vztahů 3.13 a 3.16 do rovnice 3.15 získáme funkci $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$ v separovaném tvaru. Položíme-li následně její parciální derivace dle jednotlivých parametrů rovny nule, získáme výsledné vztahy pro odhad nových parametrů. Postup výpočtu viz [Psutka06]. Ty mají následující tvar:

pravděpodobnost přechodu:

$$\hat{a}_{i,j} = \frac{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_{t-1}(i), s_t(j)|\mathbf{\Phi})}{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_{t-1}(i)|\mathbf{\Phi})} = \\ = \frac{\sum_{t=1}^{T} \gamma_t(i,j)}{\sum_{t=1}^{T} \sum_{k=1}^{I} \gamma_t(i,k)} \tag{3.17}$$

váha komponenty:

$$\hat{c}_{im} = \frac{\sum_{t=1}^{T} \zeta_t(i,m)}{\sum_{k=1}^{M} \sum_{t=1}^{T} \zeta_t(i,k)} \tag{3.18}$$

vektor středních hodnot:

$$\hat{\boldsymbol{\mu}}_{im} = \frac{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_t(i), m_t = m|\mathbf{\Phi})\mathbf{x}_t}{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_t(i), m_t = m|\mathbf{\Phi})} = \\ = \frac{\sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t}{\sum_{t=1}^{T} \zeta_t(i,m)} \tag{3.19}$$

kovariační matice:

$$\hat{\mathbf{\Sigma}}_{im} = \frac{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_t(i), m_t = m|\mathbf{\Phi})(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{im})(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{im})'}{\frac{1}{p(\mathbf{X}|\mathbf{\Phi})} \sum_{t=1}^{T} p(\mathbf{X}, s_t(i), m_t = m|\mathbf{\Phi})} = \\ = \frac{\sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t(\mathbf{x}_t)'}{\sum_{t=1}^{T} \zeta_t(i,m)} - \hat{\boldsymbol{\mu}}_{im}(\hat{\boldsymbol{\mu}}_{im})' \tag{3.20}$$

Člen $\gamma_t(i,j)$ v uvedených vztazích představuje pravděpodobnostní hustotu přechodu ze stavu i do stavu j v čase t za podmínky, že daný model vygeneroval posloupnost vektorů $\mathbf{X}$. Tuto hustotu je možné vypočítat jako

$$\gamma_t(i,j) = p(s_{t-1}=i, s_t=j|\mathbf{X}_1^t, \mathbf{\Phi}) = \frac{p(s_{t-1}=i, s_t=j, \mathbf{X}_1^t|\mathbf{\Phi})}{p(\mathbf{X}_1^t|\mathbf{\Phi})} = \\ = \frac{\alpha_{t-1}(i)a_{i,j}b_j(\mathbf{x}_t)\beta_t(j)}{\alpha_T(I)}. \quad (3.21)$$

Člen $\zeta_t(i,m)$ má význam tzv. *„okupační věrohodnosti“* (z anglického *occupation likelihood*) m-té komponenty stavu i, kterou lze pro uvažovaný typ Markovových modelů definovat vztahem

$$\zeta_t(i,m) = \frac{p(s_t=i, m_t=m, \mathbf{X}|\mathbf{\Phi})}{p(\mathbf{X}|\mathbf{\Phi})}. \quad (3.22)$$

Člen $\zeta_t(i,m)$ vyjadřuje míru pravděpodobnosti, že model s parametry $\mathbf{\Phi}$, který vygeneroval celou posloupnost příznakových vektorů $\mathbf{X}$, se v čase t nacházel ve stavu i a vektor $\mathbf{x}_t$ byl vygenerován právě m-tou komponentou stavu i. Součet $\sum_{t=1}^{T}\zeta_t(i,m)$ pak představuje míru množství dat použitých pro estimaci parametrů této komponenty. Pomocí pomocí zpětné a dopředné proměnné lze $\zeta_t(i,m)$ vyjádřit rovnicí

$$\zeta_t(i,m) = \frac{\sum_{j=1}^{I}\alpha_{t-1}(j)a_{j,i}c_{im}b_{ik}(\mathbf{x}_t)\beta_t(i)}{\alpha_T(I)}. \quad (3.23)$$

Výsledné vztahy pro odhady nových parametrů lze interpretovat také slovně. Například vztah 3.17 vyjadřující odhad pravděpodobnosti přechodu ze stavu i do stavu j lze interpretovat jako poměr mezi celkovým očekávaným počtem přechodů ze stavu i do stavu j a celkovým očekávaným počtem přechodů ze stavu i do všech možných stavů. Vztah 3.18 pro odhad váhového koeficientu m-té komponenty stavu i lze interpretovat jako poměr mezi celkovou okupačních pravděpodobností této komponenty a celkovou okupační pravděpodobností stavu i. Podobně mohou být interpretovány i vztahy 3.19 a 3.20.

Při praktické implementaci Baum-Welchova algoritmu je nutné provádět škálování dopředných a zpětných proměnných, protože jejich hodnoty pro dostatečně velké T mohou lehce klesnout pod nejmenší možnou hodnotu vyjádřitelnou použitým výpočetním systémem. Dopředná i zpětná proměnná mohou být například násobeny škálovacím koeficientem S_t definovaným dle [Huang01] jako

$$S_t = \frac{1}{\sum_{i=1}^{I}\alpha_t(i)}. \quad (3.24)$$

Protože jsou hodnoty obou proměnných počítány rekurzivně a jednotlivé škálovací koeficienty díky tomu neustále dohromady násobeny, je celkový škálovací faktor

použitý pro výpočet dopředné proměnné roven v čase t hodnotě

$$Scale_t(\alpha) = \prod_{t_1=1}^{t} S_{t_1} \tag{3.25}$$

a obdobně pro výpočet zpětné proměnné hodnotě

$$Scale_t(\beta) = \prod_{t_1=t}^{T} S_{t_1}. \tag{3.26}$$

Škálovanou hodnotu dopředné proměnné $\alpha_T^{Scale}(I)$ lze pak vyjádřit rovnicí

$$\alpha_T^{Scale}(I) = Scale_T(\alpha)p(\mathbf{X}|\mathbf{\Phi}). \tag{3.27}$$

Díky tomu, že během škálování dochází s rostoucím časem k postupnému vzájemnému násobení škálovacích koeficientů, lze všechny výše odvozené vztahy pro odhad parametrů použít při realizaci škálování beze změny. Například škálovaná míra pravděpodobnosti $\gamma_t^{Scale}(i,j)$ je rovna neškálované míře $\gamma_t(i,j)$, protože škálovací faktory v čitateli a jmenovateli se vykrátí

$$\gamma_t^{Scale}(i,j) = \frac{Scale_{t-1}(\alpha)\alpha_{t-1}(i)a_{i,j}b_j(\mathbf{x}_t)\beta_t(j)Scale_t(\beta)}{Scale_T(\alpha)\alpha_T(I)} = \gamma_t(i,j). \tag{3.28}$$

Využití Baum-Welchova algoritmu pro trénování skrytých Markovových modelů je diskutováno v kapitole 3.4.

3.3.2 Estimace metodou MAP

Estimace parametrů metodou maximální aposteriorní pravděpodobnosti (Maximum A Posteriori - MAP) je založena na rozdílném principu než metoda maximální věrohodnosti. Zatímco metoda ML předpokládá, že hledané optimální parametry jsou pevné a neznámé hodnoty, metoda MAP je založena na hypotéze, že hledané parametry jsou náhodné veličiny se známým apriorním rozložením. V praxi se proto metoda MAP používá pro adaptaci parametrů modelů, neboť nedostatek dat při adaptaci je při použití této metody částečně kompenzován informací o apriorním rozložení. Z uvedených faktů také vyplývá, že pokud je apriorní rozložení parametrů rovnoměrné, což znamená, že parametry modelů jsou pevné hodnoty, je metoda MAP identická s metodou ML. V rámci metody MAP lze optimální parametry $\hat{\mathbf{\Phi}}$ najít dle vztahu

$$\hat{\mathbf{\Phi}} = \underset{\mathbf{\Phi}}{\operatorname{argmax}}(p(\mathbf{\Phi}|\mathbf{X})). \tag{3.29}$$

kde $p(\mathbf{\Phi}|\mathbf{X})$ představuje aposteriorní hustotu pravděpodobnosti parametrů $\mathbf{\Phi}$ za podmínky, že daný model vygeneroval posloupnost $\mathbf{X}$. Protože předpokládáme

znalost apriorního rozložení parametru $\mathbf{\Phi}$, použijeme Bayesův teorém a vyjádříme tuto aposteriorní pravděpodobnost jako

$$p(\mathbf{\Phi}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{\Phi})p(\mathbf{\Phi})}{p(\mathbf{X})}, \tag{3.30}$$

kde $p(\mathbf{\Phi})$ představuje apriorní rozložení pravděpodobnosti parametru $\mathbf{\Phi}$, $p(\mathbf{X})$ apriorní rozložení pravděpodobnosti $\mathbf{X}$ a $p(\mathbf{X}|\mathbf{\Phi})$ představuje věrohodnost, že data $\mathbf{X}$ byla vygenerována modelem s parametry $\mathbf{\Phi}$. Maximalizace aposteriorního rozložení pravděpodobnosti je potom dosažena změnou parametru $\mathbf{\Phi}$ tak, aby byl maximalizován výraz $p(\mathbf{X}|\mathbf{\Phi})p(\mathbf{\Phi})$, neboť $p(\mathbf{X})$ je pro všechny možné hodnoty $\mathbf{\Phi}$ konstantní.

Pro modely typu CDHMM lze odvodit vztahy pro odhad nových parametrů podobným způsobem jako v přechozí kapitole u metody ML. Detailní postup viz například [Psutka06]. Tyto vztahy mají následující tvar:

váha komponenty:

$$\hat{c}_{im} = \frac{\upsilon_{im} - 1 + \sum_{t=1}^{T} \zeta_t(i,m)}{\sum_{k=1}^{M} \left(\upsilon_{ik} - 1 + \sum_{t=1}^{T} \zeta_t(i,k)\right)} \tag{3.31}$$

vektor středních hodnot:

$$\hat{\boldsymbol{\mu}}_{im} = \frac{\tau_{im}\boldsymbol{\mu}_{im}^{nw} + \sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t}{\tau_{im} + \sum_{t=1}^{T} \zeta_t(i,m)} \tag{3.32}$$

kovariační matice:

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{im} = &\frac{\mathbf{S}_{im} + \tau_{im}(\hat{\boldsymbol{\mu}}_{im} - \boldsymbol{\mu}_{im}^{nw})(\hat{\boldsymbol{\mu}}_{im} - \boldsymbol{\mu}_{im}^{nw})'}{\eta_{im} - P + \sum_{t=1}^{T} \zeta_t(i,m)} + \\ &+ \frac{\sum_{t=1}^{T} \zeta_t(i,m)(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{im})(\mathbf{x}_t - \hat{\boldsymbol{\mu}}_{im})'}{\eta_{im} - P + \sum_{t=1}^{T} \zeta_t(i,m)}\end{aligned} \tag{3.33}$$

Matice $\mathbf{S}_{im}$, vektor $\boldsymbol{\mu}_{im}^{nw}$ a členy η_{im},τ_{im} v těchto vztazích představují parametry normálního-Wishartova apriorního rozdělení parametrů m-té komponenty stavu i, P je dimenze matice $\mathbf{\Sigma}$ (počet příznaků) a člen υ_{im} reprezentuje parametr Dirichletova apriorního rozložení pro váhu m-té komponenty stavu i. Určení těchto takzvaných „hyperparametrů“ představuje v rámci metody MAP teoreticky asi nejsložitější problém (viz například [Psutka06]).

Zřejmě nejjednodušší možnost, jak tyto parametry odhadnout, je použít přímo parametry vhodných, již existujících modelů, které byly natrénovány na datech s podobnou charakteristikou. Tento postup se používá velmi často pro účely adaptace na mluvčího (viz kapitola 4.2), kdy jsou za některé hyperparametry dosazeny přímo parametry původního neadaptovaného modelu a zbylé mají funkci volitelné adaptační váhy. V praxi se navíc díky malému množství dostupných dat adaptují nejčastěji pouze vektory středních hodnot a výpočet ostatních hyperparametrů pro

účely adaptace zbylých parametrů tak ztrácí ještě více na významu. Přesněji lze hyperparametry určit tím způsobem, že se nejprve rozdělí všechna dostupná trénovací data na několik skupin, například dle jednotlivých řečníků či přenosového kanálu, a následně se pro každou skupinu vypočtou jednotlivé parametry modelů, které tím pádem představují konkrétní realizace z předpokládaných apriorních rozložení. Na základě nich se nakonec odhadnou hodnoty hyperparametrů.

Odvozené vztahy 3.31 - 3.33 pro odhady parametrů metodou MAP mají význam váženého součtu mezi hodnotami apriorních parametrů a hodnotami parametrů odhadnutých metodou ML z dat použitých pro estimaci. Vliv parametrů odhadnutých metodou ML se přitom zvyšuje s rostoucím množstvím dat. Pro $\sum_{t=1}^{T} \zeta_t(i,m) \rightarrow \infty$ je odhad parametrů metodou MAP totožný s odhadem parametrů metodou ML.

3.4 Trénování modelů pro rozpoznávání

3.4.1 Trénování celoslovních modelů

Pro jednoduchost uvažujme nejprve systém pracující s celoslovními Markovovými modely, které obsahují pouze jednu Gaussovu komponentu v každém stavu. Nejjednodušší metodou trénování parametrů je postup založený na přiřazení všech rámců trénovaného slova ke stavům jeho modelu pomocí Viterbiho algoritmu:

Postup trénování metodou Viterbiho přiřazení

krok1: *vytvoření prvotních odhadů parametrů*

Rámce všech realizací uvažovaného slova (celkem R realizací) jsou rovnoměrně přiřazeny ke stavům trénovaného modelu. K i-tému stavu modelu je tak přiřazeno celkem N_i rámců (označených symbolem $\mathbf{x}^{in}$) a podle následujících vztahů se provede prvotní odhad parametrů:
vektor středních hodnot: $\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{n=1}^{N_i}\mathbf{x}^{in}$
vektor rozptylů (diagonální kovariační matice): $\boldsymbol{\sigma}_i^2 = \frac{1}{N_i}\sum_{n=1}^{N_i}(\mathbf{x}^{in}-\boldsymbol{\mu}_i)^2$
pravděpodobnosti přechodů: $a_{i,i+1} = \frac{R}{N_i}$
pravděpodobnosti setrvání: $a_{i,i} = 1 - a_{i,i+1}$

krok2: *iterační optimalizace parametrů*

Rámce všech realizací uvažovaného slova jsou přiřazeny ke stavům trénovaného modelu na základě přiřazení Viterbiho algoritmem (pomocí pole zpětných ukazatelů). Podle vztahů uvedených v kroku 1 se vypočítají nové hodnoty parametrů. Krok 2 se opakuje, dokud se dostatečně zvyšuje věrohodnost, že trénovaný model vygeneroval všechna slova určená pro jeho trénování (dokud se zlepšují parametry modelu).

Popsaný postup (viz [Nouza97]) se pak opakuje pro všechna trénovaná slova a lze ho rozšířit i pro případ trénování Markovových modelů s M-komponentami, kdy jsou všechny rámce přiřazené k danému stavu nejprve rozděleny iteračním algoritmem K-means do M shluků. Každý z těchto shluků pak reprezentuje jednu komponentu. Z rámců přiřazených k jednotlivým shlukům jsou pak určeny střední hodnoty a rozptyly odpovídajících komponent. Váhové koeficienty všech komponent daného stavu jsou vypočteny jako poměr počtu rámců přiřazených k odpovídajícímu shluku a celkového počtu všech rámců přiřazených k danému stavu. Podrobný popis algoritmu K-means lze najít například v [MacQueen67].

Složitější ale efektivnější metodou iterační optimalizace parametrů je postup založený na metodě maximální věrohodnosti (kapitola 3.3.1). Na rozdíl od Viterbiho algoritmu, kdy je jeden rámec k danému stavu buď přiřazen nebo nepřiřazen,

jsou při metodě maximální věrohodnosti rámce promluvy přiřazeny pomocí pravděpodobnosti ke všem stavům modelu. Trénování metodou maximální věrohodnosti je proto přesnější a výsledné modely dávají při rozpoznávání lepší výsledky, neboť ve skutečnosti žádný foném přesně nekončí nebo nezačíná na hranici jednoho konkrétního rámce.

3.4.2 Trénování modelů fonémů

Trénování modelů jednotlivých fonémů je v principu stejné jako trénování modelů jednotlivých slov. Při trénovaní modelů fonémů (hlásek) se na rozdíl od rozpoznávání nerozlišuje mezi tím, zda promluva obsahuje pouze jedno nebo více slov. Zpracovává se vždy jako proud jednotlivých hlásek. Pořadí hlásek v tomto proudu musí samozřejmě být předem známo, musí existovat fonetický přepis promluvy určené pro trénování. Na základě fonetického přepisu lze pak sestavit jeden „fiktivní“ model celé promluvy jako řetězec modelů jednotlivých hlásek. Viterbiho algoritmem je pak nalezeno optimální přiřazení mezi rámci a stavy modelu celé promluvy. Jednotlivé rámce promluvy jsou tak zároveň přiřazeny i ke skutečným stavům jednotlivých hlásek, neboť jeden stav modelu promluvy odpovídá ve skutečnosti jednomu stavu modelu jedné konkrétní hlásky. V jedné iteraci trénování jsou tentokrát zpracovány všechny promluvy najednou a nové hodnoty parametrů modelů jednotlivých hlásek jsou vypočteny až z rámců přiřazených k danému stavu ze všech promluv určených pro trénování. Obdobně lze použít také Baum-Welchův algoritmus. Pro jednotlivé stavy modelu celé promluvy, a tím i pro jednotlivé stavy odpovídajících hlásek, jsou vypočteny čitatelé a jmenovatelé vztahů 3.17 - 3.20, které jsou pro každý stav modelu každé hlásky akumulovány v pomocných proměnných a nové hodnoty parametrů jsou vypočteny opět až po zpracování všech promluv.

Při trénování modelů fonémů je možné použít hned několik způsobů, jak inicializovat počáteční hodnoty parametrů a provádět iterační trénovaní. První nejjednodušší možností je použít pro inicializaci přímo parametry modelů natrénovaných v minulosti na podobné trénovací množině a reestimaci provádět následně efektivním Baum-Welchovým algoritmem. Pokud nejsou takové modely k dispozici, lze, podobně jako u modelů celých slov, rovnoměrně přiřadit rámce promluvy ke stavům modelů. Alternativně lze použít speciální množinu trénovacích dat, u kterých jsou přesně známy, respektive expertem určeny, hranice jednotlivých fonémů v řečovém signálu. V obou případech je pak provedeno několik iterací trénování Viterbiho přiřazením a následně je opět aplikován Baum-Welch. Počet komponent lze určit podobně jako u celoslovních modelů expertně, případně lze vyjít z počtu menšího a postupně dělit komponenty s největší vahou dokud se zlepšují výsledky rozpoznávání. Další možností, jak inicializovat parametry, je použít takzvaný *„plochý start“* (z anglického *flat start*), kdy jsou parametry všech modelů nastaveny na globální hodnoty vypočtené ze všech promluv určených pro trénování. Po této inicializaci následuje většinou opět Baum-Welchova reestimace.

Obecně se přitom rozlišují dvě základní kategorie fonémů - monofony a trifony.

Jako monofony označujeme takové modely, které nezávisí na kontextu, v němž je uvažujeme. Naopak jako trifony označujeme modely závislé na kontextu, u kterých rozlišujeme předchozí a následující hlásku. Například pro češtinu existuje celkem čtyřicet různých monofonů a teoreticky čtyřicet na třetí různých trifonů, neboť každý monofon může mít zároveň čtyřicet různých předchůdců a čtyřicet různých fonémů po něm může následovat. Ve skutečnosti je však počet trifonů nutných pro reprezentaci daného jazyka samozřejmě vždy mnohem menší než počet monofonů na třetí, protože některé kombinace hlásek se v daném jazyce vyskytují jen velmi zřídka nebo dokonce vůbec.

Výše uvedené trénovací strategie je možné uplatnit obecně pro jakékoli fonémové jednotky - monofony i trifony. Natrénovat modely trifonů je ovšem v praxi složitější než natrénovat monofony, neboť v trénovací množině se téměř nikdy nevyskytují všechny trifony. Z toho důvodu se při trénování trifonů používají postupy vedoucí k modelům se sdílenými parametry. Akusticky blízké trifony se slučují do obecnějších skupin pomocí metod shlukové analýzy.

V principu existují dvě základní metody shlukování, které se pro trénování trifonů používají. Jedná se o *„shlukování řízené daty“* (z anglického *data-driven clustering*) a *„shlukování založené na použití rozhodovacího stromu“* (z anglického *decision tree based clustering*). Protože v rámci této disertační práce byly použity pouze modely monofonů (vysvětlení viz kap. 6.2), nebudou zde obě uvedené metody detailně popsány. Čtenáře lze odkázat na množství literatury (např. [Young00] či [Psutka06].

KAPITOLA 4

PRINCIPY NEJČASTĚJI POUŽÍVANÝCH ADAPTAČNÍCH METOD

4.1 Členění adaptačních metod obecně

Metody adaptace na mluvčího se obecně dělí dle několika různých kritérií:

1. **znalosti přepisu respektive textu adaptační promluvy**

   - *Řízená adaptace, též adaptace s učitelem (supervised adaptation)*
     K dispozici je fonetický přepis promluvy, který je vytvořený nejčastěji člověkem a tudíž v principu správný.
   - *Neřízená adaptace, či adaptace bez učitele (unsupervised adaptation)*
     Fonetický přepis promluvy k dispozici není, ale lze ho vytvořit automaticky pomocí rozpoznávače řeči. Následkem toho může ovšem obsahovat více chyb.

2. **způsobu použití adaptačních dat**

   - *Postupná (inkrementální) adaptace (incremental adaptation)*
     Systém se adaptuje postupně s tím, jak přicházejí nová adaptační data.
   - *Dávková adaptace (batch adaptation)*
     Pro adaptaci jsou použita všechna adaptační data najednou.

3. **závislosti na obsahu adaptační promluvy**

   - *Adaptace závislá na textu*
     Pro adaptaci danou metodou musí být vždy použita stejná promluva odpovídající jednomu konkrétnímu textu.
   - *Adaptace nezávislá na textu*
     Při tomto typu adaptace je možné použít jakoukoli promluvu.

4. **typu adaptovaných parametrů**

   - *Adaptace akustického modelu*
     Cílem je upravit parametry akustického modelu používáného pro rozpoznávání řeči.
   - *Transformace (normalizace) vektoru příznaků*
     Cílem je transformovat přímo vektory příznaků vypočtené z rozpoznávaného řečového signálu.

