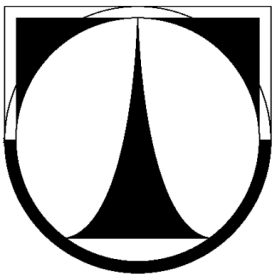
TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových
inženýrských studií



ŘÍZENÁ A NEŘÍZENÁ ADAPTACE NA MLUVČÍHO V SYSTÉMECH ROZPOZNÁVÁNÍ ŘEČI

AUTOREFERÁT DISERTAČNÍ PRÁCE

2007 PETR ČERVA

ŘÍZENÁ A NEŘÍZENÁ ADAPTACE NA MLUVČÍHO V SYSTÉMECH ROZPOZNÁVÁNÍ ŘEČI

AUTOREFERÁT 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

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

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.

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

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 3 disertační práce 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 obsažen v disertační práci v kapitole 3.

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

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.

KAPITOLA 4

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.

4.1 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.

4.2 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.

4.3 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.

4.3.1 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 4.3.5 a 4.4.1.

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.

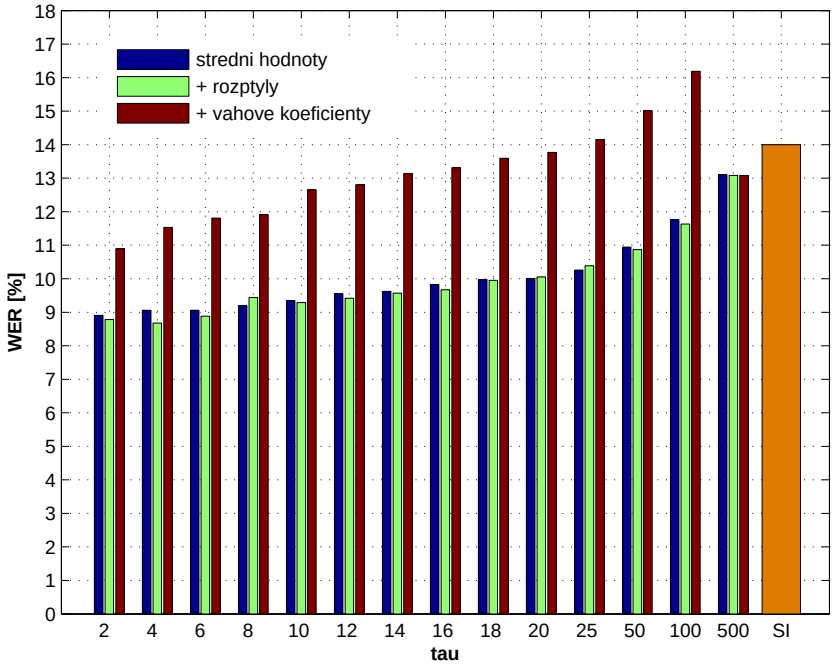
Experimentální výsledky dosažené použitím vytvořené sady slov a různých adaptačních technik jsou obsahem následujících kapitol 4.3.2 až 4.3.4. Cílem přitom bylo otestovat jednotlivé metody pro různé hodnoty jejich parametrů a ověřit, na jakou *hladinu* lze jejich aplikací snížit chybovost rozpoznávání. 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 4.3.2 až 4.3.4 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 %.

4.3.2 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 4.1: 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. 4.1. 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.

4.3.3 Adaptace metodou MLLR