Výše popsané metody se samozřejmě v praxi kombinují a použití konkrétní metody adaptace závisí zejména na množství a typu dat, která jsou k dispozici v dané konkrétní úloze. V systému pro diktování izolovaných slov nebo spojité řeči se používají nejčastěji metody řízené dávkové adaptace. Například v komerčních diktovacích systémech má každý uživatel vytvořen vlastní profil, jehož součástí je i sada adaptovaných modelů. Pro jejich vytvoření musí přitom uživatel nadiktovat připravený text a výsledná promluva je posléze najednou použita pro adaptaci. Naopak u telefonního dialogového systému, se kterým uživatel pracuje třeba jen několik minut až sekund, je užitečné využít pro adaptaci co nejrychleji jakoukoli promluvu. V tomto případě je tedy výhodné použít některou z inkrementálních metod.

V následujícím textu jsou popisované adaptační metody rozděleny na několik skupin dle principu jejich funkce (viz například [Woodland99]), přičemž důraz je kladen na ty, které byly využity v rámci této práce. Jedná se zejména o metody adaptace akustického modelu.

4.2 Metody typu MAP

Jak název napovídá, jsou metody typu MAP založeny na estimaci parametrů z adaptační promluvy metodou maximální aposteriorní pravděpodobnosti (kapitola 3.3.2). Jako hodnoty parametrů apriorních rozložení jsou přitom pro účely adaptace většinou používány přímo odpovídající hodnoty parametrů modelu, který je adaptován - nejčastěji jde o model nezávislý na mluvčím (SI). Zbylé apriorní parametry mají význam volitelné adaptační váhy.

Například nejvíce používaný vztah pro odhad vektoru středních hodnot adaptovaného systému lze vyjádřit přepsáním vztahu 3.32 jako

$$\hat{\boldsymbol{\mu}}_{im}^{SA} = \frac{\tau_{im}\boldsymbol{\mu}_{im}^{SI} + \sum_{t=1}^{T}\zeta_t(i,m)\mathbf{x}_t}{\tau_{im} + \sum_{t=1}^{T}\zeta_t(i,m)}, \tag{4.1}$$

kde symbol $\sum_{t=1}^{T}\zeta_t(i,m)$ značí okupační věrohodnost m-té komponenty stavu i vypočtenou z adaptačních dat $\mathbf{X}$ a člen τ_{im} má význam adaptační váhy. Její hodnota souvisí s množstvím dat použitých pro trénování uvažované komponenty SI modelu. Pokud by hodnota τ_{im} dané komponenty byla přímo rovna její celkové okupační věrohodnosti vypočtené ze všech trénovacích promluv, vyjadřoval by

vztah 4.1 odhad středních hodnot této komponenty vypočtený metodou ML dohromady z adaptačních i trénovacích promluv.

Pravdivost tohoto tvrzení lze jednoduše ukázat. Označme symbolem $\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)$ okupační věrohodnost vypočtenou při trénování SI systému z trénovacích promluv. Pro $\tau_{im} = \sum_{t=1}^{T} \zeta_t^{TREN}(i,m)$ dostaneme

$$\hat{\boldsymbol{\mu}}_{im}^{SA} = \frac{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)\boldsymbol{\mu}_{im}^{SI} + \sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t}{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m) + \sum_{t=1}^{T} \zeta_t(i,m)}, \tag{4.2}$$

kde

$$\hat{\boldsymbol{\mu}}_{im}^{SI} = \frac{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)\mathbf{x}_t^{TREN}}{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)}. \tag{4.3}$$

Nakonec, dosazením vztahu 4.3 do 4.2 vyjde

$$\hat{\boldsymbol{\mu}}_{im}^{SI} = \frac{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)\mathbf{x}_t^{TREN} + \sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t}{\sum_{t=1}^{T} \zeta_t^{TREN}(i,m) + \sum_{t=1}^{T} \zeta_t(i,m)}. \tag{4.4}$$

Odvozený vztah 4.4 představuje odhad vektoru středních hodnot vypočtený metodou maximální věrohodnosti dohromady z adaptačních i trénovacích promluv. V některých pramenech, kde není vysvětlen princip odhadu parametrů metodou MAP obecně, pak bývá adaptace metodou MAP odvozena jednoduše opačným postupem: Ze vztahu 4.4 je nejprve odvozen vztah 4.2 a v něm výraz $\sum_{t=1}^{T} \zeta_t^{TREN}(i,m)$ nahrazen adaptační vahou τ_{im}.

Během adaptace je pak samozřejmě vždy používána nižší hodnota τ_{im}, než která odpovídá skutečné hodnotě celkové okupační věrohodnosti. Střední hodnoty SI systému tím pádem mají nižší vliv, než když by byl pro adaptaci použit vztah 4.4 (respektive než když by byla adaptační data pouze přidána do rozsáhlé trénovací databáze SI systému). Hodnota τ_{im} přitom může být pro účely adaptace pro všechny komponenty a stavy konstantní, typicky v rozmezí 1 až 20.

Vztah 4.1 pro adaptované vektory středních hodnot pak může být chápan jako vážený součet s parametrem

$$\frac{\tau_{im}}{\tau_{im} + \sum_{t=1}^{T} \zeta_t(i,m)} \tag{4.5}$$

mezi původními středními hodnotami SI systému a novými hodnotami získanými metodou MLE pouze z adaptačních promluv.

Pomocí váhového koeficientu τ_{im} lze podobně definovat i vztahy pro nové hodnoty kovariační matice

$$\hat{\boldsymbol{\Sigma}}_{im}^{SA} = \frac{\sum_{t=1}^{T} \zeta_t(i,m)\mathbf{x}_t(\mathbf{x}_t)'}{\tau_{im} + \sum_{t=1}^{T} \zeta_t(i,m)} + \\ + \frac{\tau_{im}\left(\boldsymbol{\Sigma}_{im}^{SI} + \boldsymbol{\mu}_{im}^{SI}(\boldsymbol{\mu}_{im}^{SI})'\right)}{\tau_{im} + \sum_{t=1}^{T} \zeta_t(i,m)} - \boldsymbol{\mu}_{im}^{SA}(\boldsymbol{\mu}_{im}^{SA})' \tag{4.6}$$

a váhu příslušné komponenty

$$\hat{c}_{im}^{SA} = \frac{\tau_{im} + \sum_{t=1}^{T} \zeta_t(i,m)}{\sum_{k=1}^{M} \left(\tau_{ik} + \sum_{t=1}^{T} \zeta_t(i,k) \right)}. \tag{4.7}$$

Z uvedených vztahů vyplývá, že čím vyšší je množství dat použitých pro adaptaci, tím nižší vliv na výslednou hodnotu adaptovaného parametru má původní parametr SI systému. Naopak vliv parametrů SI systému se zvyšuje s rostoucí hodnotou váhového koeficientu τ_{im}.

Pro $\tau_{im} = 0$ nebo $\sum_{t=1}^{T} \zeta_t(i,m) \rightarrow \infty$ je adaptace metodou MAP totožná s odhadem parametrů z adaptačních promluv metodou MLE a tedy s natrénováním na řečníkovi závislého (SD) systému.

Při praktickém použití metody MAP se ale nejčastěji adaptují pouze vektory středních hodnot, pro adaptaci rozptylů a váhových koeficientů jednotlivých komponent většinou není k dispozici dostatek dat. Změnou hodnot rozptylů by se navíc mohla výrazně snížit robustnost systému.

Největší výhodou metody MAP je skutečnost, že díky svému principu jako jediná z používaných adaptačních technik konverguje s rostoucím množstvím dat k teoreticky nejlepšímu SD modelu. Je proto často považována za základní adaptační metodu a používá se zejména v případech, kdy je pro adaptaci k dispozici více promluv v trvání deseti až několika desítek minut. Naopak její největší nevýhodu představuje fakt, že adaptovány mohou být pouze parametry modelů, jejichž realizace jsou součástí adaptační promluvy. Modely řečových jednotek, které nejsou v adaptačních promluvě obsaženy, nejsou adaptovány vůbec a modely, pro které je množství adaptačních dat příliš malé, jsou po aplikaci metody MAP adaptovány nedostatečně.

4.2.1 Predikce modelů založená na regresi (RMP)

Pro odstranění problému nedostatečně adaptovaných parametrů bylo vyvinuto několik nadstavbových regresních technik, které jsou označovány jako metody typu RPM (Regression Based Model Prediction - predikce modelů založená na regresi). Jejich princip spočívá v tom, že se snaží najít vzájemně korelované parametry adaptovaného modelu, mezi korelovanými parametry pak odvodit lineární regresní vztahy a ty následně využít pro dodatečnou adaptaci těch parametrů, pro které nebyl k dispozici dostatek dat. Adaptace je přitom většinou aplikována pouze pro vektory středních hodnot.

Příkladem techniky typu RMP je metoda WNR (Weighted Neighbor Regression - regrese s vážením sousedů) [Lei00]. Při jejím použití je nejprve pro každou komponentu SI systému pomocí Mahalanobisovy vzdálenosti nalezeno celkem N komponent (*sousedů*), které jsou jí nejbližší. Po adaptaci metodou MAP jsou pak všechny komponenty rozděleny podle množství použitých dat na zdrojové, pro které byl k dispozici dostatek dat, a cílové, pro které k dispozici dostatek

dat nebyl. Následně jsou pro každou cílovou komponentu vybrány z množiny nejbližších komponent pouze ty, které byly označeny jako zdrojové. Mezi nimi je pak spočítána korelace. Pokud je vypočtená hodnota korelace vyšší než minimální požadovaná, je pro střední hodnoty vybraných zdrojových komponent určena rovnice regresní přímky (metodou vážených nejmenších čtverců) a její pomocí jsou adaptovány střední hodnoty dané cílové komponenty. V [Lei00] jsou popsány další varianty této metody, které se liší podle typu použité regresní metody.

Obdobným příkladem je metoda *svazování parametrů* publikovaná v [Železný01]. Regresní vztahy jsou v rámci ní hledány ve dvourozměrném prostoru, v němž každá osa reprezentuje jeden stejný parametr (střední hodnotu) dvou různých komponent daného akustického modelu a každý bod různého řečníka. Pro hledání regresních vztahů proto musí být k dispozici několik sad SD modelů nebo alespoň několik sad dobře adaptovaných modelů, které v uvažovaném prostoru představují několik konkrétních bodů. Pro každou dvojici parametrů je pak spočítána korelace a metodou nejmenších čtverců nalezena rovnice regresní přímky. Při aplikaci celé metody je pak také nejprve provedena adaptace metodou MAP a následně je opět provedeno rozdělení adaptovaných parametrů na cílové a zdrojové. Pro každý cílový parametr je potom nalezen nejvíce korelovaný parametr ze skupiny zdrojových parametrů a následně je tento cílový parametr adaptován použitím odpovídajícího předem vypočteného regresního vztahu.

4.2.2 Strukturální MAP (SMAP)

Další nadstavbový přístup k technice MAP představuje metoda označovaná jako SMAP (Structural MAP - strukturální MAP) [Shinoda97]. Stejně jako výše popsané metody RPM se i SMAP používá pro zvýšení rychlosti adaptace. V rámci SMAP jsou všechny Gaussovy komponenty systému rozděleny do stromové struktury, přičemž apriorní parametry jsou pro každý uzel výsledného stromu vypočteny na základě komponent obsažených v předcházejícím uzlu vyšší úrovně. Během adaptace lze pak vytvořený strom procházet a adaptaci je možné provádět na různých úrovních stromu dle aktuálně dostupného množství adaptačních dat. Uvedený postup adaptace pomocí hierarchické stromové struktury je podrobně popsán v následující kapitole.

4.3 Metody založené na lineární transformaci

Princip uvedených technik spočívá v lineární transformaci parametrů původních modelů tak, aby nové adaptované modely více odpovídaly charakteristikám hlasu daného mluvčího. Výhoda tohoto přístupu spočívá ve zvýšení rychlosti adaptace, neboť v případě, že je adaptačních promluv málo, může být použita jedna transformace najednou pro několik akusticky blízkých Gaussových komponent. Množina akusticky blízkých komponent, sdílejících jednu společnou transformaci, je přitom označována jako regresní třída. Svazování komponent je možné provádět přes celé

stavy jednotlivých modelů nebo lépe přímo přes jednotlivé komponenty různých stavů.

4.3.1 Maximálně věrohodná lineární regrese (MLLR)

Asi nejrozšířenější a základní technikou ze skupiny přístupů založených na lineární transformaci je metoda MLLR (Maximum Likelihood Linear Regression - maximálně věrohodná lineární regrese) [Leggetter95]. V rámci metody MLLR je prováděna transformace parametrů modelů tak, aby byla maximalizována věrohodnost toho, že pro adaptaci použitá posloupnost vektorů příznaků $\mathbf{X} = \{\mathbf{x}_1,\ \mathbf{x}_2,\ \ldots,\ \mathbf{x}_T\}$ byla vygenerována daným modelem. Maximalizace věrohodnosti je přitom dosažena použitím techniky EM podobně jako v rámci estimace metodou ML nebo MAP.

Metoda MLLR se používá zejména pro transformaci vektorů středních hodnot. Pro jednoduchost odvození transformace těchto parametrů nejprve uvažujme, že regresní třída, pro kterou je transformace hledána, obsahuje pouze jednu Gaussovu komponentu odpovídající m-té komponentě j-tého stavu modelu jednoho konkrétního fonému. Vztah pro transformaci vektorů středních hodnot lze potom vyjádřit jako

$$\boldsymbol{\mu}_{jm}^{SA} = \mathbf{A}\boldsymbol{\mu}_{jm}^{SI} + \mathbf{b}, \tag{4.8}$$

kde $\mathbf{A}$ je hledaná transformační matice o rozměrech $P \times P$, $\mathbf{b}$ je P-rozměrný vektor posunutí a P je počet příznaků počítaných z řečového signálu.

Rovnice 4.8 bývá pro jednoduchost uváděna ve tvaru

$$\boldsymbol{\mu}_{jm}^{SA} = \mathbf{W}\boldsymbol{\xi}_{jm}^{SI}, \tag{4.9}$$

kde $\mathbf{W}$ je matice o rozměrech $P \times (P+1)$ a $\boldsymbol{\xi}_{jm} = [\omega,\ \mu_{jm_1},\ \mu_{jm_2},\ \ldots,\ \mu_{jm_P}]^T$ je rozšířený vektor středních hodnot s posunutím ω.

Pro odvození výpočtu transformační matice $\mathbf{W}$ metodou EM pro modely typu CDHMM s multimodálním Gaussovým rozložením připomeňme nejprve vztah 3.15 pro výpočet pomocné funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = \sum_{\mathbf{S}\in\mathbf{\Psi}} \sum_{\mathbf{K}\in\mathbf{\Omega}^T} \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi})}{p(\mathbf{X}|\mathbf{\Phi})} log\ p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\hat{\mathbf{\Phi}}).$$

Ten je možné s pomocí 3.13 upravit do tvaru

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = \sum_{\mathbf{S}\in\mathbf{\Psi}} \sum_{\mathbf{K}\in\mathbf{\Omega}^T} \frac{p(\mathbf{X}, \mathbf{S}, \mathbf{K}|\mathbf{\Phi})}{p(\mathbf{X}|\mathbf{\Phi})} \Big(\sum_{t=1}^{T} log\ \hat{a}_{s_{t-1},s_t} + \\ + \sum_{t=1}^{T} log\ \hat{b}_{s_t k_t}(\mathbf{x}_t) + \sum_{t=1}^{T} log\ \hat{c}_{s_t k_t} \Big). \tag{4.10}$$

Vzhledem k tomu, že hledáme transformační matici pro vektory středních hodnot jedné konkrétní komponenty, lze rovnici 4.10 přepsat do tvaru

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = c + \sum_{\mathbf{S}\in\mathbf{\Psi}} \sum_{\mathbf{K}\in\mathbf{\Omega}^T} \sum_{t=1}^{T} \frac{p(\mathbf{X}, s_t = j, k_t = m|\mathbf{\Phi})}{p(\mathbf{X}|\mathbf{\Phi})} log\ \hat{b}_{jm}(\mathbf{x}_t). \quad (4.11)$$

Použitím vztahu 3.22 definujícího okupační věrohodnost konkrétní komponenty lze předchozí vztah upravit do podoby.

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = c + \sum_{t=1}^{T} \zeta_t(j,m) log\ \hat{b}_{jk}(\mathbf{x}_t). \quad (4.12)$$

Rozepsáním členu $log\ \hat{b}_{jk}(\mathbf{x}_t)$ v předchozí rovnici potom dostaneme

$$Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = c - \frac{1}{2} \sum_{t=1}^{T} \zeta_t(j,m) \Big[P\ log(2\pi) + log|\mathbf{\Sigma}_{jm}| + \\ + log(c_{jm}) + (\mathbf{x}_t - \mathbf{W}\boldsymbol{\xi}_{jm})' \mathbf{\Sigma}_{jm}^{-1} (\mathbf{x}_t - \mathbf{W}\boldsymbol{\xi}_{jm}) \Big]. \quad (4.13)$$

Maximum pomocné funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$ pak vzhledem k hledané transformační matici najdeme tak, že nejprve vypočteme její derivaci dle $\mathbf{W}$

$$\frac{d}{d\mathbf{W}} Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi}) = \sum_{t=1}^{T} \zeta_t(j,m) \mathbf{\Sigma}_{jm}^{-1} \Big[\mathbf{x}_t - \mathbf{W}\boldsymbol{\xi}_{jm} \Big] \boldsymbol{\xi}_{jm}' \quad (4.14)$$

a poté tuto derivaci položíme rovnu nule

$$\sum_{t=1}^{T} \zeta_t(j,m) \mathbf{\Sigma}_{jm}^{-1} \Big[\mathbf{x}_t - \mathbf{W}\boldsymbol{\xi}_{jm} \Big] \boldsymbol{\xi}_{jm}' = 0. \quad (4.15)$$

Úpravou předchozího vztahu dostaneme finální rovnici ve tvaru

$$\sum_{t=1}^{T} \zeta_t(j,m) \mathbf{\Sigma}_{jm}^{-1} \mathbf{x}_t \boldsymbol{\xi}_{jm}' = \sum_{t=1}^{T} \zeta_t(j,m) \mathbf{\Sigma}_{jm}^{-1} \mathbf{W} \boldsymbol{\xi}_{jm} \boldsymbol{\xi}_{jm}'. \quad (4.16)$$

Řešení této rovnice v uzavřeném tvaru existuje pouze pro modely s diagonální kovariační maticí. Jeho odvození můžeme nyní jednoduše provést i pro obecnější případ, kdy je transformační matice sdílena celkem M různými komponentami. Nejprve upravíme rovnice 4.16 pro případ více komponent

$$\sum_{m=1}^{M} \sum_{t=1}^{T} \zeta_t(m) \mathbf{\Sigma}_{m}^{-1} \mathbf{x}_t \boldsymbol{\xi}_{m}' = \sum_{m=1}^{M} \sum_{t=1}^{T} \zeta_t(m) \mathbf{\Sigma}_{m}^{-1} \mathbf{W} \boldsymbol{\xi}_{m} \boldsymbol{\xi}_{m}'. \quad (4.17)$$

Nyní označíme pravou stranu předchozí rovnice symbolem $\mathbf{Y}$ ($\mathbf{Y}$ bude matice o rozměrech $P \times (P+1)$ a upravíme ji do tvaru

$$\mathbf{Y} = \sum_{m=1}^{M} \mathbf{V}^m \mathbf{W} \mathbf{D}^m, \tag{4.18}$$

kde

$$\mathbf{V}^m = \sum_{t=1}^{T} \zeta_t(m) \mathbf{\Sigma}_m^{-1} \tag{4.19}$$

a

$$\mathbf{D}^m = \boldsymbol{\xi}_m \boldsymbol{\xi}_m^{'}. \tag{4.20}$$

Protože $\mathbf{\Sigma}_m$ je diagonální a $\mathbf{D}^m$ je symetrická, platí pro prvky matice $\mathbf{Y}$ že

$$y_{i,j} = \sum_{q=1}^{P+1} w_{i,q} g_{j,q}^{(i)}, \tag{4.21}$$

kde $g_{j,q}^{(i)}$ jsou prvky matice $\mathbf{G}^{(i)}$ o rozměrech $(P+1) \times (P+1)$, které jsou definovány jako

$$g_{j,q}^{(i)} = \sum_{m=1}^{M} v_{i,i}^r d_{j,q}^m. \tag{4.22}$$

Označíme-li nyní levou stranu rovnice 4.17 symbolem $\mathbf{Z}$, musí platit, že

$$z_{i,j} = y_{i,j} = \sum_{q=1}^{P+1} w_{i,q} g_{j,q}^{(i)}. \tag{4.23}$$

Protože $\mathbf{Z}$ ani $\mathbf{G}^{(i)}$ nezávisí na hledané transformační matici, lze vyjádřit řešení pro i-tý řádek matice $\mathbf{W}$ jako

$$\mathbf{w}_i^{'} = \mathbf{G}^{(i)^{-1}} \mathbf{z}_i^{'}. \tag{4.24}$$

Výpočet inverze matice $\mathbf{G}^{(i)}$ z výsledného vztahu 4.24 je v praxi netriviální, neboť tato matice je často špatně podmíněná. V praxi je proto třeba tvořit regresní třídy tak, aby obsahovaly dostatečný počet komponent, pro které musí být zároveň dostatek adaptačních dat, aby jejich příspěvky nebyly nulové. Pro výpočet inverze potom bývá používána metoda SVD (Singular Value Decomposition - rozklad na singulární hodnoty) [Press02].

Uvedený postup odvození lze podobně jako v kapitole 3.3.2 jednoduše rozšířit i pro případ, že adaptační data jsou tvořena posloupností několika promluv. Ve všech výsledných vztazích se pak kromě proměnné t sčítá také přes jednotlivé promluvy.

4.3.1.1 Diagonální transformace

Protože výpočet plné transformační matice dle výše odvozených vztahů je poměrně náročný, používá se v některých aplikacích, například když je k dispozici pouze malé množství dat, transformační matice ve tvaru

$$\mathbf{W} = \begin{pmatrix} w_{1,1} & w_{1,1} & \dots & 0 \\ w_{2,1} & 0 & \dots & 0 \\ w_{n,1} & 0 & \dots & w_{n,n+1} \end{pmatrix} \tag{4.25}$$

Každý prvek nového transformovaného vektoru pak vnikne tak, že se vynásobí odpovídající původní hodnota a přičte se k ní konstanta dle vztahu

$$\mu_i^{SA} = \omega w_{i,1} + w_{i+1,i}\mu_i^{SI}. \tag{4.26}$$

V případě, že $\omega = 0$, může být transformační matice čtvercová a diagonální a celý výpočet se ještě více urychlí.

Teoreticky lze navíc ještě hledanou transformaci omezit pouze na vybrané příznaky (například statické) a tím ještě více redukovat počet koeficientů, které je nutné vypočítat z adaptačních dat. V praxi ale tato redukce bohužel téměř vždy vede k signifikantnímu zhoršení rozpoznávacího skóre [Leggetter95].

4.3.1.2 Transformace ostatních parametrů modelu

Metodou MLLR mohou být teoreticky transformovány i ostatní parametry modelů, nejen vektory středních hodnot. Protože ale pravděpodobnosti přechodů mezi stavy a váhové koeficienty jednotlivých komponent mají jen malý vliv na úspěšnost rozpoznávání, má smysl rozšířit adaptaci pouze na kovariační matice.

V principu existují dvě možnosti, jak transformaci kovariační matice vyjádřit. V prvním možném případě je hledána transformace ve tvaru

$$\mathbf{\Sigma}_m^{SA} = \mathbf{B}_m \mathbf{H} \mathbf{B}_m^{'}, \tag{4.27}$$

kde $\mathbf{H}$ je hledaná transformace a $\mathbf{B}_m$ představuje Choleskiho faktor matice $\mathbf{\Sigma}_m^{SI}$. Výhodou je, že řešení uvedené rovnice metodou ML je jednoduché (viz například [Gales98]. Nevýhodu představuje skutečnost, že adaptovaná kovariační matice je plná, ačkoli původní matice apriorního modelu je diagonální. Při rozpoznávání je pak proto pomalejší výpočet logaritmu věrohodnosti, že daná komponenta vygenerovala konkrétní vektor příznaků

$$log\ p(\mathbf{x}_t|\mathbf{W}, \mathbf{H}, \boldsymbol{\mu}_m^{SI}, \mathbf{\Sigma}_m^{SI}) = log\ N(\mathbf{x}_t|\boldsymbol{\mu}_m^{SA}, \mathbf{\Sigma}_m^{SA}). \tag{4.28}$$

Symbol $N(\mathbf{.})$ ve vztahu 4.28 přitom představuje Gaussovo pravděpodobnostní rozložení dané komponenty.

Ve druhém případě je transformace hledána ve tvaru

$$\mathbf{\Sigma}_m^{SA} = \mathbf{H} \mathbf{\Sigma}_m^{SI} \mathbf{H}^{'}. \tag{4.29}$$

Řešení uvedené rovnice je pak nutné provádět iteračním výpočtem, nicméně výhodou je, že během rozpoznávání je možné počítat věrohodnosti efektivně s původní diagonální kovariační maticí neadaptovaného systému, neboť platí (viz [Gales98])

$$log\, p(\mathbf{x}_t|\mathbf{W},\mathbf{H},\boldsymbol{\mu}_m^{SI},\boldsymbol{\Sigma}_m^{SI}) = log\, N(\mathbf{H}^{-1}\mathbf{x}_t|\mathbf{H}^{-1}\boldsymbol{\mu}_m^{SA},\boldsymbol{\Sigma}_m^{SI}) - \frac{1}{2}log(|\mathbf{H}|^2). \quad (4.30)$$

4.3.1.3 MLLR s omezením - CMLLR

Kromě klasické metody MLLR popsané výše, nazývané někdy jako MLLR bez omezení (unconstrained MLLR), existuje také varianta s omezením (constrained MLLR - CMLLR) [Digalakis95]. V rámci CMLLR je hledána jedna stejná transformační matice pro všechny komponenty systému a tato je poté aplikována na vektory středních hodnot i kovariační matice.

$$\begin{aligned}\boldsymbol{\mu}_m^{SA} &= \mathbf{A}\boldsymbol{\mu}_m^{SI} - \mathbf{b}\\ \boldsymbol{\Sigma}_m^{SA} &= \mathbf{A}\boldsymbol{\Sigma}_m^{SI}\mathbf{A}'.\end{aligned} \quad (4.31)$$

Řešení rovnice 4.31 se opět provádí metodou EM, přičemž funkci $Q(\hat{\boldsymbol{\Phi}},\boldsymbol{\Phi})$ je v tomto případě možné upravit do tvaru

$$Q(\hat{\boldsymbol{\Phi}},\boldsymbol{\Phi}) = c - \frac{1}{2}\sum_{t=1}^{T}\zeta_t(m)\Big[log|\boldsymbol{\Sigma}_m^{SI}| - log(|\mathbf{A}|^2) + (\hat{\mathbf{x}}_t - \boldsymbol{\mu}_m^{SI})'\boldsymbol{\Sigma}_m^{SI^{-1}}(\hat{\mathbf{x}}_t - \boldsymbol{\mu}_m^{SI})\Big], \quad (4.32)$$

kde $\hat{\mathbf{x}}_t$ je vektor příznaků transformovaný dle vztahu

$$\hat{\mathbf{x}}_t = \mathbf{A}^{-1}\mathbf{x}_t + \mathbf{A}^{-1}\mathbf{b}. \quad (4.33)$$

Z uvedených vztahů vyplývá, že CMLLR odpovídá aplikaci lineární transformace přímo na vektory příznaků a bývá proto někdy zařazována mezi techniky typu FSA (Feature Space Adaptation - adaptace v prostoru příznaků). V této skutečnosti spočívá i největší výhoda metody CMMLR, neboť při adaptaci na mluvčího není nutné počítat a ukládat adaptované parametry jednotlivých modelů, ale stačí pouze aplikovat vypočtenou transformaci na vektory příznaků. Výpočet logaritmu věrohodnosti během rozpoznávání je ovšem nutné počítat jako:

$$log\, p(\mathbf{x}_t|\mathbf{A},\mathbf{b},\boldsymbol{\mu}_m^{SI},\boldsymbol{\Sigma}_m^{SI}) = log\, N(\mathbf{A}^{-1}\mathbf{x}_t + \mathbf{A}^{-1}\mathbf{b}|\boldsymbol{\mu}_m^{SI},\boldsymbol{\Sigma}_m^{SI}) - \frac{1}{2}log(|\mathbf{A}|). \quad (4.34)$$

Metoda CMLLR se proto používá zejména v rámci trénování metodou SAT (kapitola 4.3.3), které je založeno na použití trénovacích dat od velkého množství různých mluvčích.

Nevýhodu techniky CMLLR představuje skutečnost, že transformační matici je nutné hledat iteračním výpočtem [Gales98] a že tato metoda dává sama o sobě podobné výsledky jako MLLR bez omezení.

4.3.1.4 Tvorba regresních tříd

Jak již bylo uvedeno v úvodu této kapitoly, patří mezi největší výhody metody MLLR skutečnost, že jedna transformační matice může být vypočtena najednou pro několik Gaussových komponent systému, které spolu sdílí jednu regresní třídu. Možností, jak vytvořit jednotlivé regresní třídy, je přitom hned několik.

V případě, že je k dispozici malé množství adaptačních dat, mohou všechny komponenty sdílet pouze jednu společnou regresní třídu. V případě většího množství dat je naopak možné vytvořit jednu třídu pro každý konkrétní foném nebo je možné rozdělit všechny modely do jednotlivých fonetických tříd: na explozívy, frikativy atd.

Problém nalezení vhodných regresních tříd vzhledem k množství dostupných adaptačních dat lze vyřešit také aplikací algoritmu klastrování [Young00]. V tomto případě jsou regresní třídy automaticky uspořádány do binárního stromu, kde každý uzel reprezentuje skupinu akusticky podobných komponent, přičemž ve vrcholu stromu jsou obsaženy všechny komponenty. Výhoda uvedeného přístupu spočívá v tom, že při adaptaci je možné regresní strom postupně procházet a transformační matice je možné vygenerovat pouze pro uzly, pro které je k dispozici dostatek adaptačních dat a splňují jednu z následujících podmínek - jsou buď uzly koncovými nebo mají potomka, pro něhož byl adaptačních dat nedostatek.

Další informace o metodě MLLR mohou být nalezeny například v [Huang01], [Young00], [Leggetter95] nebo [Matsoukas97].

4.3.2 Kombinace metod MAP a MLLR

V praxi se jako výhodné ukazuje použít metodu MLLR v kombinaci s metodou MAP, přičemž způsobů, jak obě techniky zkombinovat, existuje hned několik. Například v rámci metody označované jako MAPLR (Maximum A Posteriori Linear Regression - maximálně aposteriorní lineární regrese) [Chesta99] je možné pro výpočet transformační matice použít namísto metody maximální věrohodnosti metodu maximální aposteriorní pravděpodobnosti. Zmíněný přístup má ovšem tu nevýhodu, že nejprve je třeba odhadnout apriorní pravděpodobnosti rozložení $p(\mathbf{W})$.

Jednodušší a efektivnější možnost kombinace obou metod spočívá v postupu, kdy jsou nejprve pomocí MLLR odhadnuty nové hodnoty vektorů středních hodnot, které jsou následně použity jako apriorní pro metodu MAP. Díky prvotnímu použití MLLR pro jednotlivé regresní třídy tak mohou být adaptovány i modely řečových jednotek, které nejsou obsaženy v adaptačních promluvách a které nemohou být normálně pomocí MAP adaptovány. Následným použitím metody MAP je pak zaručena konvergence všech adaptovaných modelů k SD modelům.

4.3.3 Trénování s adaptací na mluvčího (SAT)

Metoda označovaná jako SAT (Speaker Adaptive Training - trénování s adaptací na mluvčího) [Anastakos96] představuje alternativní postup pro trénování akustic-

kého modelu nezávislého na mluvčím. Je založena na hypotéze, že velká variabilita (hodnoty rozptylů) akustické modelu nezávislého na mluvčím je způsobena dvěma faktory - fonetickou odlišností jednotlivých hlásek a rozdílností hlasových charakteristik jednotlivých mluvčích z trénovací databáze, která nezávisí na informačním obsahu řečového signálu. Cílem SAT je tuto rozdílnost mezi řečníky během trénovaní potlačit a vytvořit přesnější akustické modely s menšími hodnotami rozptylů - tzv. *„kompaktní modely“* (z anglického *compact model*).

Metoda SAT je proto založena na základním předpokladu, že je možné rozdělit trénovací databázi promluv dle jednotlivých mluvčích. Počet všech mluvčích se označuje symbolem N. Dále se předpokládá, že promluvy $\mathbf{X}^{(n)}$ jsou pro každého mluvčího generovány jeho specifickým modelem respektive transformací $\mathbf{G}^{(n)}(\mathbf{\Phi}_c)$, kde $\mathbf{\Phi}_c$ označuje hledané *kompaktní* parametry, které, na rozdíl od běžně uvažovaných parametrů, nevystihují variabilitu hlasových charakteristik jednotlivých mluvčích. Jinak řečeno, transformace $\mathbf{G}^{(n)}$ mapuje kompaktní parametry $\mathbf{\Phi}_c$ na parametry závislé na konkrétním mluvčím podobně, jako například metoda MLLR transformuje na řečníkovi nezávislé modely na modely konkrétního mluvčího. Množina optimálních transformací $\hat{\mathbf{\Theta}} = \{\hat{\mathbf{G}}^{(1)}, \hat{\mathbf{G}}^{(2)}, ..., \hat{\mathbf{G}}^{(N)}\}$ charakterizující jednotlivé mluvčí je přitom v rámci SAT hledána současně s modely $\hat{\mathbf{\Phi}}_c$

$$(\hat{\mathbf{\Phi}}_c, \hat{\mathbf{\Theta}}) = \underset{\mathbf{\Phi},\mathbf{\Theta}}{\operatorname{argmax}} \prod_{n=1}^{N} p(\mathbf{X}^{(n)}|\mathbf{G}^{(n)}(\mathbf{\Phi}_c)). \tag{4.35}$$

V praxi se v rámci metody SAT hledají optimální parametry vektorů středních hodnot a rozptylů, ostatní parametry jsou odhadnuty klasicky metodou ML. Transformace charakterizující jednotlivé mluvčí se hledá ve stejném tvaru jako v rámci metody MLLR. Hledání optimálních parametrů je založeno na algoritmu EM a pomocná funkce $Q(\hat{\mathbf{\Phi}}_c, \mathbf{\Phi}_c)$ má přitom vzhledem k jednotlivým Gaussovým komponentám systému tvar

$$Q(\hat{\mathbf{\Phi}}_c, \mathbf{\Phi}_c) = \sum_{n,t,j,m=1}^{N,T_n,I,M_j} \zeta_t^{(n)}(j,m) log N(\mathbf{x}_t^{(n)}|\hat{\mathbf{A}}^{(n)}\hat{\boldsymbol{\mu}}_{jm}^c + \hat{\mathbf{b}}^{(n)}, \hat{\mathbf{\Sigma}}_{jm}^c), \tag{4.36}$$

kde $N(\mathbf{.})$ symbolizuje Gaussovo pravděpodobnostní rozložení dané komponenty.

Přímá maximalizace funkce $Q(\hat{\mathbf{\Phi}}, \mathbf{\Phi})$ vede na soustavu složitých nelineárních rovnic. V praxi se proto používá postup, v jehož průběhu jsou jednotlivé optimální parametry hledány postupně, přičemž zbylé dosud neodhadnuté parametry jsou fixovány. Nejprve se zafixují střední hodnoty a rozptyly kompaktního modelu a odhadnou se jednotlivé transformace. Následně se opět zafixují rozptyly a pomocí nalezených transformací jsou vypočítány nové střední hodnoty. V posledním kroku jsou odhadnuty hodnoty rozptylů.

Vztahy pro odhad jednotlivých parametrů a transformací za předpokladu diagonálních kovariačních matic je možné najít v [Anastakos96]. Z výsledků uvedených ve zmiňovaném článku vyplývá, že SI (kompaktní) modely vytvořené uvedeným způsobem dávají stejné výsledky jako modely natrénované metodou ML.

Výhoda trénování metodou SAT se projeví až v případě, že výsledný model adaptujeme na konkrétního mluvčího. Adaptovaný kompaktní model by měl totiž při rozpoznávání dávat relativně cca o 10 procent lepší výsledky, než model vzniklý adaptací standardního SI modelu.

4.3.3.1 Metoda FSA-SAT

Popsaná nejobecnější varianta metody SAT se v praxi příliš nepoužívá, neboť je spojena s přílišnými výpočetními a paměťovými nároky [Matsoukas97]. Mnohem populárnější je proto metoda označovaná jako FSA-SAT [Gales96]. Jak název napovídá, jedná se variantu metody SAT založenou na transformaci vektorů příznaků - tedy užití metody MLLR s omezením (CMLLR). Celý proces estimace parametrů je pak jednodušší: pro každého mluvčího se nejprve vypočte transformační matice, pomocí které se transformují jeho adaptační promluvy a pomocí metody ML se pak z transformovaných dat všech mluvčích odhadnou nové hodnoty parametrů kompaktních modelů. Ty se pak opět použijí pro nový výpočet jednotlivých transformačních matic a celý uvedený postup se několikrát opakuje.

4.4 Metody založené na shlukování modelů mluvčích

Principem metod založených na *shlukování* (modelů) *mluvčích* (z anglického *Speaker Clustering* - SC) je vytvořit ve fázi trénování systému několik sad modelů pro různé skupiny (klastry respektive shluky) tzv. *referenčních* mluvčích a tyto sady modelů pak během adaptace vhodně využít pro odhad parametrů modelu neznámého mluvčího.

Nejjednodušší formou tohoto přístupu je aplikace na pohlaví závislých (gender dependent - GD) modelů, které mohou být použity namísto SI modelů dvěma způsoby. Jednak přímo pro rozpoznávání řeči a pak jako apriorní při adaptaci (viz kap. 6.3.5).

4.4.1 Trénovaní s výběrem mluvčích (SST)

Metoda SST (Speaker Selection Training - trénování s výběrech mluvčích) [Padmanabhan98] představuje limitní variantu technik založených na shlukování mluvčích, neboť v rámci SST je každý shluk reprezentován právě jedním referenčním mluvčím. Adaptace na neznámého mluvčího probíhá ve dvou krocích:

1. Nejprve je proveden výběr N referenčních mluvčích, jejichž data respektive modely budou použity pro adaptaci.

2. Poté je vytvořen finální adaptovaný model.

Oba uvedené kroky jsou podrobně rozebrány v následujících podkapitolách.

4.4.1.1 Výběr referenčních mluvčích

Pro tento účel je možné použít celou řadu různých strategií, z nichž dvě nejefektivnější jsou popsány níže.

Pokud je k dispozici přepis adaptačních dat, ať již předem připravený či vytvořený rozpoznávačem řeči, je možné použít všechny modely referenčních mluvčích pro výpočet věrohodnosti, že právě daný model vygeneroval adaptační data, a následně vybrat skupinu N mluvčích, jejichž modely měly věrohodnost nejvyšší. V tomto případě je ovšem nutné během trénování systému vytvořit SD modely pro všechny referenční mluvčí. K tomuto účelu se v praxi často používá některá z adaptačních metod (např. MAP či MLLR), neboť většinou není k dispozici dostatek dat pro klasické trénování.

Pokud fonetický přepis k dispozici není vůbec, je možné založit hledání nejbližších mluvčích na metodách identifikace řečníka. Pro tento účel bývají většinou používány modely typu GMM (Gaussian Mixture Model - gaussovské mixturové modely), které je opět nutné vytvořit pro všechny mluvčí předem ve fázi trénování.

4.4.1.2 Tvorba adaptovaného modelu

Je-li vybrána skupina N nejbližších mluvčích, existuje opět několik možných způsobů, jak vytvořit finální adaptovaný model.

V případě, že jsou během trénování SD modelů referenčních mluvčích uloženy okupační věrohodnosti všech komponent, je možné sečíst komponenty modelů N mluvčích s vahou, která pro každou komponentu každého mluvčího odpovídá podílu, kde v čitateli je věrohodnost daného komponenty daného mluvčího a ve jmenovateli je součet věrohodností této komponenty přes všech N mluvčích [Yoshizawa01]. Popsaný způsob je pak ekvivalentní s trénováním na datech referenčních mluvčích, přičemž výhodou je, že tato data nemusí být distribuována. Uvedený postup lze navíc použít, i když není k dispozici fonetický přepis adaptačních dat neznámého mluvčího.

V případě, že je fonetický přepis k dispozici (může být vytvořen třeba i automaticky rozpoznávačem řeči), je vhodné založit kombinaci modelů na některé z estimačních technik, například metodě ML či MAP. V tomto případě se často kombinují pouze vektory středních hodnot, ostatní parametry modelů jsou nastaveny na hodnoty apriorního modelu (například nezávislého na mluvčím). Pro každého mluvčího přitom může být vypočtena pouze jedna globální adaptační váha nebo je možné použít regresní třídy.

Kombinaci vektorů středních hodnot metodou ML lze odvodit podobně (viz [Huang02]) jako metodu MLLR. Uvažujme, že všechny komponenty všech modelů jednotlivých referenčních mluvčích jsou pro každého mluvčího rozděleny do několika vzájemně si odpovídajících regresních tříd. Regresní třídy přitom mohou být vytvořeny binárním regresním stromem vypočteným z apriorního modelu. Každý (m-tý) adaptovaný vektor středních hodnot náležící do jedné konkrétní re-

gresní třídy, která obsahuje celkem M komponent, lze pak vyjádřit jako

$$\boldsymbol{\mu}_m^{SA} = \mathbf{S}_m \boldsymbol{\lambda}, \tag{4.37}$$

kde $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, ..., \lambda_N]$ je hledaný vektor váhových koeficientů a $\mathbf{S}_m = [\boldsymbol{\mu}_m^1, \boldsymbol{\mu}_m^2, ..., \boldsymbol{\mu}_m^N]$ je matice sestavená ze vzájemně si odpovídajících vektorů středních hodnot náležících jednotlivým referenčním mluvčím. Pro nalezení vektoru $\boldsymbol{\lambda}$ pomocí metody ML se opět využívá algoritmus EM. Funkce $Q(\hat{\boldsymbol{\Phi}}, \boldsymbol{\Phi})$ má vzhledem k faktu, že hledáme nové vektory středních hodnot jako kombinaci vektorů existujících, následující tvar:

$$Q(\hat{\boldsymbol{\Phi}}, \boldsymbol{\Phi}) = c - \frac{1}{2} \sum_{k=1}^{M} \sum_{t=1}^{T} \zeta_t(k) \Big[(\mathbf{x}_t - \mathbf{S}_k \boldsymbol{\lambda})' \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_t - \mathbf{S}_k \boldsymbol{\lambda}) \Big], \tag{4.38}$$

kde $\zeta_t(k)$ je okupační pravděpodobnostní hustota k-té komponenty dané regresní třídy. Ta je vypočtena z adaptačních dat $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_1, ..., \mathbf{x}_T]$ neznámého mluvčího pomocí apriorního (například SI) modelu s kovariační maticí $\boldsymbol{\Sigma}_k$. Roznásobením a upravením rovnice 4.38 dostaneme

$$\begin{aligned} Q(\hat{\boldsymbol{\Phi}}, \boldsymbol{\Phi}) = c - \frac{1}{2} \sum_{k=1}^{M} \sum_{t=1}^{T} \zeta_t(k) \Big[& \mathbf{x}_t' \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_t \\ & -2\mathbf{x}_t' \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k \boldsymbol{\lambda} + \boldsymbol{\lambda}' \mathbf{S}_k' \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k \boldsymbol{\lambda} \Big]. \end{aligned} \tag{4.39}$$

Maximum pomocné funkce $Q(\hat{\boldsymbol{\Phi}}, \boldsymbol{\Phi})$ vzhledem k hledanému vektoru váhových koeficientů najdeme tak, že nejprve vypočteme její derivaci dle $\boldsymbol{\lambda}$

$$\frac{d}{d\boldsymbol{\lambda}} Q(\hat{\boldsymbol{\Phi}}, \boldsymbol{\Phi}) = -\frac{1}{2}(2\mathbf{U}\boldsymbol{\lambda} - 2\mathbf{v}), \tag{4.40}$$

kde

$$\mathbf{C} = \sum_{k=1}^{M} \sum_{t=1}^{T} \zeta_t(k) \mathbf{S}_k' \boldsymbol{\Sigma}_k^{-1} \mathbf{S}_k \tag{4.41}$$

a

$$\mathbf{d} = \sum_{k=1}^{M} \sum_{t=1}^{T} \mathbf{S}_k' \boldsymbol{\Sigma}_k^{-1} \zeta_t(k) \mathbf{x}_t. \tag{4.42}$$

Položíme-li vypočtenou derivaci rovnu nule, je možné přímo vyjádřit vektor $\boldsymbol{\lambda}$ jako

$$\boldsymbol{\lambda} = \mathbf{C}^{-1} \mathbf{d}. \tag{4.43}$$

Výhodou je, že odvozený vztah je výpočetně mnohem jednodušší než rovnice pro výpočet transformační matice metodou MLLR.

4.4.2 Trénovaní s adaptací a shlukováním mluvčích (CAT)

Metoda CAT (Cluster Adaptive Training - trénování s adaptací a shlukováním modelů mluvčích) ve své podstatě představuje zobecnění techniky SST. Adaptace na neznámého mluvčího zde probíhá podobně. Nové hodnoty vektorů středních hodnot jsou určeny pomocí lineární kombinace vektorů středních hodnot náležících modelům několika tzv. *„shluků“* respektive *„klastrů“* či skupin mluvčích z trénovací databáze. Váhové koeficienty pro zmíněnou kombinaci jsou počítány metodou ML. Ostatní parametry se neadaptují. Možnosti jak vytvořit a reprezentovat zmíněné shluky jsou přitom v podstatě dvě.

První jednodušší spočívá v rozdělení všech trénovacích promluv do několika skupin. Následně jsou pak pro každou skupinu (shluk) trénovacích promluv určeny metodou ML vektory středních hodnot, přičemž ostatní parametry jsou vypočteny globálně z celé trénovací databáze. Jednotlivé shluky jsou v tomto případě reprezentovány přímo svým modelem.

Ve druhém efektivnějším případě je každý shluk reprezentován pouze maticí transformující parametry obecného tzv. *„kanonického modelu”* (z anglického *canonical model*) na modely daného shluku. Tento princip je obdobný jako u metody SAT, kde se pracuje s obecným kompaktním modelem a množinou matic transformujících tento model na model odpovídající hlasovým charakteristikám každého mluvčího z trénovací databáze. Stejně jako u SAT jsou pak i v rámci CAT hledány tyto transformace a kanonický model najednou v jednom estimačním procesu.

Další informace o metodě CAT a její podrobný popis lze nalézt v [Gales98-1].

4.4.3 Metoda vlastních hlasů (EV)

V rámci metody EV (EigenVoices - vlastní hlasy) [Kuhn96] je opět prováděna pouze adaptace vektorů středních hodnot, tentokrát vážením tzv. *„kanonických“ mluvčích* (vlastních hlasů). Tito mluvčí, respektive jejich modely, jsou nalezeny tak, že se nejprve vytvoří množina SD modelů pro všechny mluvčí z trénovací databáze a následně se všechny střední hodnoty všech vzniklých modelů uspořádají do rozsáhlé matice, na kterou je nakonec aplikována PCA (Principal Component Analysis - metoda hlavních komponent). Váhové koeficienty pro lineární kombinaci hlavních komponent (vlastních hlasů) výsledného rozkladu jsou pak nalezeny metodou MLED (Maximum Likelihood Eigen-Decomposition) [Kuhn96]. Ta je v principu identická s metodou používanou pro účel lineární kombinace v rámci SST nebo CAT.

Výhodou metody vlastních hlasů je její schopnost pracovat s velmi malým množstvím adaptačních dat. Pro adaptaci stačí mít k dispozici pouze několik sekund promluvy. Metoda vlastních hlasů je proto vhodná pro rychlou adaptaci menších systémů, pracujících s omezeným slovníkem. Bohužel je ale její implementace náročná pro složitější systémy pracující s velkým počtem parametrů akustického modelu. Kromě numerických problémů při aplikaci algoritmu PCA je pro rozsáhlejší systémy složité také vytvořit prvotní sady SD modelů pro jednotlivé

mluvčí.

Metoda EV se také někdy používá jako alternativní postup pro nalezení kanonického modelu v rámci techniky CAT.

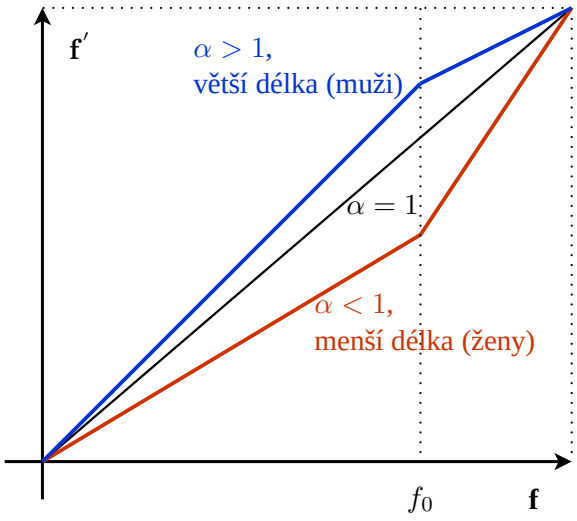
4.5 Metody normalizace dle mluvčího

Většina doposud popsaných adaptačních metod měnila hodnoty parametrů Markovových modelů tak, aby lépe vystihovaly charakteristiku řeči mluvčího, na nějž se systém adaptoval. Pro svoji funkci vyžadovaly fonetický přepis použitých adaptačních promluv a to ať již vytvořený člověkem či automaticky rozpoznávačem řeči.

Naproti tomu metody normalizace dle mluvčího pracují přímo s příznakovými vektory vypočtenými z řečového signálu a při jejich použití často není nutné znát obsah promluvy. Nejjednodušší formou tohoto přístupu je široce používaná metoda odečítání kepstrálního průměru [Huang01], která má ale pouze malý normalizační efekt.

4.5.1 Normalizace délky řečového traktu (VTLN)

Lepší výsledky než odečítáním kepstrálního průměru lze dosáhnout aplikací metody VTLN (Vocal Tract Length Normalization - normalizace délky řečového traktu) [Zhan97]. Jejím cílem je potlačit ty rozdíly v řečových charakteristikách různých mluvčích, které jsou způsobené rozdílnou délkou jejich řečového traktu. Tato délka ovlivňuje zejména hodnoty formantových frekvencí, které jsou důležité při tvorbě znělých hlásek.



Obrázek 4.1: Metoda VTLN - příklad po částech lineární warpovací funkce.

Princip metody VTLN tedy spočívá v nalezení optimální transformace stupnice frekvenční osy používané během procesu parametrizace signálu, aby transformované pozice formantů daného mluvčího odpovídaly průměrným pozicím vypočítaným univerzálně přes velkou populaci různých řečníků a tím i parametrům akustického modelu natrénovaného nezávisle na řečníkovi. Místo slova transformace se přitom v kontextu VTLN často používá pojem *„borcení“* respektive *„warpování“* (z anglického *warping*) frekvenční osy.

Warpovací funkce frekvenční osy se obvykle definuje jako po částech lineární (viz obr. 4.1) popřípadě bilineární (viz např. [Psutka06]) a to tak, aby nebylo nutné určovat hodnoty transformovaných frekvencí mimo původní frekvenční rozsah. V obou případech je warpovací funkce

$$f^{'} = w_{\alpha}(f) \tag{4.44}$$

závislá na faktoru α, který je různý pro každého mluvčího. Cílem adaptace metodou VTLN je právě určení tohoto parametru na základě dostupných adaptačních dat. Zlomová frekvence f_0, při které se mění sklon warpovací funkce, se přitom volí tak, aby odpovídala minimálně frekvenci třetího formantu mluvčího s teoreticky přesně průměrnou délkou řečového traktu ($\alpha = 1$).

V praxi se pro nalezení optimálního warpovacího faktoru používá nejčastěji postup, během kterého je v omezeném intervalu postupně měněna hodnota α, vypočítávány nové vektory příznaků z adaptační promluvy a následně určována věrohodnost, že dané příznakové vektory byly vygenerovány na mluvčím nezávislým akustickým modelem. Warpovací faktor odpovídající nejvyšší dosažené hodnotě věrohodnosti je pak pro daného mluvčího vybrán jako optimální. V případě, že není k dispozici přepis adaptační promluvy, lze pro určení optimálního warpovacího faktoru použít sadu GMM modelů natrénovaných pro různé hodnoty α.

KAPITOLA 5

METODY HODNOCENÍ ÚSPĚŠNOSTI ROZPOZNÁVÁNÍ ŘEČI A ADAPTACE

Hodnocení úspěšnosti rozpoznávání řeči

Pro účely definice veličin vyjadřujících úspěšnost adaptace na mluvčího je nejprve třeba připomenout míry používané pro hodnocení úspěšnosti rozpoznávání řeči:

V případě rozpoznávání spojité řeči je textový výstup z klasifikátoru porovnán s referenčním textem algoritmem DTW (Dynamic Time Warping - dynamické borcení času) [Sakoe78], přičemž slova rozpoznaná správně jsou označena jako HIT (H) a slova rozpoznaná špatně jako SUBSTITUCE (S). Slova, která v přepisu chybí, jsou zahrnuta mezi DELECE (D). Ta, která v rozpoznaném textu naopak přebývají, jsou označena jako INZERCE (I). Celkový počet slov v referenčním textu pak bývá označován symbolem N. Příklad porovnání referenčního a rozpoznaného textu je znázorněn v tab. 5.1.

V případě rozpoznávání izolovaných slov je situace jednoduší. Pouze se vyhodnocuje, bylo-li dané slovo rozpoznáno správně (HIT), nebo nikoli (pak jde o SUBSTITUCI).

Pomocí výše definovaných veličin lze vypočítat tři základní míry úspěšnosti rozpoznávání řeči: přesnost (ACC - accuracy), korektnost (CORR - correctness) a chybovost respektive procento chybně rozpoznaných slov (WER - Word Error Rate).

Veličina WER uvádí procento slov, jež je nutné opravit, to znamená doplnit, přepsat nebo smazat, aby bylo dosaženo referenčního textu. Je definována jako:

$$WER = \frac{S + D + I}{N} 100\ \% \tag{5.1}$$

ACC je doplňkem WER do 100 %:

$$ACC = 100\% - WER \tag{5.2}$$

CORR udává pouze procento správně rozpoznaných slov:

$$CORR = \frac{H}{N} 100\% \tag{5.3}$$

U uvedeného příkladu v tab. 5.1 vychází ACC=62,5 %, WER=37,5 % a CORR=62,5 %.

reference:	uložit	ji	můžeme	teprve	tehdy	když	jí	-
rozpoznaný text:	důležitý	-	můžeme	teprve	tehdy	když	jí	od
ohodnocení:	S	D	H	H	H	H	H	I

Tabulka 5.1: Ukázka porovnání referenčního a automaticky rozpoznaného textu metodou DTW.

Hodnocení úspěšnosti adaptace na mluvčího

Úspěšnost adaptace na mluvčího je posuzována dosaženým zlepšením veličiny ACC nebo častěji dosaženým snížením chybovosti systému (veličina WERR - Word Error Rate Reduction). Zlepšení přesnosti či snížení chybovosti může být přitom udáváno v absolutních číslech nebo může být vztaženo relativně k původní hodnotě neadaptovaného systému.

Výhodou relativní míry je to, že lépe vystihuje známý fakt, že je složitější zlepšit výsledky systému, jehož původní přesnost rozpoznávání je vysoká (respektive chybovost nízká). Například zvýšit úspěšnost rozpoznávání z 90 % na 95 %, absolutně pouze o 5 %, se může zdát na první pohled jednodušší, než zvýšit úspěšnost systému z hladiny 50 % na 75 %, absolutně o 25 %, ačkoli ve skutečnosti je první případ pro řešení spíše o něco obtížnější. Naproti tomu relativní zlepšení je v obou případech 50 %.

KAPITOLA 6

PRAKTICKÉ ASPEKTY ADAPTACE NA MLUVČÍHO SE ZNÁMOU IDENTITOU

Cílem této kapitoly je popsat praktické metody, které byly navrženy, použity a vyhodnoceny pro účely adaptace na mluvčího, jehož identita je v době rozpoznávání jeho promluvy známa. Jde tedy například o uživatele diktovacího systému, který má vytvořen svůj uživatelský profil, jehož součástí je i adaptovaný akustický model. Řeč tak bude především o metodách řízené adaptace, neboť jak již bylo naznačeno, v případě, že je mluvčí během rozpoznávání řeči znám, je většinou možné získat od něj předem akustická data, jejichž textový přepis může být připraven. To lze v praxi zajistit například tak, že každý uživatel diktovacího systému musí po jeho instalaci přečíst stejný text.

6.1 Vytvořený adaptační software

Jedním z prvních a důležitých cílů této disertační práce, jehož splnění zabralo nemálo času, bylo vytvořit vlastní adaptační software, který by mohl být distribuován spolu s existujícími rozpoznávacími systémy vyvinutými na TUL, které jsou dlouhodobě používaný jednou konkrétní osobou. Jedná se především o systém MyVoice pro hlasové ovládání počítače, systém MyDictate pro diktování izolovaných slov a jeho obdobu pro diktování plynulé. Důvod pro vývoj vlastního softwaru je přitom ten, že programy používané pro adaptaci ve většině světových laboratoří mají licenčně omezené použití (typicky například software HTK) nebo nabízejí pouze omezené spektrum funkcí.

Jako první byl vytvořen program obsahující vlastní implementaci Baum-Welchova algoritmu pracujícího s modely fonémů, který je nedílnou součásti většiny adaptačních metod. Následně byl tento program rozšířen o modul umožňující provádět adaptaci všech parametrů Markovových modelů metodou MAP.

Jako druhá přišla na řadu metoda MLLR, která byla implementována tak, aby bylo možné volit si ze tří alternativních forem regresního stromu:

1. **binární regresní strom**
   Je vytvořený ze všech komponent všech stavů předloženého akustického

modelu automaticky pomocí klastrování. Jako kritérium pro klastrování je použita Euklidovská vzdálenost mezi vektory středních hodnot jednotlivých komponent.

2. **expertní regresní strom**
   Je vytvořený z daného modelu na základě expertních znalostí o daném jazyce - všechny fonémy daného jazyka mohou být například rozděleny do několika skupin na základě jejich fonetické podobnosti.

3. **kombinace obou předchozích způsobů**
   V tomto případě je několik počátečních uzlů stromu vytvořeno na základě expertních znalostí a tyto uzly jsou následně automaticky rozděleny do binární struktury.

Kromě adaptace vektorů středních hodnot umožňuje výsledný software adaptovat metodou MLLR také hodnoty rozptylů dle vztahu 4.27. Numericky obtížný výpočet inverzní matice ve vztahu 4.24 je přitom vyřešen aplikací metody SVD (Singular Value Decomposition) dle [Press02].

6.2 Metody používané pro zpracování a modelování řečového signálu

V rámci Laboratoře počítačového zpracování se autor této práce podílel na celé řadě rozsáhlých experimentů na různých typech úloh, jejichž cílem bylo pokaždé najít nejlepší možnou metodu pro zpracování a modelování řečového signálu. Na základě výsledků všech experimentů pak byly stanoveny níže popsané standardy, které se nyní používají ve většině systémů vyvíjených na TUL a byly proto aplikovány i v rámci všech experimentů prezentovaných v této disertační práci.

Zpracování a parametrizace signálu

Zpracování akustického signálu je prováděno standardně metodou MFFC (Mel-Frequency Cepstral Coefficients - melovské frekvenční kepstrální koeficienty) [Huang01], přičemž použitý příznakový vektor obsahuje celkem 39 parametrů - prvních 13 MFFC koeficientů a jejich první a druhé diference. Vzorkovací frekvence je 16 kHz.

Struktura používaných akustických modelů

Jako akustické modely slouží třístavové levopravé Markovovy modely českých monofonů [Nouza97-1] a několika ruchů. Těchto celkem 48 modelů obsahuje v každém stavu maximálně 100 Gaussových komponent, přičemž jejich skutečný počet závisí pro každý stav na množství dat dostupných během trénování. Výstupní pravděpodobnostní hustota každé komponenty je spojitá s diagonální kovariační

maticí. Trénovací řečová databáze obsahuje cca 50 hodin promluv namluvených několika sty různými mluvčími.

Důvod, proč je akustické modelování založeno právě na monofonech s velkým počtem komponent v každém stavu a nikoli na trifonech, které by měly být dle teoretických předpokladů obecně přesnější, není ten, že by snad vytvořené systémy a navržené adaptační metody nemohly s trifony pracovat, ale pouze dosažené experimentální výsledky. Za použití přesného jazykového modelu a rozsáhlého slovníku vychází chybovost rozpoznávání u většiny systému vyvinutých na TUL téměř stejně s monofony i trifony, a používat trifony, jejichž počet je mnohem větší, rozpoznávání s nimi pomalejší a adaptace náročnější, pak nedává žádný praktický smysl.

6.3 Úloha rozpoznávání izolovaných slov

Dalším cílem disertační práce bylo najít nejlepší adaptační techniku, kterou by bylo možné pro češtinu prakticky aplikovat v úloze rozpoznávání izolovaných slov (IWSR - Isolated-Word Speech Recognition) a kterou by šla adaptace provádět při pevně daném počtu speciálně vybraných adaptačních slov, neboť tato konfigurace nejvíce odpovídá charakteru uvažované úlohy a možným aplikacím (hlasové ovládání, diktování).

Pro tento účel bylo experimentováno s různými variantami dvou nejvýznamějších adaptačních přístupů, metody MAP a metody MLLR. V rámci jednotlivých experimentů byl použit diktovací systém vyvinutý v Laboratoři počítačové zpracování na TUL [Nouza05]. Jeho slovník obsahoval 500 tisíc nejčastějších českých slov a systém pracoval s unigramovým jazykovým modelem.

6.3.1 Navržená strategie tvorby sady adaptačních slov

Prvním úkolem bylo stanovit výše zmíněnou sadu adaptačních slov, na které by mohlo být provedeno srovnání jednotlivých metod a která by se poté v jednotlivých systémech pro adaptaci skutečně využívala. Obecně přitom platí, že slova by měla být do každé adaptační sady vybíraná dle následujících důležitých kritérií:

1. S ohledem na frekvenční analýzu českého jazyka, aby byla pokryta nejčastěji se vyskytující slova.
2. Aby byly zastoupeny všechny uvažované fonémy v co nejrůznějším kontextu.
3. Aby byla zastoupena všechna důležitá řídicí a klíčová slova daného systému.
4. Aby sada obsahovala také slova, která jsou jen obtížně rozpoznatelná, například předložky a spojky.
5. Aby vybraná slova byla pokud možno jednoduše a jednoznačně vyslovitelná.

6. Aby celkový počet slov byl co nejmenší, neboť není vhodné uživatele zbytečně obtěžovat dlouhotrvajícím čtením slov.

V rámci provedených experimentů byl počet adaptačních slov nakonec nastaven na hodnotu 300, protože namluvení uvedeného množství netrvá více než deset minut a zároveň je 300 slov dostatečných z hlediska množství dat potřebného pro kvalitní adaptaci. Experimenty s různým množstvím adaptačních dat jsou obsahem kapitol 6.3.7 a 6.4.1.

Aby adaptační sada splňovala všechna výše uvedená kritéria, byly navrženy dvě odlišné strategie, jak do ní přidávat slova:

První **strategie** zajišťovala **pokrytí důležitých slov**:

- Třikrát byla přidána všechna česká slova obsahující pouze jeden foném, například slovo “a”.
- Dvakrát byly přidány všechny důležité řídicí povely daného systému, například VYMAŽ_SLOVO.
- Ze slovníku rozpoznávače bylo vybráno třicet slov s největší frekvencí výskytu (hodnotou unigramového faktoru) a větším počtem fonémů než jedna.

Následně byl navržen algoritmus zajišťující pokrytí všech fonémů, přičemž slova byla podle tohoto algoritmu vybírána ručně:

Algoritmus zajišťující pokrytí všech fonémů

krok1: *Na všech doposud vybraných slovech byla spočítána četnost výskytu jednotlivých fonémů a byl vybrán foném s nejnižší četností.*

krok2: *Do adaptační sady bylo přidáno slovo s nejvyšší frekvencí výskytu, které zároveň nejvíce vyhovovalo všem třem následujícím podmínkám:*

1. *obsahovalo daný monofon,*
2. *co nejvíce se lišilo od slov, která byla již v adaptační sadě obsažena,*
3. *mělo jednoznačnou a jednoduchou výslovnost.*

Oba dva předchozí kroky byly poté opakovány tak dlouho, dokud nebylo vybráno stanovené množství 300 slov.

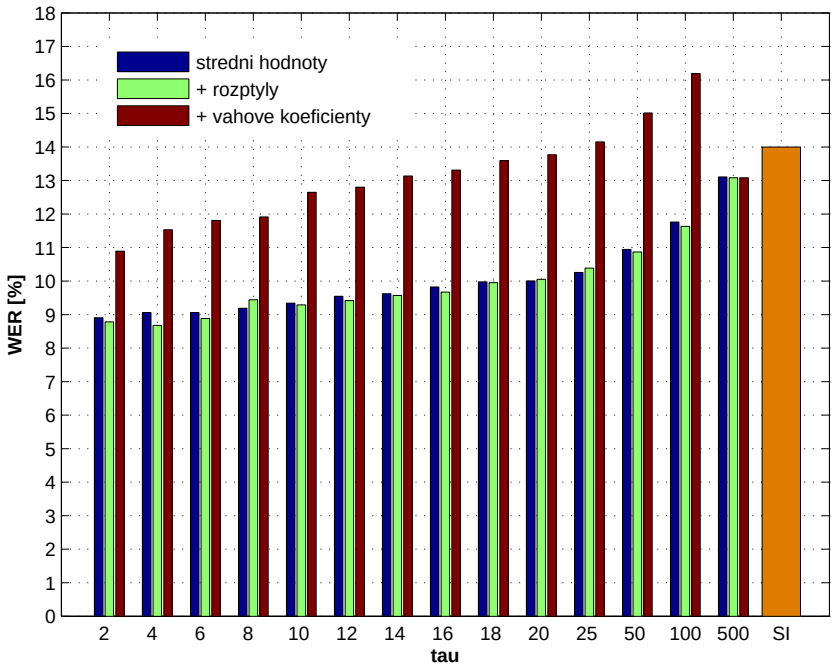
Experimentální výsledky dosažené použitím vytvořené sady slov a různých adaptačních technik jsou obsahem následujících kapitol 6.3.2 až 6.3.6. Cílem přitom bylo otestovat jednotlivé metody pro různé hodnoty jejich parametrů a ověřit,

na jakou *hladinu* lze jejich aplikací snížit chybovost rozpoznávání. Při prezentaci výsledků uvnitř textu je proto dána přednost přehlednějšímu souhrnnému grafickému znázornění, protože výsledky různých variant téže metody se od sebe většinou liší číselně jen v řádech desetin procenta, což je vzhledem k velikosti testovací množiny málo významný rozdíl. Pro úplnost jsou přesto všechna čísla uvedena v příslušných tabulkách v příloze v závěru práce.

Experimenty 6.3.2 až 6.3.6 byly vyhodnoceny na testovací databázi obsahující celkem 3929 slov, která byla namluvena 4 mluvčími - dvěma muži a dvěma ženami. Každý nadiktoval dva články. Jeden se zaměřením na sport a druhý na domácí zpravodajství. Průměrná základní chybovost rozpoznávání při použití SI modelů byla pro tyto mluvčí 14 %, přičemž pro nejhoršího činila 21 % a pro nejlepšího 6,8 %. V testovací množině byli zastoupeni mluvčí s různě dobrou výslovností. Počet slov mimo slovník rozpoznávače byl při všech experimentech menší než 1 %.

6.3.2 Adaptace metodou MAP

První z provedených experimentů byl zaměřen na adaptaci metodou MAP. Adaptace byla prováděna s různou hodnotou váhového koeficientu τ, která byla stejná pro všechny adaptované parametry všech komponent systému.



Obrázek 6.1: IWSR - výsledky adaptace různých parametrů metodou MAP s odlišnými hodnotami adaptačního váhového koeficientu τ.

Adaptovány byly nejprve pouze vektory středních hodnot, poté střední hodnoty a rozptyly a nakonec byla adaptace rozšířena i na váhové koeficienty jednotlivých komponent. Jako apriorní byly použity parametry modelu nezávislého na mluvčím, který byl natrénován metodou maximální věrohodnosti.

Výsledky experimentu jsou znázorněny na obr. 6.1 a podrobně uvedeny v příloze A.3. Byly vypočítány průměrem přes všechny 4 mluvčí. Ukazují, že adaptaci lze při uvažovaném množství adaptačních dat provádět pouze pro vektory středních hodnot a rozptyly. Rozšíření adaptace ze středních hodnot na rozptyly přitom ale dává pouze zanedbatelně lepší výsledky. Naopak rozšíření adaptace i na váhové koeficienty komponent vede ke zvýšení chybovosti rozpoznávání.

V souhrnu lze říci, že adaptací metodou MAP lze za použití 300 adaptačních slov snížit procento chyb systému z hladiny 14 % na hladinu 9 % - tedy relativně cca o 35 %. Adaptační váhový koeficient τ je přitom vhodné nastavit na hodnoty v rozsahu 2 - 20, přičemž rozdíly pro jednotlivé hodnoty v uvedeném rozmezí lze zanedbat.

6.3.3 Adaptace metodou MLLR

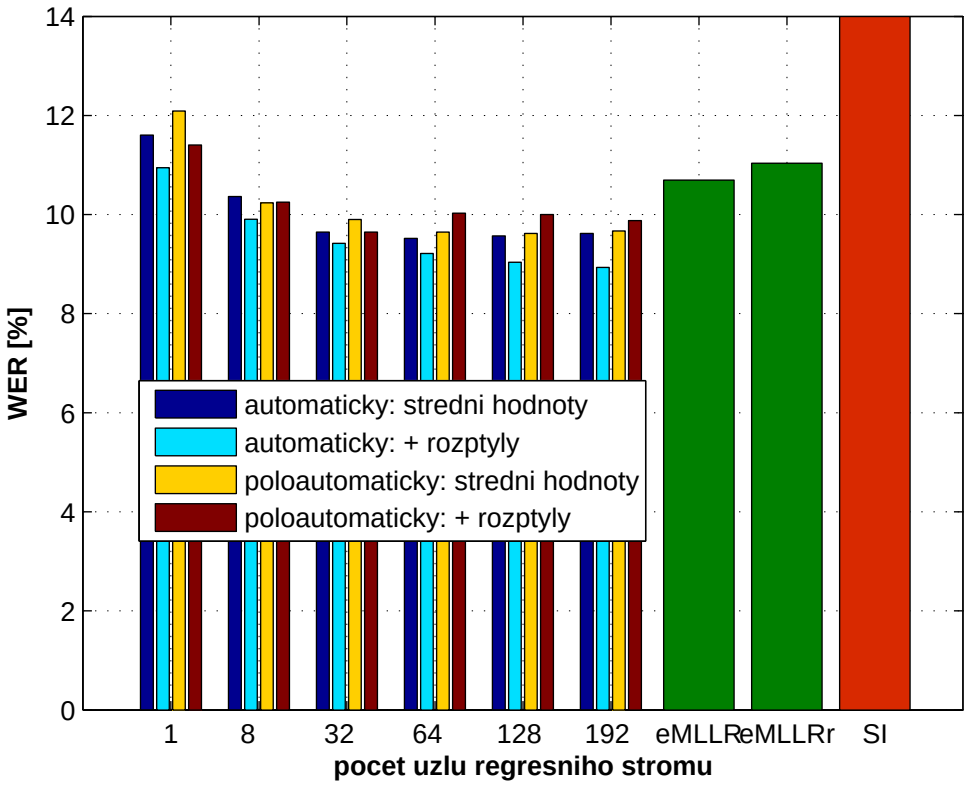
V pořadí druhý experiment byl zaměřen na adaptaci metodou MLLR. Adaptace byla prováděna s použitím regresního stromu vytvořeného třemi různými způsoby tak, jak to umožňuje vytvořený adaptační software:

1. Plně automaticky - pomocí klastrování.
2. Poloautomaticky - první dva uzly byly inicializovány rozdělením všech fonémů na hlásky a zbylé ruchy. Následně bylo opět aplikováno klastrování.
3. Expertně - rozdělením všech modelů na 8 kategorií dle fonetické podobnosti. Vytvořené kategorie jsou uvedeny v tab. 6.1.

skupina	obsažené modely
samohlásky	a,á,e,é...
znělé frikativy	z,ž,v...
neznělé frikativy	s,š,f...
znělé explozívy	b,g,d...
neznělé explozívy	p,t,ť...
zbylé hlásky	k,l,m...
ruchy	nádech, ehm...
ticho	model ticha

Tabulka 6.1: Rozdělení českých monofonů do akusticky blízkých skupin.

V prvních dvou případech měl automaticky vytvořený strom různý počet uzlů od 1 do 256. V případě jednoho uzlu byla přitom hledána pouze jedna společná globální transformace pro všechny komponenty systému. Ve všech třech případech byly adaptovány nejprve střední hodnoty a poté byla adaptace rozšířena i na rozptyly.



Obrázek 6.2: IWSR - výsledky adaptace různých parametrů metodou MLLR při použití několika typů regresních stromů.

Experiment byl vyhodnocen na stejné testovací databázi jako v předchozí kapitole a uváděné výsledky (viz obr. 6.2 a příloha A.1) byly opět vypočteny průměrem přes všechny 4 testovací mluvčí. Zkratka “eMLLR” ve výsledném grafu odpovídá variantě MLLR, kdy byly všechny monofony rozděleny do 8 skupin a následně byly adaptovány vektory středních hodnot. Zkratka “eMLLRr” pak označuje stejný postup s tím, že adaptovány byly tentokrát i rozptyly.

Výsledky experimentu ukázaly, že metodou MLLR lze při uvažované adaptační sadě dosáhnout jen zanedbatelně horších výsledků než metodou MAP. Chybovost systému se opět podařilo snížit z hladiny 14 % na cca 9 %. Z jednotlivých zkoumaných variant MLLR dopadl nejlépe postup, při kterém byl použit plně automaticky vytvořený binární regresní strom, přičemž počet uzlů stromu byl větší než 32. Pouze

při této variantě se navíc ukázalo vhodné rozšířit adaptaci i na hodnoty rozptylů. U ostatních variant mělo toto rozšíření negativní vliv.

Z porovnání druhého sloupce (počet uzlů je osm) s variantou expertní MLLR (celkový počet uzlů je také osm) navíc vyplývá, že provedené ruční rozdělení modelů do regresních tříd dává horší výsledky než automatický přístup pomocí klastrování.

6.3.4 Adaptace kombinací metod MAP a MLLR

V rámci experimentu zaměřeného na kombinaci metody MAP a metody MLLR bylo použito takové nastavení parametrů obou metod, které dávalo v předchozích experimentech nejlepší výsledky. Nejprve tedy byla provedena adaptace metodou MLLR za použití binárního regresního stromu, přičemž počet uzlů regresního stromu byl 128. Následně byly parametry transformované metodou MLLR použity jako apriorní pro metodu MAP, přičemž ta byla aplikována s různou hodnotou adaptačního váhového koeficientu τ stejnou pro všechny adaptované parametry.

V první fázi byly uvedeným postupem adaptovány pouze střední hodnoty a až poté i rozptyly. Z dosažených výsledků (tab. 6.2) vyplývá, že *kombinace obou přístupů nepřinesla v uvažované úloze žádné další významné zlepšení*. Chybovost adaptovaného systému je opět na hladině 9 %. I v tomto případě se ale jako pozitivní ukázalo rozšíření adaptace ze středních hodnot i na rozptyly.

hodnota τ adaptované parametry	2	4	10	16	20	25	50	100
střední hodnoty	9,4	9,2	9,1	9,2	9,3	9,3	9,4	9,5
střední hodnoty + rozptyly	9,1	8,8	8,8	8,8	8,9	8,9	8,9	9,0

Tabulka 6.2: IWSR - WER [%] po adaptaci různých parametrů kombinací metod MAP a MLLR při použití odlišných hodnot adaptačního váhového koeficientu τ (SI WER = 14,0 %).

6.3.5 Vliv použití GD modelů jako apriorních parametrů

Cílem dalšího experimentu bylo ověřit, jakých výsledků lze dosáhnout, jsou-li jako apriorní parametry pro adaptaci použity modely závislé na pohlaví mluvčího. Ty by totiž měly odpovídat charakteristikám hlasu každého mluvčího více než modely nezávislé na mluvčím, které byly použity ve všech předchozích experimentech. Experiment byl proveden s metodou MAP, metodou MLLR i použitím kombinace obou metod, přičemž ve všech případech bylo použito nejlepší nastavení jejich parametrů dle předchozích experimentů. Metoda MAP tak byla aplikována s váhovým koeficientem nastaveným na hodnotu 4, v rámci MLLR byl opět použit binární regresní strom se 128 uzly a při kombinaci obou metod byl adaptační váhový koeficient nastaven na hodnotu 10.

Výsledky experimentu jsou uvedeny v tab. 6.3. Symbol “r” zde značí, že adaptace byla pro danou metodou prováděna nejen pro vektory středních hodnot, ale také pro rozptyly. Chybovost rozpoznávání za použití SI a GD modelů přitom byla 14 % respektive 12,6 %.

apriorní parametry	MAP	MAPr	MLLR	MLLRr	MLLRaMAP	MLLRaMAPr
SI	9,1	8,5	9,1	9,1	9,1	8,8
GD	9,1	8,6	9,1	8,9	8,7	8,5

Tabulka 6.3: IWSR - hodnoty WER [%] po adaptaci různými metodami za použití na pohlaví závislých (GD) a nezávislých (SI) modelů jako apriorních parametrů.

Z uvedených výsledků vyplývá, že *GD modely jako apriorní parametry mají na výsledky adaptace v dané úloze pozitivní vliv*, zejména při použití kombinace metod MAP a MLLR. *Dosažené zlepšení ovšem není velké*.

Dále je možné konstatovat, že nejlepší výsledky lze při použití předem připravené sady 300 adaptačních slov dosáhnout aplikací metody MAP, a to ať už samostatně nebo v kombinaci s metodou MLLR. V tomto případě je ovšem vhodné zvolit jako apriorní parametry GD modely.

6.3.6 Vliv použité sady adaptačních slov

Ve všech předchozích experimentech byla pro adaptaci použita sada 300 speciálně vybraných adaptačních slov (viz kap. 6.3.1). Cílem následujícího experimentu, provedeného na stejné testovací množině, bylo ukázat, jaký vliv má použití těchto slov oproti adaptaci na běžném textu.

Každý že čtyř testovacích mluvčích pro tento účel nadiktoval jeden novinový text čítající 300 slov. Namluvená slova pak byla použita pro adaptaci různými metodami podobně jako v předchozím experimentu. Jako apriorní parametry byly použity GD modely. Výsledky experimentu jsou uvedeny v tab. 6.4.

	MAP	MAPr	MLLR	MLLRr	MLLRaMAP	MLLRaMAPr
adaptační sada	9,1	8,6	9,1	8,9	8,7	8,5
novinový článek	9,1	9,1	9,3	9,5	9,3	9,5

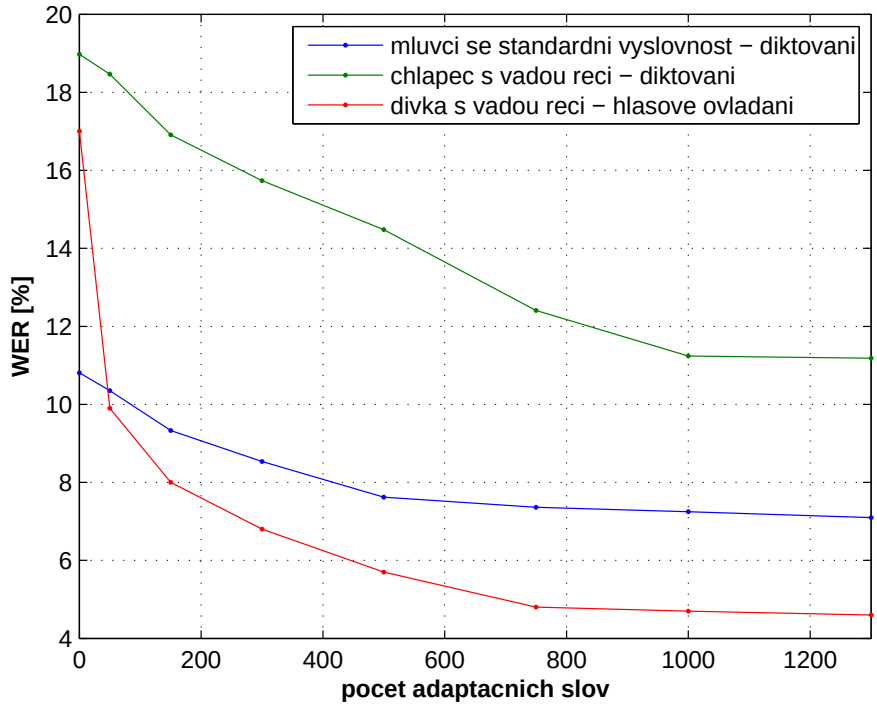
Tabulka 6.4: IWSR - porovnání hodnot WER [%] po adaptaci založené na použití běžného textu a speciálně připravené sady adaptačních slov.

Z výsledků vyplývá, že *použití speciálně vybraných slov vede u všech metod k lepším výsledkům*. U kombinace metod MAP a MLLR dokonce o celé jedno procento. V případě použití běžného textu chybovost systému neklesla ani v jednom

případě pod hranici 9 %. Zároveň se v tomto případě ukázalo, že není vhodné provádět adaptaci rozptylů. Výsledky jsou pak horší, než když jsou adaptovány pouze vektory středních hodnot.

6.3.7 Adaptace na mluvčího s vadou řeči

Důležitou aplikační oblastí, kde nacházejí metody adaptace své uplatnění, je bezpochyby problematika adaptace na hlas mluvčích s vadou řeči. Ne že by snad tito mluvčí byli typickými uživateli systémů rozpoznávání řeči, spíše naopak, ale problémy s výslovností se bohužel často vyskytují u motoricky handicapovaných lidí (například quadruplegiků), pro které může být rozpoznávání řeči velice užitečné. Problém se špatnou výslovností nastává například u osob, jejichž handicap je spojen se zvýšeným svalovým napětím v těle, které negativně ovlivňuje i funkci jejich řečových orgánů. Následující experiment (viz obr. 6.3 a příloha A.2) proto ukazuje, jakých výsledků lze pomocí adaptace dosáhnout právě u osob s motorickým handicapem doprovázeným mírnou vadou řeči.



Obrázek 6.3: IWSR - porovnání úspěšnosti adaptace na mluvčího se standardní výslovností a handicapované osoby s vadou řeči.

Experiment byl proveden na základě zvukových záznamů získaných od handicapované dívky, která již více než dva roky úspěšně pracuje se systémem MyVoice [Nouza05-1] pro hlasové ovládání počítače, a handicapovaného chlapce, který již několik měsíců testuje obdobný software pro hlasové diktování do počítače [Cerva07]. Charakter řeči dívky (quadruplegičky) lze označit jako dýchavičný vlivem nedostatečné funkce plic. U chlapce se při vyslovování jednotlivých slov negativně projevuje zvýšená svalová tenze.

Použité nahrávky byly zaznamenány během praktického používání obou zmíněných programů pomocí funkce automatického ukládání nahrávek. Následně byla provedena jejich analýza a fonetický a textový přepis. Celkem tak bylo pro adaptaci na každého z mluvčích připraveno až 1300 slov a dalších 1500 slov bylo použito pro testování. Adaptace byla provedena použitím kombinace metod MAP a MLLR, GD modelů jako apriorních parametrů a adaptovány byly pouze střední hodnoty.

Pro srovnání byl experiment s diktovacím systémem proveden i pro mluvčího s průměrně dobrou výslovností. U systému hlasového ovládání není třeba za běžných okolností žádnou adaptaci provádět, neboť chybovost systému je díky charakteru úlohy standardně nižší než 3 %.

Z výsledků experimentů je patrné, že *chybovost rozpoznávání u mluvčích s vadou řeči klesá s rostoucím množstvím adaptačních dat pomaleji než u mluvčích se standardní výslovností*. Zatím co pro adaptaci v diktovacím systému lze pro běžného mluvčího použít 300 až maximálně 500 slov, v případě handicapované osoby je třeba slov 1000. Rovněž u jednoduššího systému MyVoice byla chybovost dostatečně snížena až po použití více než 600 slov. *U obou handicapovaných osob došlo k vysoké relativní redukci chybovosti*. U systému MyVoice z hladiny 17 % na hladinu 5 % (tedy o 70 %), u diktovacího systému z 19 % na 11 % (o více než 40 %).

Celkově lze tedy říci, že adaptace má pro osoby s vadou řeči větší význam než pro ostatní mluvčí. Bohužel ji ale nelze použít v případech, kdy je řeč dané osoby až příliš nesrozumitelná.

6.3.8 Adaptace na mluvčího a mezijazyková adaptace

Cílem posledního experimentu provedeného v rámci úlohy rozpoznávání izolovaných slov je ukázat (spíše pro zajímavost), jak lze pomocí adaptace na mluvčího zlepšit výsledky systému, v kterém jsou v průběhu rozpoznávání používány akustické modely natrénované původně pro odlišný jazyk, a jehož slovník vznikl pouze namapováním těchto původních modelů na fonémy daného nového jazyka. V rámci této disertační práce je takový systém označován jako systém vzniklý *mezijazykovou adaptací* (z anglického *cross-lingual adaptation*).

První experiment (viz tab. 6.5) byl proveden na systému hlasového ovládání MyVoice, který byl pokusně transformován tak, aby umožňoval hlasové ovládání handicapovaným osobám i ve španělštině, přičemž v současné době probíhá jeho transformace na praktičtější a nám bližší jazyk - slovenštinu. Přesný postup provedené transformace je pak popsán v článku [Callejas07]. V rámci tohoto expe-

rimentu namluvil španělský rodilý mluvčí (dívka) španělskou obdobu české sady 300 adaptačních slov a dalších více než 1000 slov pro testování.

	GD modely	SA modely
WER [%]	9,0	3,4

Tabulka 6.5: IWSR - porovnání chybovosti španělské verze systému MyVoice (vytvořeného mezijazykovou adaptací z češtiny) před a po adaptaci na mluvčího.

V rámci druhého složitějšího experimentu (viz tab. 6.6) pak byla změřena i úspěšnost adaptace v úloze rozpoznávání izolovaných slov s velkým slovníkem. Pro tento účel byla rozšířena sada adaptačních slov na celkem 620 položek a bylo nadiktováno 1200 nových slov pro testování. Experiment byl proveden s různě velkým slovníkem čítajícím 10 až 146 tisíc nejčastějších španělských slov.

počet slov ve slovníku	10 000	46 000	85 000	146 000
GD modely	56,2	49,8	49,5	49,0
SA modely	34,6	22,0	21,1	20,3

Tabulka 6.6: IWSR - porovnání hodnot WER [%] pro španělský diktovací systém (vytvořený mezijazykovou adaptací z češtiny) před a po adaptaci na mluvčího.

Výsledky obou experimentů ukázaly, že *adaptace na mluvčího je v systémech vytvořených mezijazykovou adaptací velice prospěšná* respektive téměř nezbytná, neboť *kromě adaptace na charakteristiky konkrétního mluvčího dochází i k adaptaci na výslovnost daného jazyka*. Tím pádem se podařilo velmi významně snížit chybovost rozpoznávání obou systémů (relativně cca o 60 %) až na hladinu, která téměř odpovídá průměrnému českému mluvčímu s neadaptovaným akustickým modelem.

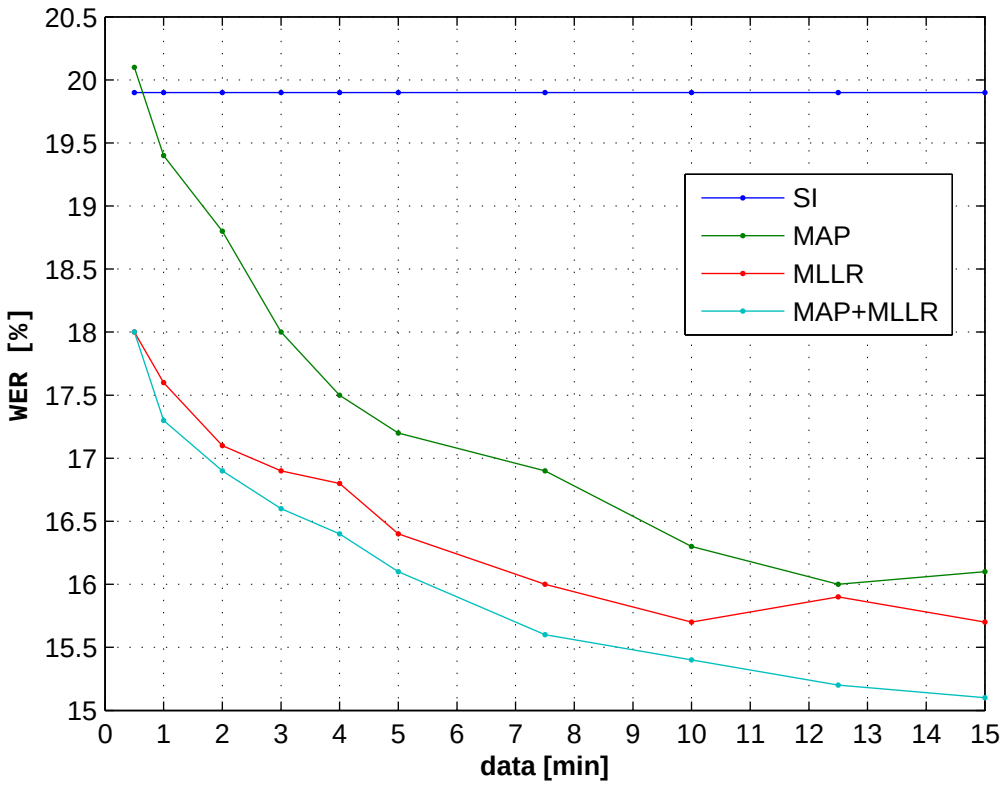
6.4 Úloha rozpoznávání plynulé řeči

Po úloze rozpoznávání izolovaných slov byla další fáze experimentálního vývoje a vyhodnocování zaměřena na složitější úlohu rozpoznávání plynulé řeči. Cílem bylo najít sadu nejvhodnějších technik pro adaptaci na mluvčího v systému pro plynulé rozpoznávání řeči (Continuous Speech Recognition - CSR), který byl vyvinutý pro češtinu v Laboratoři počítačového zpracování řeči na TU v Liberci. Jeho slovník obsahuje 312 490 nejčastějších českých slov a jazykový model je založen na vyhlazeném bigramovém modelu, který je vypočítán z textového korpusu obsahujícího více než 3 GB textů. Zmíněný systém nachází své uplatnění zejména v rámci softwaru pro plynulé diktování a softwaru pro přepis mluvených zvukových záznamů, například televizních a rozhlasových pořadů.

6.4.1 Porovnání úspěšnosti vybraných metod

V pořadí první provedený experiment je opět zaměřen na srovnání jednotlivých adaptačních metod, tentokrát ovšem v závislosti na množství použitých dat, protože vliv konkrétních parametrů jednotlivých metod na celkové výsledky je, podobně jako v předchozí úloze, v průměru jen velmi malý. Znalost závislosti jednotlivých metod na množství dat je navíc důležitá pro použití adaptace v systému pro přepis televizních a rozhlasových pořadů, kde je pro jednotlivé mluvčí, na než je prováděna adaptace, k dispozici velice rozdílné množství dat, která jsou navíc i různě foneticky bohatá (viz kap. 7).

Experimenty byly provedeny pro metodu MAP, MLLR a kombinaci obou metod, přičemž jejich parametry byly vždy nastaveny na hodnoty, které se ukázaly jako nejlepší v předchozích kapitolách. Adaptovány byly pouze střední hodnoty, neboť pro adaptaci rozptylů nebyl ve všech testovaných případech k dispozici dostatek dat a adaptační věty (běžné novinové texty) navíc nebyly vytvořeny podle žádných speciálních kritérií, což znamená, že v nich všechny fonémy nemusely být dostatečně zastoupeny.



Obrázek 6.4: CSR - porovnání výsledků adaptace různými metodami pro různé množství použitých adaptačních dat (od 0,5 do 15 min).

Testovací databáze obsahovala celkem 1127 vět, které byly namluveny celkem 6 mluvčími. Počáteční úroveň chybovosti rozpoznávání při použití SI modelů se pro jednotlivé mluvčí lišila v intervalu 9 % až 29 %. V testovací množině tak byli opět zastoupeni mluvčí s různě kvalitní výslovností. Uvedené množství testovacích vět obsahovalo více než 16 000 slov a délka všech vět dohromady byla 130 minut.

Výsledky experimentu (viz obr. 6.4 a příloha A.4) potvrdily teoretická očekávání, že *chybovost rozpoznávání lze při různém množství adaptačních dat snížit nejvíce použitím metody MLLR a následnou aplikací metody MAP*. Výhodou tohoto přístupu je totiž skutečnost, že modely neobsažené v adaptační promluvě jsou adaptovány díky shlukování do regresních tříd v rámci metody MLLR a modely, pro které je adaptačních dat dostatek, jsou následně adaptovány metodou MAP, která zajišťuje konvergenci k teoreticky nejlepšímu SD modelu. Například *při použití 10 minut, které odpovídají v průměru cca 110 adaptačním větám, byla chybovost rozpoznávání snížena z hladiny 20 % na hladinu 15 %* - tedy relativně cca o 25 %. Uvedené výsledky rovněž odpovídají výsledkům uváděným v [Huang01] pro angličtinu a v [Železný01] pro češtinu, i když zde jsou výsledky hůře porovnatelné, neboť jako akustické modely byly používány trifony a tak hlavně metoda MAP dávala v principu horší výsledky.

Výsledky dále ukázaly, že ani u jedné z metod nemá příliš smysl použít větší množství dat než zmíněných 10 minut, neboť dodatečně dosažené zlepšení je pak malé v porovnání s pracností namlouvání adaptačního textu. U metody MAP pak jako u jediné došlo při použití malého množství dat (0,5 min) k mírnému zhoršení rozpoznávacího skóre.

6.4.2 Redukce počtu komponent adaptovaného systému

Systém natrénovaný jako na řečníkovi nezávislý obsahuje většinou velké množství komponent, jejichž účelem je pokrýt řečové charakteristiky co největšího počtu různých mluvčích, aby byl systém co nejrobustnější. Po adaptaci na konkrétního mluvčího je ovšem počet Gaussových komponent systému pro daného mluvčího zbytečně vysoký [Hui00], neboť prostor příznaků každého mluvčího je omezený dle charakteristik jeho hlasu. Nabízí se proto možnost, zkusit omezit počet komponent adaptovaného systému a tím ještě více snížit jeho chybovost popřípadě paměťové či výpočetní nároky.

V rámci této disertační práce bylo pro redukci málo významných komponent navrženo kritérium založené na použití Baum-Welchova algoritmu. Pro každou komponentu každého stavu modelu je pomocí tohoto algoritmu na adaptačních datech vypočítána okupační věrohodnost, která vyjadřuje míru množství dat použitých pro adaptaci dané komponenty a zároveň i míru věrohodnosti, s jakou tato komponenta pokrývá charakteristiky hlasu daného mluvčího. Výpočet okupačních věrohodností všech komponent je nutné provést v rámci úlohy řízené adaptace u většiny adaptačních technik a tento krok proto sebou nenese žádné zvýšené výpočetní nároky. Porovnáním vypočtené hodnoty okupační věrohodnosti dané komponenty s průměrnou hodnotou vypočtenou přes všechny komponenty daného stavu

lze pak nezávisle na množství dat určit, zda je daná komponenta pro daného mluvčího málo významná a zda může být odstraněna.

Matematicky lze kritérium pro odstranění m-té komponenty stavu s daného modelu vyjádřit následovně:

$$\zeta_t(i,m) < \lambda\, \bar{\zeta}_t(i,m) \tag{6.1}$$

kde $\zeta_t(i,m)$ je okupační věrohodnost dané komponenty vypočtená dle vztahu 3.21, $\bar{\zeta}_t(i,m)$ je průměrná okupační věrohodnost stavu s vypočtená přes všechny jeho komponenty a λ je heuristicky určená konstanta. Uvedené kritérium bylo experimentálně ověřeno na testovací databázi z předchozí kapitoly.

V rámci prvního experimentu byly použity adaptované akustické modely obsahující až 100 komponent v každém stavu. Jejich skutečný počet závisel pro každý stav na množství dat dostupných během trénování. U těchto modelů, s celkem 14328 komponentami, byla poté provedena redukce málo významných komponent pro různé hodnoty λ až na úroveň systému, který byl natrénován s 64 komponentami na každý stav a obsahoval celkem 9295 komponent.

V tab. 6.7 jsou dosažené výsledky porovnány s adaptovanými modely o 100 komponentách na stav. Uváděná redukce chybovosti rozpoznávání je **relativní**.

hodnota λ	0,03	0,05	0,06	0,07	0,08	0,1	0,2	1,0
počet komponent	9260	9023	8920	8835	8667	8524	7822	5089
relativní redukce WER [%]	2,0	2,7	3,7	2,9	2,9	2,5	2,1	-8,8
redukce počtu komponent [%]	35,4	37,0	37,7	38,3	39,5	40,5	45,4	64,5
red. výpočetního času [%]	2,2	2,4	2,3	2,6	2,9	2,8	3,7	7,7

Tabulka 6.7: CSR - porovnání výsledků adaptovaného systému o 100 komponentách na stav před a po provedení redukce málo významných Gaussových komponent.

Z výsledků experimentu vyplývá, že chybovost systému s redukovaným počtem komponent je nižší oproti původnímu adaptovanému systému relativně jen cca o 2,8%. Dosažené zlepšení je tedy malé, i když statisticky signifikantní na hladině 5% (dle testu NIST MAPSSWE [NIST]). Největší přínos redukce málo významných komponent tak spočívá ve skutečnosti, že výsledné modely mají cca o 40 % méně komponent a proto je potřeba i o 40 % méně paměťové kapacity na jejich uchování. To může být přínosné například v systému pro přepis televizních a rozhlasových pořadů, který pracuje s modely několika set klíčových mluvčích. Kromě toho vede redukce počtu komponent také k malému snížení výpočetní náročnosti během rozpoznávání - cca o 3%.

V pořadí druhé tabulce (tab. 6.8) byly modely s redukovaným počtem komponent porovnány s modely, které byly natrénovány s 64 komponentami na stav

a poté standardně adaptovány. Celkový počet komponent obou typů modelů byl přitom téměř stejný.

hodnota λ	0,03	0,05	0,06	0,07	0,08	0,1	0,2	1,0
počet komponent	9260	9023	8920	8835	8667	8524	7822	5089
relativní redukce WER [%]	5,5	6,1	7,1	6,3	6,3	6,0	5,6	-4,9
redukce počtu komponent [%]	0,4	2,9	4,0	4,9	6,8	8,3	15,8	45,3
red. výpočetního času [%]	1,3	1,5	1,4	1,7	2,0	1,9	2,8	6,9

Tabulka 6.8: CSR - porovnání adaptovaného systému o 100 komponentách na stav po provedení redukce málo významných komponent s adaptovaným systémem o 64 komponentách na stav.

Z dosažených výsledků vyplývá zajímavý a přínosný závěr, že z hlediska úspěšnosti rozpoznávání je lepší vytvořit systém s větším počtem komponent, ten adaptovat a následně odstranit jeho málo významné složky, než adaptovat systém s odpovídajícím nižším počtem komponent.

6.4.3 Porovnání efektivity řízené a neřízené adaptace

Kromě klasické úlohy řízené adaptace, kde byly praktické aspekty jednotlivých metod zdokumentovány v předchozích experimentech, je cílem této kapitoly ověřit také možnosti adaptace neřízené, protože pouze neřízené metody lze použít pro adaptaci v komplexním systému pro přepis mluvených záznamů (viz. 7).

Cílem prvního experimentu je proto porovnat, jak se liší výsledky rozpoznávání po použití adaptovaného modelu, který byl vytvořen řízenou a neřízenou adaptací na stejných datech. Zatímco v prvním případě byl tedy fonetický přepis adaptačních dat vytvořen ručně člověkem, ve druhém byl přepis nahrávek vytvořen automaticky pomocí rozpoznávače řeči, přičemž během rozpoznávání byl použit akustický model nezávislý na mluvčím. Chybovost automaticky vytvořeného přepisu byla cca 20 %, což znamená, že přepis přibližně dvou slov z deseti byl zatížen chybou, která mohla vést k vytvoření méně přesného adaptovaného modelu.

Experiment (viz tab. 6.9) byl vyhodnocen za pomoci stejného rozpoznávacího systému a na stejné testovací databázi 1127 vět jako v kapitole 6.4.1. Pro adaptaci byla použita kombinace metody MAP a MLLR. Adaptovány byly střední hodnoty.

Výsledky experimentu byly překvapivě dobré v tom smyslu, že *neřízenou adaptací na stejných bylo dosaženo jen o málo horších výsledků než adaptací řízenou*. Pro menší množství dat, cca do 5 minut, byla chybovost rozpoznávání stejná. Až při použití většího množství dat se začala projevovat větší přesnost přepisu vytvořeného člověkem. Chybovost dosažená neřízenou adaptací se pak přestala snižovat a naopak se s rostoucím množstvím dat zvyšovala.

data [min]	0,5	1	2	4	5	7,5	10	12,5	15
řízená adapt.	18,0	17,3	16,9	16,4	16,1	15,6	15,4	15,2	15,1
neřízená adapt.	18,0	17,6	17,3	16,4	16,1	16,0	16,1	15,7	16,4

Tabulka 6.9: CSR - porovnání hodnot WER [%] po aplikaci řízené a neřízené adaptace při různé množství použitých adaptačních dat.

6.4.4 Kombinace řízené a neřízené adaptace

Dobré výsledky dosažené neřízenou adaptací v předchozím experimentu vedly k myšlence, zkusit zkombinovat oba dva přístupy a snížit tak dále chybovost rozpoznávání konkrétní neznámé promluvy (testovacích dat). V průběhu dalšího experimentu tak byl nejprve pro každého mluvčího vytvořen na 10 minutách ručně přepsaných adaptačních dat model odpovídající jeho hlasovým charakteristikám. Byla tedy aplikována klasická řízená adaptace. Následně byl vytvořený model použit pro přepis různého množství testovacích dat, konkrétně jedné věty až několika minut. Získaný přepis a původní adaptovaný akustický model byl pak použit pro vytvoření nového modelu, adaptovaného na konkrétní použité množství testovacích dat, který by měl dle teoretických předpokladů těmto datům více odpovídat svými parametry. Tento dvojfázově adaptovaný model byl posléze použit pro druhý rozpoznávací průchod, v rámci kterého byla měřena chybovost rozpoznávání uvedená v tab. 6.10.

testovací data [min]	jedna věta	0,5	1	2	5	10	všechna data daného mluvčího
WER [%]	17,7	17,4	17,2	17,1	16,8	16,6	16,4

Tabulka 6.10: CSR - chybovost systému při aplikaci řízené a následně neřízené adaptace.

Výsledky experimentu ukázaly, že n*avržený postup založený na kombinaci řízené a neřízené adaptace nevede bohužel ani v jednom případě k dodatečnému snížení chybovosti rozpoznávání*. Model vytvořený dvoufázovou adaptací z SI modelu, nejprve řízeně na adaptačních datech a pak ještě jednou neřízeně na testovacích datech, dával při rozpoznávání dokonce horší výsledky, než model vytvořený pouze jednofázovou neřízenou adaptací na adaptačních datech (viz předchozí experiment).

6.4.5 Adaptace na mluvčího a zvukový kanál

Cílem posledního experimentu provedeného v rámci úlohy rozpoznávání spojité řeči je ukázat, že metody adaptace na mluvčího lze efektivně použít nejen pro adaptaci na hlasové charakteristiky, ale také na prostředí, v kterém mluvčí řeč pronáší, či použité záznamové zařízení.

V rámci tohoto experimentu, jednalo se o jeden dílčí experiment z rozsáhlého

testování provedeného pro firmu Olympus, byl použit diktafon umístěný jak těsně před ústy mluvčího (konfigurace close-talk), tak i ve větší vzdálenosti, konkrétně 80 cm (konfigurace far-talk). Vybraný mluvčí pak v obou případech nadiktoval 100 vět pro testování a dalších 100 vět pro adaptaci. Tyto věty byly vytvořeny poloautomaticky podle obdobných pravidel jako adaptační sada slov v rámci kapitoly 6.3.1, přičemž důraz byl kladen na to, aby jednotlivé věty obsahovaly všechny fonémy v co nejbohatším kontextu a zároveň jejich čtení nečinilo potíže.

Při obou konfiguracích se akustické modely adaptovaly nejen na hlasové charakteristiky mluvčího, ale i na použité záznamové zařízení (využívající kompresi zvuku), přenosovou cestu ovlivněnou zejména ve druhém případě odrazy od okolních objektů, a šum prostředí, v kterém nahrávání probíhalo. Jednalo se o kancelář s několika běžícími počítači. Jako adaptační technika byla použita kombinace metod MAP a MLLR a adaptovány byly střední hodnoty.

Výsledky experimentu jsou uvedeny v tab. 6.11, kde je chybovost dosažená po použití adaptace na mluvčího porovnána s výsledky dosaženými pomocí Wienerovy filtrace (zvýrazňováním řeči).

konfigurace	původní systém	zvýrazňování řeči	adaptace na mluvčího	zvýrazňování řeči +adaptace na mluvčího
close-talk	18,9	18,5	9,7	9,5
far-talk	61,3	36,1	22,8	18,7

Tabulka 6.11: CSR - chybovost rozpoznávání [%] nahrávek z diktafonu před a po aplikaci metod adaptace a zvýrazňování řeči.

Ačkoli výsledky experimentu mají ze statistického hlediska jen malou vypovídací schopnost, protože experiment byl proveden pouze pro jednoho mluvčího a 100 vět, je z nich zřejmé, že metody adaptace na mluvčího lze skutečně využít i pro adaptaci na šum prostředí, použité záznamové zařízení a přenosovou cestu. Pomocí adaptace se při konfiguraci far-talk podařilo snížit chybovost rozpoznávání více než zvýrazňováním řeči - konkrétně z hladiny 61 % až na hladinu 23%.

KAPITOLA 7

NAVRŽENÁ METODA ADAPTACE NA MLUVČÍHO S NEZNÁMOU IDENTITOU

Výzkumné práce a experimenty provedené v předchozí kapitole ukázaly, že řízenou i neřízenou adaptací lze významně snížit chybovost rozpoznávání řeči u systémů, kde je během rozpoznávání řeči známa identita daného uživatele respektive mluvčího. Hlavním cílem této disertační práce proto bylo navrhnout metodu, která by umožnila provádět adaptaci také ve složitějších úlohách, kde není ve všech případech během zpracování řeči známa identita jednotlivých mluvčích a je prakticky nemožné ji určit automaticky. Jedná se například o úlohu automatického přepisu parlamentních debat, sportovních utkání či zpravodajských pořadů (BNT - Broadcast News Transcription).

7.1 Navržená metoda dvoufázové neřízené adaptace

Adaptační schéma navržené pro výše zmíněný účel (viz obr. 7.1 a článek [Cerva06]) vychází ze základního předpokladu, že z rozpoznávaného zvukového záznamu, například televizních zpráv, je možné automaticky vytvořit posloupnost kratších úseků (dále jen *segmentů*), které v ideálním případě obsahují pouze promluvu jednoho mluvčího. Uvedený předpoklad přitom není v praxi nesplnitelný - je možné ho s jistou mírou přesnosti zajistit například použitím metod automatické detekce změny řečníka [Zdansky05]. Cílem navržené metody je vytvořit pro každý segment daného záznamu co nejpřesnější akustický model, který by mohl být použit během procesu rozpoznávání řeči.

Stručně lze princip funkce metody popsat následujícím způsobem: ve fázi trénování systému je nejprve vytvořena množina modelů pro skupinu tzv. *referenčních* mluvčích, pro které musí být k dispozici akustická data se známým fonetickým přepisem. Proces vlastní adaptace na každý segment je pak z hlediska teorie adaptace neřízený a probíhá ve dvou fázích.

1. V rámci **první** fáze je nejprve na daném segmentu provedena automatická identifikace neznámého mluvčího a jeho pohlaví. Na základě získaných výsledků je pak z množiny všech referenčních mluvčích vybrána podskupina

N mluvčích, kteří jsou k daném mluvčímu akusticky nejblíže. Následně jsou jejich modely lineárně zkombinovány a výsledný model je použit během prvního rozpoznávacího průchodu pro vytvoření fonetického přepisu segmentu.

2. V průběhu **druhé** adaptační fáze je získaný přepis využit pro výpočet přesnějších váhových koeficientů metodou ML a lineární kombinace modelů referenčních mluvčích je provedena znovu. Výsledkem je finální adaptovaný model, který je následně aplikován během závěrečné fáze rozpoznávání řeči, kdy je vytvořen textový přepis segmentu.

Jednotlivé kroky uvedeného procesu jsou podrobně rozebrány v následujících podkapitolách.

7.1.1 Postup tvorby modelů referenčních mluvčích

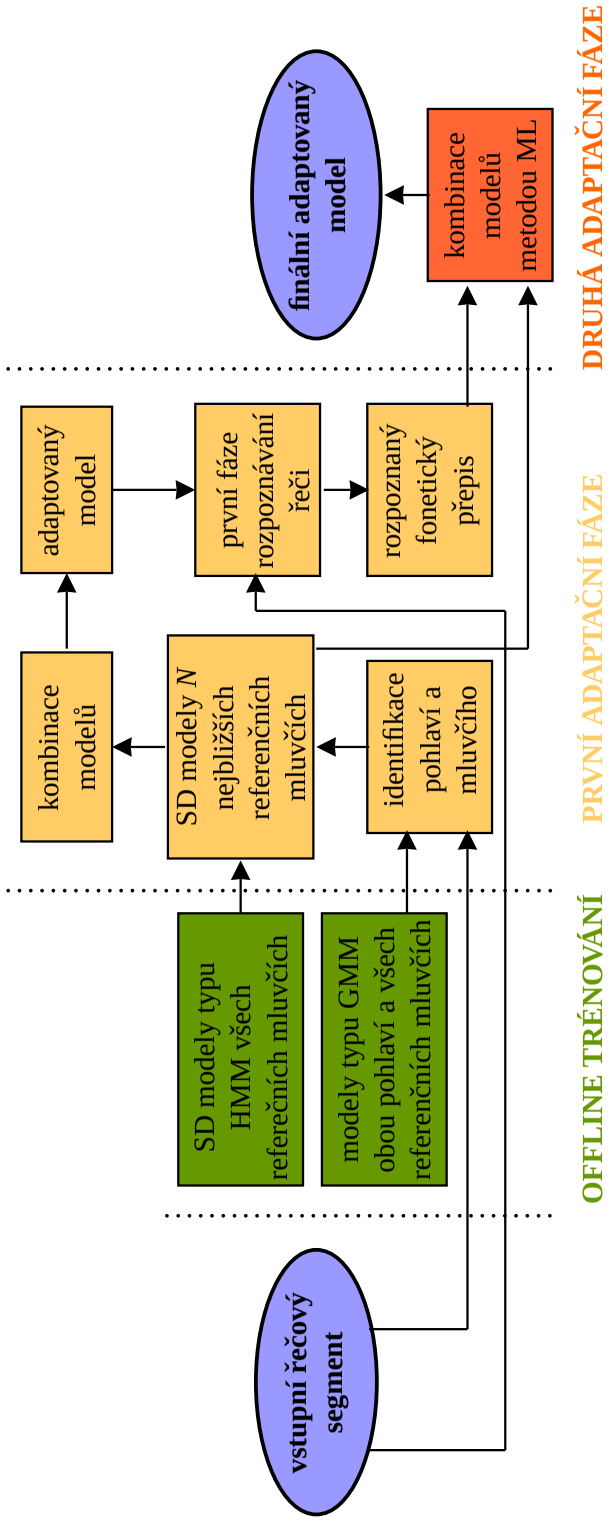
Akustické modely jsou vytvořeny pro všechny referenční mluvčí ve fázi trénování systému. Kromě modelů typu HMM jsou natrénovány i modely typu GMM sloužící pro automatickou identifikaci mluvčích a pohlaví. Zatímco GMM modely jsou natrénovány metodou ML, pro trénování Markovových modelů v našem případě nebyl, a ani obecně většinou nikdy není, k dispozici dostatek dat. Tyto modely jsou proto vytvořeny pomocí adaptace kombinací metod MAP a MLLR.

Při adaptaci jsou přitom jako apriorní brány parametry modelů závislých na pohlaví (GD) a adaptovány jsou vždy pouze střední hodnoty, protože pro adaptaci ostatních parametrů není pro všechny mluvčí k dispozici dostatek dat. Samotné GD modely jsou vytvořeny v několika iteracích standardního trénování a liší se pro obě pohlaví v počtu komponent, který závisí na množství dat použitých pro trénování ženského a mužského modelu.

Modely typu HMM dvou referenčních mluvčích stejného pohlaví se tak od sebe liší pouze v hodnotách vektorů středních hodnot. Ostatní parametry jejich modelů jsou stejné a Gaussovy komponenty jejich modelů si vzájemně odpovídají.

7.1.2 Identifikace mluvčího a výběr nejbližších mluvčích

Identifikace mluvčího je založena na použití připravených GMM modelů. Pro každého mluvčího je vypočítána věrohodnost, že jeho model vygeneroval daný segment. To samé je posléze provedeno i pro GMM modely reprezentující obě pohlaví, přičemž model s vyšší věrohodností určuje, zda je neznámý mluvčí muž či žena. Chybovost detekce pohlaví uvedeným způsobem přitom byla během všech provedených experimentů okolo 1 %, což znamená, že vliv chybovosti detekce pohlaví na celkové výsledky navržené metody je zanedbatelný.



Obrázek 7.1: BNT - schématické znázornění navržené dvoufázové neřízené adaptační metody.

Na základě výsledků všech modelů v procesu identifikace je pak vybrána skupina celkem N mluvčích, kteří mají k danému neznámému mluvčímu nejblíže (jejich modely dosáhly nejlepšího skóre). Jejich pohlaví přitom musí být stejné jako pohlaví, které bylo automaticky identifikováno. Tento požadavek přitom není jen přirozený, ale souvisí také s tím, že pouze modely mluvčích se stejným pohlavím jdou zkombinovat, neboť mají stejnou strukturu. Liší se pouze ve středních hodnotách.

7.1.3 První fáze kombinace modelů

V rámci první adaptační fáze není k dispozici žádná informace o fonetickém přepisu rozpoznávaného segmentu a adaptace je proto v principu velice obtížná. Nemůže být použita žádná metoda maximalizující věrohodnost vygenerování dat modelem a podobně. Navržený postup adaptace je proto založen na jednoduché ale robustní lineární kombinaci modelů referenčních mluvčích.

V literatuře lze pro účely lineární kombinace bez znalosti fonetického přepisu nalézt postup založený na použití okupačních věrohodností, které jsou spočítány pro všechny komponenty modelů všech referenčních mluvčích během trénování [Yoshizawa01]. Použití této metody je ale v našem případě nevhodné ze dvou důvodů.

První spočívá v tom, že hodnoty okupačních věrohodností závisí i na množství použitých dat. Protože v našem případě bylo pro jednotlivé mluvčí k dispozici velice rozdílné množství trénovacích nahrávek, od osmnácti sekund až do deseti minut, jsou hodnoty okupačních věrohodností stejných komponent různých mluvčích vzájemně neporovnatelné. Druhý problém je ten, že při tvorbě SD modelů mohla být díky nedostatku dat provedena pouze adaptace středních hodnot a ostatní parametry jsou tedy pro všechny mluvčí jednoho pohlaví stejné a nemá smysl je kombinovat.

Z těchto důvodů vychází navržená metoda lineární kombinace pouze z apriorní informace o podobnosti neznámého mluvčího k jednotlivým referenčním mluvčím a z informace o jeho pohlaví. Kombinovány jsou navíc pouze vektory středních hodnot - ostatní parametry jsou zkopírovány z odpovídajícího GD modelu. Tento způsob má dvě výhody: robustnost výsledného modelu je oproti adaptaci rozptylů jiným způsobem vysoká a modely s rozptyly z GD modelů přitom dávají při rozpoznávání signifikantně lepší výsledky než pouze SI modely (viz experiment v kap. 7.2.1).

Před samotnou kombinací středních hodnot jsou modely N nejbližších mluvčích seřazeny dle dosažené věrohodnosti ve vzestupném pořadí a následně je pro n-tý model spočítána pouze jedna globální váha dle vztahu

$$\lambda_n = \frac{n}{\sum_{j=1}^{N} j} \tag{7.1}$$

Uvedená rovnice 7.1 zajišťuje, že střední hodnoty nejbližšího mluvčího budou

mít ve výsledném modelu n-krát větší váhu než střední hodnoty mluvčího nejvzdálenějšího, dále že $\sum_{n=1}^{N} \lambda_n = 1$ a zároveň $\lambda_n > 0 \; \forall n$.

Vytvořený adaptovaný model je následně použit během prvního rozpoznávacího průchodu pro vytvoření fonetického přepisu rozpoznávaného segmentu.

7.1.4 Druhá fáze kombinace modelů

Ve druhé adaptační fázi je zužitkován vytvořený fonetický přepis pro opětovnou kombinaci středních hodnot. Ostatní parametry jsou opět zkopírovány z odpovídajícího GD modelu, protože jeden řečový segment neobsahuje dostatek dat pro jejich přesnou adaptaci.

Tentokrát ovšem není určována pouze jedna globální adaptační váha, ale všechny komponenty daného GD modelu jsou pomocí klastrování automaticky rozděleny do binárního regresního stromu. Stejně jsou rozděleny i všechny komponenty všech modelů jednotlivých referenčních mluvčích. Kombinace modelů je pak založena na metodě ML (viz kap. 4.4.1.2) a adaptační váhy jsou počítány pro různé úrovně regresního stromu na základě aktuálně dostupného množství dat.

Výsledkem druhé adaptační fáze je finální adaptovaný model, který je aplikován během druhého rozpoznávacího průchodu, kdy je cílem vytvořit výsledný textový přepis daného segmentu.

7.2 Hledání optimálních parametrů navržené metody

Hlavním volným parametrem popsané adaptační metody je bezpochyby počet referenčních mluvčích, jejichž modely se v obou adaptačních fázích mají zkombinovat.

Pro účely určení, v jakých mezích by se hodnota N měla pohybovat, bylo provedeno několik sad experimentů na vývojové řečové databázi. Ta obsahovala 2 hodiny dlouhý zvukový záznam parlamentní debaty nahraný z televize, který byl ručně (pro jednoduchost vyhodnocení) rozdělen do celkem 225 řečových segmentů tak, že každý segment obsahoval pouze promluvu jednoho mluvčího. Všechny segmenty dohromady obsahovaly celkem 13624 slov.

Pro rozpoznávání řeči byl použit stejný rozpoznávací systém jako v předchozí kapitole s tím, že jazykový model byl tentokrát vytvořen adaptací původního modelu na textovém korpusu záznamů z parlamentu [Nouza05-2] tak, aby výchozí rozpoznávací skóre bylo z hlediska jazykového modelování co nejpřesnější.

Rovněž akustický model byl natrénován na stejné řečové databázi jako v předchozí kapitole, takže během trénování nebyly použity žádné zvukové nahrávky parlamentních debat. Stejný byl i postup parametrizace řečového signálu.

Řečová databáze použitá pro trénování všech typů akustických modelů obsahovala nahrávky od více než 1000 různých mluvčích. Z nich bylo vybráno celkem 500 referenčních mluvčích (190 žen a 310 mužů), u kterých byla jednoznačně známa identita a pro které bylo k dispozici alespoň 18 sekund promluvy. Pro tyto

mluvčí pak byly pomocí adaptace vytvořeny SD, respektive přesněji SA, akustické modely.

7.2.1 První adaptační fáze

V rámci první adaptační fáze bylo experimentováno s několika různými přístupy pro kombinaci vektorů středních hodnot:

1. Metoda navržená v této práci - lineární kombinace dle vztahu 7.1, přičemž byla prováděna identifikace pohlaví a SD modely referenčních mluvčích byly vytvořeny adaptací z GD modelů.

2. Stejný postup jako v předchozím případě s tím, že byly kombinovány modely referenčních mluvčích s různým pohlavím, které byly vytvořeny adaptací z SI modelů.

3. Lineární kombinace dle článku [Yoshizawa01] na základě použití statistik uložených během tvorby SD modelů s tím, že opět nebylo bráno v úvahu pohlaví referenčních mluvčích.

Výsledky všech experimentů jsou uvedeny v tab. 7.1, která ukazuje chybovost rozpoznávání dosaženou zmíněnými postupy pro různé hodnoty N. Chybovost rozpoznávání s SI a GD modely přitom byla 26,8 % respektive 24,75 %.

hodnota N	5	25	50	75	100	150	190
navržená metoda	24,45	23,77	23,83	24,10	24,02	24,19	24,39
2. strategie	24,78	24,39	24,56	24,67	24,74	24,48	24,61
3. strategie	25,61	25,51	25,50	25,81	26,13	26,02	26,09

Tabulka 7.1: BNT - chybovost [%] přepisu parlamentních debat pro různé hodnoty N a použité metody kombinace modelů během první adaptační fáze.

Z tabulky 7.1 vyplývá, že největšího snížení chybovosti se podařilo dosáhnout pouhým použitím GD modelů - z 26,80 % na 24,75 %, tedy postupem, kdy nebyla prováděna identifikace mluvčího, ale bylo pouze detekováno jeho pohlaví. Kombinace vektorů středních hodnot měla za následek už jen menší, ale stále statisticky signifikantní (dle metodiky [NIST]) dodatečné snížení chybovosti, zejména pro hodnoty N okolo 25.

Ze srovnání první a druhé použité metody je dále zřejmé, že chybovost rozpoznávání může být snížena pouze v případech, kdy je skupina N referenčních mluvčích vybrána s ohledem na automaticky identifikované pohlaví a kdy jsou modely referenčních mluvčích vytvořeny adaptací z GD modelů.

Z porovnání prvního a třetího řádku je pak možné dojít k závěru, že navržená metoda dává signifikantně lepší výsledky než metoda založená na použití statistik

akumulovaných během procesu tvorby jednotlivých SD modelů. Výhodou navržené metody je navíc skutečnost, že není třeba žádné statistiky ukládat a že adaptace je díky použití jedné globální váhy rychlejší.

Kromě prezentovaných přístupů pak byly experimentálně odzkoušeny i další metody pro jednoduchou kombinaci vektorů středních hodnot, navržená metoda založená na identifikaci mluvčího, respektive podobnosti v akustickém prostoru, dávala ovšem vždy nejlepší výsledky.

7.2.2 Druhá adaptační fáze

Další sada provedených experimentů byla zaměřena na vyhodnocení druhé adaptační fáze. Cílem bylo porovnat použitou techniku kombinace metodou ML s neřízenou adaptací metodou MLLR. V případě MLLR byl model vytvořený v první fázi adaptace transformován za pomoci stejného binární regresního stromu, jako byl použit pro účely kombinace. Výsledky experimentu jsou dokumentovány v tab. 7.2.

hodnota N	5	25	50	75	100	150	190
metoda MLLR	23,81	23,10	23,03	23,26	23,16	23,29	23,33
kombinace založená na metodě ML	23,39	22,45	21,78	21,58	21,49	21,02	20,83

Tabulka 7.2: BNT - chybovost [%] přepisu parlamentních debat pro různé hodnoty N a použité metody adaptace během druhé adaptační fáze.

Výsledky dokazují, že kombinace vektorů středních hodnot referenčních mluvčích vede k lepším výsledkům než adaptace metodu MLLR, přičemž dosažené zlepšení se zvyšuje s rostoucí hodnotou N. Je to z toho důvodu, že v případě vyšší hodnoty N je vybrán větší počet referenčních mluvčích, pro které mohou být vypočítány optimální adaptační váhy. Výpočet vah metodou ML je navíc rychlejší než výpočet transformační matice u metody MLLR.

7.3 Experimentální vyhodnocení

7.3.1 Ručně segmentovaná data

Experimentální vyhodnocení navržené adaptační metody bylo nejprve provedeno na ručně segmentovaných nahrávkách. Jednalo se o databázi nahrávek zpravodajských pořadů pořízenou v rámci projektu COST278 [Vandecatseye04]. Konkrétně byly použity 2 hodiny nahrávek zpravodajství z Českého rozhlasu, které obsahovaly 16 677 slov a 3 hodiny nahrávek televizních zpráv ze stanic Nova, Prima a ČT1, které obsahovaly celkem 29 887 slov. Všechny nahrávky byly ručně rozděleny do několika stovek segmentů tak, aby obsahovaly pokud možno promluvu pouze jednoho mluvčího.

Tím, že databáze použitá pro tvorbu SD modelů referenčních mluvčích obsahovala také nahrávky z televize a rozhlasu, nastala u některých vytvořených testovacích segmentů situace, že daný mluvčí byl současně zastoupen také v množině referenčních mluvčích. V tomto případě by pak výsledný adaptovaný model vznikl kombinací i z SD modelu daného mluvčího, což by mohlo nadsazovat výsledky metody. Pro zajištění maximální objektivity provedeného experimentu byla proto v každé takové situaci daná osoba z databáze referenčních mluvčích dočasně vyjmuta.

V rámci experimentu byl parametr N nastaven na hodnoty, které se ukázaly jako nejlepší během předchozích experimentů na vývojové databázi. To znamená, že během první adaptační fáze bylo N rovno 25 a ve druhé 190. Výsledky experimentu jsou uvedeny v tab. 7.3. Vyplývá z nich, že *aplikací navržené metody lze snížit chybovost automatického přepisu oproti použití SI modelů relativně cca o 20 %*. Toto číslo lze považovat za velmi dobré z toho důvodu, že adaptace byla prováděna neřízeně pro každý segmen, přičemž délka jednotlivých segmentů se lišila v rozmezí několika jednotek až několika desítek sekund.

typ pořadu	SI modely	SA modely	relativní snížení chybovosti [%] oproti SI modelům
rozhlasové zprávy	19,45	15,03	22,7
televizní zprávy	22,96	19,04	17,0

Tabulka 7.3: BNT - chybovost přepisu různých pořadů [%] po aplikaci celé navržené dvoufázové adaptační metody.

Cenou za dosažené zlepšení je více než dvojnásobný výpočetní čas celého přepisu, který je dán dvoufázovým rozpoznáváním. V případě nutnosti rozpoznávat rychleji, například v online režimu kvůli titulkování, lze provádět pouze první adaptační fázi. Relativní dosažené zlepšení je pak zhruba poloviční a srovnatelné například s odlišnou metodu používanou v systému pro online indexaci vybraných satelitních pořadů [Liu05].

Významné zrychlení adaptace lze dosáhnout tím, že je v prvním rozpoznávacím průchodu použit menší slovník, obsahující řádově jen desítky tisíc slov. Právě na tento aspekt se zaměřuje následující podkapitola, kde je navíc navržená metoda otestována v reálném systému pro přepis mluvených zvukových záznamů, to jest bez ruční segmentace člověkem.

7.3.2 Reálný systém pro přepis zvukových nahrávek

Pro účely finálního ověření navržené adaptační metody byl použit systém ATT (Audio Transcription Toolkit) používaný v rámci Laboratoře počítačového zpracování řeči pro přepis různých zvukových záznamů, nejčastěji televizních a rozhlasových pořadů. Testovací databáze obsahovala 4 televizní zprávy, které byly nahrány

ze stanic ČT1, NOVA, PRIMA a ČT24. Jejich délka byla 90 minut a obsahovaly dohromady 13 759 slov. Množina referenčních mluvčích a řečová databáze použitá pro trénování GD modelů byla stejná jako v předchozí kapitole. Pouze jazykový model byl od doby provedení předchozích experimentů aktualizován.

Použitá nejnovější verze systému ATT neprovádí segmentaci rozpoznávaného zvukového záznamu, ale celý záznam je rozpoznáván jako celek s tím, že je v akustickém signálu prováděna detekce řeči, pohlaví mluvčího a změn řečníka. Při nalezení změny řečníka není tedy proces rozpoznávání přerušen, nedojde k segmentaci, ale pouze se aktualizuje akustický model. Rozpoznávání řeči je přitom založeno na Viterbiho časově synchronním dekodéru a běží oproti detektoru akustických změn s malým zpožděním. *Výhodou zmíněného přístupu je skutečnost, že umožňuje eliminovat chyby rozpoznávání vznikající v důsledku nepřesnosti automatické segmentace*. Bližší informace o popsané rozpoznávací strategii lze nalézt v článku [Zdansky07].

Chybovost modulu pro detekci řeči a pohlaví mluvčího byla na použité testovací databázi 0,78 % respektive 1,19 %. Detektor změny mluvčího byl založen na metodě BINSEG [Zdansky06]. Nerozpoznal přibližně 13 % změn mluvčího, které měly být detekovány, a v 17 % všech detekovaných změn naopak ve skutečnosti změna mluvčího nenastala.

Výsledky prvního provedeného experimentu jsou uvedeny v tab. 7.4. V rámci experimentu byla zvlášť vyhodnocena adaptace po první a druhé fázi rozpoznávání řeči. Kromě identifikace mluvčích byla navíc před první fází rozpoznávání řeči prováděna také jejich verifikace. V případě, že byl některý mluvčí během verifikace přijat, byl pro rozpoznávání řeči v prvním průchodu použit přímo jemu odpovídající SD model. V opačném případě, když byli všichni referenční mluvčí během verifikace zamítnuti, byl pro rozpoznávání použit GD model rozpoznaného pohlaví. Pro úplnost zbývá dodat, že EER (Equal Error Rate) během verifikace mluvčího byl 12,5 %.

	1. fáze adaptace			2. fáze adaptace	
adaptační metoda	GD modely	SD modely	kombinace modelů	MLLR	kombinace modelů pomocí metody ML
WER [%]	21,30	21,1	20,86	21,65	18,73
rel. snížení WER [%]	8,7	9,6	10,6	7,2	19,8

Tabulka 7.4: BNT - chybovost přepisu televizních zpráv po aplikaci navržené adaptační metody v reálném systému pro přepis zvukových záznamů (SI WER = 23,34 %).

Z výsledků experimentů vyplývá, že použitím verifikace mluvčího nelze dosáhnout signifikantně lepších výsledků než v případě, kdy je prováděna pouze detekce pohlaví. O něco lepší jsou výsledky v situaci, když že je namísto výběru jednoho konkrétního modelu prováděna lineární kombinace modelů nejbližších mluvčích.

Značně lepší výsledky pak přináší lineární kombinace založená na znalosti fo-

netického přepisu a metodě maximální věrohodnosti. *Dosažené relativní zlepšení 20 % je stejné, jako v případě manuálně segmentovaných dat*. Naopak aplikace metody MLLR dává v případě reálného neideálního systému horší výsledky, než jednofázová adaptace. *Vzhledem k chybovosti detekce změny mluvčího se* proto *kombinace modelů nejbližších mluvčích jeví jako robustnější než metoda MLLR.*

počet slov ve slovníku [tis.] během první fáze rozpoznávání	312	200	100	50	10
WER [%] po 1. fázi rozpoznávání	23,34	27,28	29,01	32,84	55,26
WER [%] po 2. fázi rozpoznávání	18,73	18,76	19,00	19,08	19,03

Tabulka 7.5: BNT - úspěšnost neřízené dvoufázové adaptace v závislosti na velikosti slovníku během první fáze rozpoznávání řeči.

Poslední provedený experiment (viz tab. 7.5) ukazuje, jak závisí úspěšnost navržené adaptační metody na velikosti slovníku použitého během první fáze rozpoznávání řeči. V druhé finální fázi rozpoznávání byl vždy použit největší dostupný slovník obsahující 312 tisíc slov.

Výsledky ukazují, že *z pohledu chybovosti rozpoznávání výrazně horší fonetický přepis vede pouze k zanedbatelně malému snížení celkové úspěšnosti dvoufázové adaptace. Rychlost celého přepisu je ovšem v případě použití menšího slovníku výrazně zvýšena* a největší nevýhodu navržené metody, více než dvojnásobnou výpočetní náročnost, lze tak uvedeným způsobem téměř eliminovat.

KAPITOLA 8

ZÁVĚR

V rámci této disertační práce se autor zabýval metodami umožňujícími provádět řízenou i neřízenou adaptaci akustického modelu na mluvčího s předem známou i neznámou identitou.

Úloha adaptace na mluvčího se známou identitou

V rámci úlohy adaptace na mluvčího, jehož identita je v době zpracování jeho promluvy známa, bylo cílem najít vhodný praktický přístup, jenž by umožňoval provádět řízenou adaptaci v již existujicích systémech vyvinutých na TUL, které jsou dlouhodobě používány jednou konkrétní osobou. Jedná se např. o systém pro hlasové ovládání PC či diktování. Z tohoto důvodu byla navržena sada speciálních adaptačních slov a byly provedeny srovnávací experimenty se známými a nejčastěji používanými technikami - metodou MAP a metodou MLLR. Ty byly navíc implementovány do podoby softwaru, jenž může být distribuován spolu s cílovým rozpoznávacím systémem. Všechny experimentálně dosažené výsledky lze shrnout následujícím způsobem:

V úloze rozpoznávání izolovaných slov lze při použití dostatečného počtu speciálně vybraných adaptačních slov aplikovat obě metody, případně jejich kombinaci, téměř se stejným výsledkem. Jako apriorní parametry je přitom výhodné použít hodnoty modelů natrénovaných jako na pohlaví závislých. Adaptaci je možné kromě vektorů středních hodnot provádět také pro rozptyly. Rozšíření na další parametry však již k lepším výsledkům nevede. Rozdíly v rámci jednotlivých metod lze pro různé hodnoty jejich parametrů nastavených v rozumném rozmezí zanedbat.

Konkrétně bylo použitím sady 300 adaptačních slov dosaženo zlepšení chybovosti diktovacích systémů z hladiny 14 % na hladinu 8,5 % (tedy relativně o 40 %). Počet slov ve slovníku klasifikátoru byl přitom 500 tisíc. Obdobných výsledků se podařilo dosáhnout také u motoricky handicapovaných osob s vadou řeči - u diktovacího systému došlo ke snížení chybovosti relativně o 40 % a v systému hlasového ovládání PC pak dokonce o 70 %. Počet použitých adaptačních slov musel být ale v obou případech dvojnásobný.

Jako velice přínosné se dále ukázalo použít stejnou techniku v systému vytvořeném mezijazykou adaptací, kdy kromě adaptace na hlasové charakteristiky konkrétního řečníka dochází zároveň také k adaptaci na odlišnou výslovnost jednotlivých fonémů daného jazyka.

V úĺoze rozpoznávání plynulé řeči se jako jednoznačně nejlepší ukázalo použití kombinace metod MAP a MLLR, které dávalo nejlepší výsledky pro libovolné množství adaptačních dat. I zde platí, že je výhodnější založit adaptaci na apriorních modelech vytvořených trénováním zvlášť na mužských a ženských datech. Optimální množství dat, jež by mělo být pro adaptaci použito, je přitom 10 minut.

V rámci provedených experimentů pak došlo při tomto množství dat ke snížení chybovosti rozpoznávače z hladiny 20 % na hladinu 15 %, tedy relativně o 25 %. Rozpoznávání přitom probíhalo se slovníkem o 312 tisících položkách. Dále bylo experimentálně ověřeno, že počet komponent adaptovaného sytému je pro jednoho konkrétního mluvčího v porovnání s SI modelem zbytečně velký a že málo významné komponenty lze efektivně odstranit pomocí jednoduchého kritéria.

Kromě toho byla také provedena sada experimentů zaměřených na porovnání účinnosti řízené a neřízené adaptace na stejných datech. Z dosažených výsledků vyplynulo, že při menším množství dat (několik minut záznamů) je možné dosáhnout neřízenou adaptací jen o málo horších výsledků než adaptací řízenou.

Úloha adaptace na mluvčího s neznámou identitou

V rámci úlohy adaptace na mluvčího, jehož identita není v době zpracování jeho promluvy známa, byla navržena vlastní dvoufázová neřízená adaptační technika, založená na principech metod CAT a SST. Jejím cílem je umožnit provádět adaptaci v komplexním systému pro přepis zvukových záznamů, zejména televizních a rozhlasových pořadů.

Jednotlivé fáze navrženého adaptačního schématu byly ověřeny v celé řadě rozsáhlých experimentů na různých typech pořadů. V rámci těchto experimentů přitom byla použita kromě ručně segmentovaných dat i data zpracovaná reálným přepisovacím systémem využívajícím modul automatické detekce změny řečníka.

Dosažené výsledky ukázaly, že pomocí navržené metody je možné snížit chybovost přepisu různých pořadů relativně o 20 % a že největší nevýhodu navržené metody - dvě fáze rozpoznávání řeči - lze částečně eliminovat tím, že je během prvního rozpoznávacího průchodu použit jen velmi malý slovník obsahující pouze deset tisíc slov. Navržená metoda navíc dávala lepší výsledky než neřízená adaptace pomocí MLLR a ukázala se oproti ní i robustnější vzhledem k chybovosti modulu detekce změny řečníka.

Shrnutí přínosů k rozvoji vědního oboru

V práci je

- podán jednotný výklad základních principů většiny používaných adaptačních metod, zejména přístupů založených na přímé adaptaci akustického modelu;
- formou spoluautorství vytvořeno a anotováno několik menších řečových databází pro účely testování adaptačních metod v různých úlohách - počínaje

hlasovým ovládáním PC a konče úlohou přepisu televizních a rozhlasových pořadů

- navržen praktický postup, jak provádět adaptaci na mluvčího v úloze rozpoznávání izolovaných slov a plynulé řeči za předpokladu, že je známa identita mluvčího, jehož promluva je rozpoznávána;
- popsán postup tvorby speciální sady adaptačních slov, jejíž pomocí lze dosáhnout lepších výsledků než aplikací běžného textu;
- navržena metoda pro redukci málo významných komponent adaptovaného systému a ověřena její účinnost;
- porovnána efektivita adaptace na mluvčího v řízeném a neřízeném režimu;
- experimentálně ověřena možnost použití metod adaptace na mluvčího v dalších netypických aplikacích, např. v systému vytvořeném mezijazykovou adaptací a pro účely adaptace na zvukový kanál;
- navržena nová metoda dvoufázové neřízené adaptace, která může být aplikována v případě, že není známa identita mluvčího, jehož promluva je zpracovávána;
- provedeno experimentální nalezení optimálních parametrů této metody na vývojové řečové databázi;
- experimentálně ověřena účinnost nové metody na rozsáhlé testovací databázi v reálném systému pro přepis zvukových televizních a rozhlasových pořadů.

Shrnutí přínosů pro praxi

Všechny vytvořené programy a metody navržené v rámci této disertační práce nachází své uplatnění v reálných systémech vyvíjených v Laboratoři počítačového zpracování řeči na TUL. Jedná se například o systém MyVoice pro hlasové ovládání počítače, který má v současné době již několik desítek uživatelů z řad českých motoricky handicapovaných osob, a kde je adaptace na mluvčího klíčová zejména pro osoby trpící vadou řeči. Umožňuje jim totiž používat zmíněný systém a tím i celý počítač obdobným způsobem, jako s ním pracují běžní uživatelé. Dále je možné adaptaci využít v systému MyDictate vyvinutém pro účely diktování po slovech a jeho modernější období vyvíjené v současné době pro diktát plynulý.

Kromě těchto již klasických a například pro angličtinu běžně dostupných systémů, lze implementované a navržené metody aplikovat i v komplexním systému pro přepis zvukových nahrávek, který je možné použít pro monitorování televizních nebo rozhlasových kanálů, přepis rozsáhlých zvukových archivů nebo rozpoznávání parlamentních debat či záznamů ze soudních síní.

Literatura

Citovaná literatura

[Anastakos96] Anastakos T. et al.: *A Compact Model for Speaker Adaptive Training.* In Proceedings of ICSLP96, Philadelphia, 1996.

[Boulianne06] Boulianne G., Beaumont J.-F., Boisvert M., Brousseau J., Cardinal P., Chapdelaine C., Comeau M., Ouellet P., Osterrath F.: *Computer-assisted closed-captioning of live TV broadcasts in French.* In Proceedings of InterSpeech2006, Pittsburgh (USA), 2006.

[Callejas07] Callejas Z., Nouza J., Cerva P. and López-Cózar R.: *MyVoice goes Spanish. Cross-lingual adaptation of a voice controlled PC tool for handicapped people.* XXIII Congreso de la Sociedad Espan ola para el Procesamiento del Lenguaje Natural (SEPLN07), 10 - 12 September 2007, Sevilla, Spain. In print.

[Cerva06] Cerva P. , Nouza J. and Silovsky J.: *Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination.* In Proceedings of InterSpeech2006, Pittsburgh (USA), pp. 2326-2329, 2006.

[Cerva07] Cerva P. , Nouza J.: *Design and Development of Voice Controlled Aids for Motor-Handicapped Persons.* In Proceedings of InterSpeech2007, Antwerp (Belgium), pp. 2521-2524, 2007.

[Dempster77] Dempster A.P., Laird N.M., Rubin D.B.: *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of Royal Statistical Society, vol. B 39, pp. 1-38, 1977.

[Diany05] Diany B., Nguyen L., Guo X., and Xu D.: *The BBN Mandarin Broadcast News Transcription System.* In Proceedings of InterSpeech2005, Lisboa (Portugal), 2005.

[Digalakis95] Digalakis V., Rtischev D., and NeumeyerL.G.: *Speaker Adaptation using Constrained Estimation of Gaussian Mixtures.* IEEE Trans. On Speech and Audio Processing, Vol.3, n.3, pp. 357-366, 1995.

[Gales96] Gales M.J.F., Woodland P.C.: *Mean and Variance Adaptation Within the MLLR Framework.* Computer Speech & Language, Vol. 10, pp. 249-264, 1996.

[Gales98] Gales M.J.F.: *Maximum likelihood linear transformations for HMM-based speech recognition.* Computer Speech and Language, vol. 12, no. 2, pp. 75-98, 1998.

[Gales98-1] Gales M.J.F.: *Cluster Adaptive Training for Speech Recognition.* IEEE Transactions on Speech and Audio Processing, Vol. 2, pp. 291-298, 1944.

[Gauvain04] Gauvain J.L., Lee C.H.: *Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains.* IEEE Trans. SAP, Vol. 2, pp. 291-298, 1994.

[Hajek96] Hajek D., Nouza J.: *Speaker Adaptation in HMM Based Speech Recognition.* In Proceedings of Radioelektronika96, pp.328-331, Brno, April 1996.

[Huang90] Huang X.D., Ariki Y., Jack M.A.: *Hidden Markov models for speech recognition.* Edinburgh University Press 1990.

[Huang01] Huang X.D., Acero A., Hon H.W.: *Spoken Language Processing.* Prentice Hall 2001.

[Huang02] Huang C., Chen T., Chang E.: *Adaptive Model Combination for Dynamic Speaker Selection Training.* In Proceedings of ICSLP2002, vol. 1, pp. 65-68, 2002.

[Hui00] Hui Y. Pascale F., Taiyi H.: *Principal Mixture Speaker Adaptation for Improved Continuous Speech Recognition.* In Proceedings of ICSLP2000, vol.1, pp. 774-777.

[Chesta99] Chesta C., Siohan O., Lee C.-H.: *Maximum a Posteriori Linear Regression for Hidden Markov Model Adaptation.* In Proceedings of EuroSpeech99, volume 1, pp. 211-214, 1999.

[Kuhn96] Kuhn R., Nguyen P., Junqua J.C., Goldwasser L., Niedzielski N., Fincke S., Field K., Contolini M.*Eigenvoices for Speaker Adaptation.* In Proceedings of InterSpeech1998, pp. 1771- 1774, Sydney (Australia), 1998.

[Leggetter95] Leggetter C. J., WOODLAND P. C. *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.*, Computer Speech & Language, Vol. 9, pp. 171-185, 1995.

[Lei00] Lei H., Jian W., Ditang F., Wenhu W. *Speaker Adaptation Based on Combination of MAP Estimation and Weighted neighbor regression.* In Proceedings of ICASSP2000, Volume II, pp. 981-984, Instanbul (Turkey), 2000.

[Liu05] Liu D., Kiecza D., Srivastava A., Kubala F.: *Online Speaker Adaptation and Tracking for Real-Time Speech Recognition.* In Proceedings of InterSpeech05, pp. 281-284, Lisbon (Portugal), 2005.

[MacQueen67] MacQueen J. B.*Some Methods for classification and Analysis of Multivariate Observations*. In Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.

[Matsoukas97] Matsoukas S. Schwartz R., Jin H., Nguyen L. *Practical Implementations of Speaker-Adaptive Training.* DARPA Speech Recognition Workshop, Chantilly VA, 1997.

[McTait05] McTait K., Adda-Decker M. *The 300K LIMSI German Broadcast News Transcription System.* In Proceedings of InterSpeech2005, Lisboa (Portugal), 2005.

[NGU05] Nguyen L., Xiang B., Afify M., Abdou S., Matsoukas S., Schwartz R., and Makhoul J.: *The BBN RT04 English Broadcast News Transcription System.* In Proceedings of InterSpeech2005, Lisboa (Portugal), 2005.

[NIST] *Webové stránky organizace NIST - National Institute of Standards and Technology.* Dostupné na WWW: <http:www.nistgovspeechtestssigtestmapsswe.htm>.

[Nouza97] Nouza J. (editor): *Počítačové zpracování řeči.* TUL Liberec 1997.

[Nouza97-1] Nouza J., Psutka J., Uhlíř J.: *Phonetic Alphabet for Speech Recognition of Czech.* Radioengineering, vol.6, no.4, pp. 16-20, 1997.

[Nouza05] Nouza J.: *Discrete and Fluent Voice Dictation in Czech Language.*Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 273-280, 2005.

[Nouza05-1] Nouza J., Nouza T., Červa P.: *A Multi-Functional Voice-Control Aid for Disabled Persons.* In Proceedings of Specom 2005, pp. 715-718, Patras (Greece).

[Nouza05-2] Nouza J., Červa P., Zdansky J., Kolorenc J., David P.: *Towards Automatic Transcription of Parliament Speech.* In Proceedings of Electronic Speech Signal Processing, pp. 237-244, Prague, Czech Republic, 2005.

[Nouza06] Nouza J., Zdansky J., Cerva P., Kolorenc J.: *Continual On-line Monitoring of Czech Spoken Broadcast Programs.* In Proceedings of Interspeech06, Pittsburgh (USA), 2006.

[Padmanabhan98] Padmanabhan M., Bahl L., Nahamoo D., Picheny M.: *Speaker Clustering and Transformation for Speaker Adaptation in Speech Recognition*

Systems. IEEE Transactions on Speech and Audio Processing, vol. 6, n1, pp. 71-77, 1998.

[Press02] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P.: *Numerical recipes in C.* Cambridge university press. 2002.

[Psutka06] Psutka J., Műller L., Matoušek J., Radová V.: *Mluvíme s počítačem česky*. Academia. 2006.

[Sakoe78] Sakoe H., Chiba S.: *Dynamic Programming Algorithm Optimization for Spoken Word Recognition.* IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 26, Issue 1, pp. 43-49, 1978.

[Shinoda97] Shinoda K., Lee C-H.: *Structural MAP Speaker Adaptation Using Hierarchical Priors.* In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 381- 387, 1997.

[Vandecatseye04] Vandecatseye A. et al.: *The COST278 pan-European Broadcast News Database.* In Proceedings of LREC04, Lisbon (Portugal), 2004.

[Viterbi67] Viterbi A.J.: *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Transactions on Information Theory 13(2), pp. 260–269, April 1967.

[Woodland99] Woodland P.C.: *Speaker Adaptation: Techniques and Challenges.* In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, Keystone, 1999.

[Yoshizawa01] Yoshizawa S., Baba A., Matsunami K. et al: *Evaluation on Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers*, In Proceedings of Eurospeech2001, vol. 2, pp.1219-1222, 2001.

[Young00] Young S., Kershaw D., Odell J., Woodland P., Ollason D., Valtchev V.: *The HTK Book.* Microsoft Corporation 2000.

[Zdansky05] Zdansky J.: *Metody detekce změny mluvčího v akustickém signálu.* Disertační práce, TU Liberec, 2005.

[Zdansky06] Zdansky J.: *BINSEG: An Efficient Speaker Based Segmentation Technique.* In Proceedings of InterSpeech06, Pittsburgh (USA), 2006.

[Zdansky07] Zdansky J., Cerva P., Silovsky J., Nouza J.: *Acoustic Model Management Strategies for Improved Automatic Transcription of Broadcast Programs.* In Proceedings of Specom 2007, Moscow, Russia. In print.

[Zhan97] Zhan P., Westphal M.: *Speaker Normalization Based on Freguency Warping.* In Proceedings of ICASSP97, Munich (Germany), 1997.

[Železný01] Železný M.*Adaptace systému rozpoznávání plynulé češtiny na konkrétního řečníka,* Disertační práce, ZČU Plzeň 2001.

Seznam vlastních prací

[1] Cerva P. , Nouza J.: *Design and Development of Voice Controlled Aids for Motor-Handicapped Persons*, In Proceedings of InterSpeech2007, Antwerp (Belgium), pp. 2521-2524, 2007.

[2] Callejas Z., Nouza J., Cerva P. and López-Cózar R.: *MyVoice goes Spanish. Cross-lingual adaptation of a voice controlled PC tool for handicapped people* XXIII Congreso de la Sociedad Espan ola para el Procesamiento del Lenguaje Natural (SEPLN07), 10 - 12 September 2007, Sevilla, Spain. In print.

[3] Zdansky J., Cerva P., Silovsky J., Nouza J.: *Acoustic Model Management Strategies for Improved Automatic Transcription of Broadcast Programs.* In Proceedings of Specom 2007, Moscow, Russia. In print.

[4] Cerva P. , Nouza J. and Silovsky J.: *Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination*, In Proceedings of InterSpeech2006, Pittsburgh (USA), pp. 2326-2329, 2006.

[5] Nouza J., Zdansky J. Cerva P., Kolorenc J.: *Continual On-line Monitoring of Czech Spoken Broadcast Programs*, In Proceedings of Interspeech2006, pp. 1650-1653, Pittsburgh (USA), 2006.

[6] Nouza J., Zdansky J., Cerva P., Kolorenc J.: *A System for Information Retrieval from Large Records of Czech Spoken Data*, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE, pp. 485-492, Springer Berlin, 2006.

[7] Cerva P., Nouza J., Kolorenc J.: *Improved Transcription of Czech Parliament Speeches by Acoustic and Language Model Adaptation*, In Proceedings of Specom2006, pp. 103-106, St. Petersburg (Russia), 2006.

[8] Kolorenc J., Nouza J., Cerva P.: *Multi-words in the TV/radio News Transcription System*, In Proceedings of Specom2006, St. Petersburg (Russia), 2006.

[9] Boril H., Cerva P., Zdansky J., Kolorenc J.: *Lombard Speech Recognition: A Comparative Study*, In Proceedings of 16th Czech-German Workshop „Speech Processing“, Prague (Czech Republic), 2006.

[10] Silovsky J., Cerva P.: *Study on Speaker Recognition Aided Broadcast Streams Transcription*, In Proceedings of 16th Czech-German Workshop „Speech Processing“, Prague (Czech Republic), 2006.

[11] Nouza J., Zdansky J., David P., Cerva P., Kolorenc J., Nejedlova D.: *Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon*, In Proceedings of Interspeech2005, pp. 1681-1684, Lisbon (Portugal), 2005.

[12] Cerva P., Nouza J.: *Supervised and Unsupervised Speaker Adaptation in Large Vocabulary Continuous Speech Recognition of Czech*, LECTURE NOTES IN ARTIFICIAL INTELLIGENCE 3658, pp. 203-210, Springer Berlin, 2005.

[13] Cerva P., David P., Nouza J.: *Acoustic Modeling Based on Speaker Recognition and Adaptation for Improved Transcription of Broadcast Programs*, In Proceedings of Specom2005, pp. 183-186, Patras (Greece), 2005.

[14] Nouza J., Nouza T., Červa P.: *A Multi-Functional Voice-Control Aid for Disabled Persons.* In Proceedings of Specom 2005, pp. 715-718, Patras (Greece).

[15] Cerva P.: *Reduction of Unimportant Gaussian Components in Speaker Adapted Continuous Speech Recognition Systems*, In Proceedings of 7th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS), Tolouse (France), 2005.

[16] David P., Cerva P., Nouza J.: *Optimized Configuration of Speaker Recognition System for Broadcast News Transcription*, In Proceedings of 7th International Workshop on Electronics, Control, Modelling, Measurement and Signals (ECMS), Tolouse (France), 2005.

[17] Cerva P.: *Study on Different Speaker Adaptation Approaches in Isolated-Word Speech Recognition of Czech.* In Proceedings of 14th Czech-German Workshop „Speech Processing“, pp. 61-65, Prague (Czech Republic), 2004.

[18] David P., Cerva P., Nouza J.: *Speaker Recognition Applied for Enhanced Broadcast News Transcription.* In Proceedings of 14th Czech-German Workshop „Speech Processing“, pp. 72-76, Prague (Czech Republic), 2004.

[19] Nouza J., Cerva P., Zdansky J., Kolorenc J.: *Towards automatic transcription of parliament speech*, In Proceedings of Electronic Speech Signal Processing 2005, pp. 237-244, Prague (Czech Republic), 2005.

[20] Holada M., Nouza J., Cerva P., Nouza T.: *Distributed Recognition Used as Platform for Public Testing of Speech Technology Applications*, In Proceedings of n ASIDE2005 ISCA ITRW and COST278 Final Workshop on Applied Spoken Language Interaction in Distributed Environments, Aalborg (Danmark), 2005.

[21] Cerva P., Nouza J.: *MAP Based Speaker Adaptation in Very Large Vocabulary Speech Recognition of Czech*, RadioeEngineering, pp. 42-46, Vol. 13, No 3, September 2004.

[22] Cerva P, Nouza J.: *Map Based Speaker Adaptation in Large Vocabulary Speech Recognition of Czech Language*, In Proceedings of Radioelektronika 2004, Bratislava (Slovak Republic), 2004.

[23] Cerva P., Skoda J., Nouza J.: *Building and Annotating Large Speech Databases for Automatic Speech Recognition*, In Proceedings of Radioelektronika 2004, Bratislava (Slovak Republic), 2004.

DODATEK A

TABULKY

Úloha rozpoznávání izolovaných slov

počet uzlů regresního stromu	0	8	32	64	128	192
binární regresní strom: střední hodnoty						
WER [%]	11,6	10,4	9,7	9,5	9,6	9,6
binární regresní strom: střední hodnoty a rozptyly						
WER [%]	10,9	9,9	9,4	9,2	9,0	8,9
binární regresní strom s expertní inicializací: střední hodnoty						
WER [%]	12,1	10,2	9,9	9,6	9,6	9,7
binární regresní strom s expertní inicializací: střední hodnoty a rozptyly						
WER [%]	11,4	10,3	9,6	10,0	10,0	9,9

Tabulka A.1: IWSR - výsledky adaptace různých parametrů metodou MLLR při použití několika typů regresních stromů (SI WER = 14,0 %).

počet adapt. slov	0 = GD model	50	150	300	500	750	1000	1300
mluvčí se standardní výslovností: diktování								
WER [%]	10,8	10,4	9,3	8,5	7,6	7,3	7,2	7,1
mluvčí s vadou řeči: diktování								
WER [%]	19,0	18,5	16,9	15,7	14,5	12,4	11,2	11,2
mluvčí s vadou řeči: hlasové ovládání								
WER [%]	17,0	9,9	8,0	6,8	5,7	4,8	4,7	4,6

Tabulka A.2: IWSR - porovnání úspěšnosti adaptace na mluvčího se standardní výslovností a handicapované osoby s vadou řeči.

hodnota τ	2	4	6	8	10	12	14	16	18	20	25	50	100	500
střední hodnoty														
WER [%]	8,9	9,1	9,1	9,2	9,3	9,5	9,6	9,8	10,0	10,0	10,3	10,9	11,8	13,1
střední hodnoty+rozptyly														
WER [%]	8,8	8,7	8,9	9,4	9,3	9,4	9,6	9,7	10,0	10,0	10,4	10,9	11,7	13,1
střední hodnoty+rozptyly+váhové koeficienty jednotlivých komponent														
WER [%]	10,9	11,5	11,8	11,9	12,7	12,8	13,1	13,3	13,6	13,8	14,2	15,0	16,1	13,1

Tabulka A.3: IWSR - výsledky adaptace různých parametrů modelů metodou MAP při odlišných hodnotách adaptačního váhového koeficientu τ (SI WER = 14,0 %).

adaptační data [min]	0,5	1	2	3	4	5	7,5	10	12,5	15
metoda MAP										
WER [%]	20,1	19,4	18,8	18,0	17,5	17,2	16,9	16,3	16,0	16,1
metoda MLLR										
WER [%]	18,0	17,6	17,1	16,9	16,8	16,4	16,0	15,7	15,9	15,7
kombinace metody MAP a MLLR										
WER [%]	18,0	17,3	16,9	16,6	16,4	16,1	15,6	15,4	15,2	15,1

Tabulka A.4: CSR - porovnání výsledků adaptace různými metodami pro různé množství použitých adaptačních dat (SI WER = 19,9 %).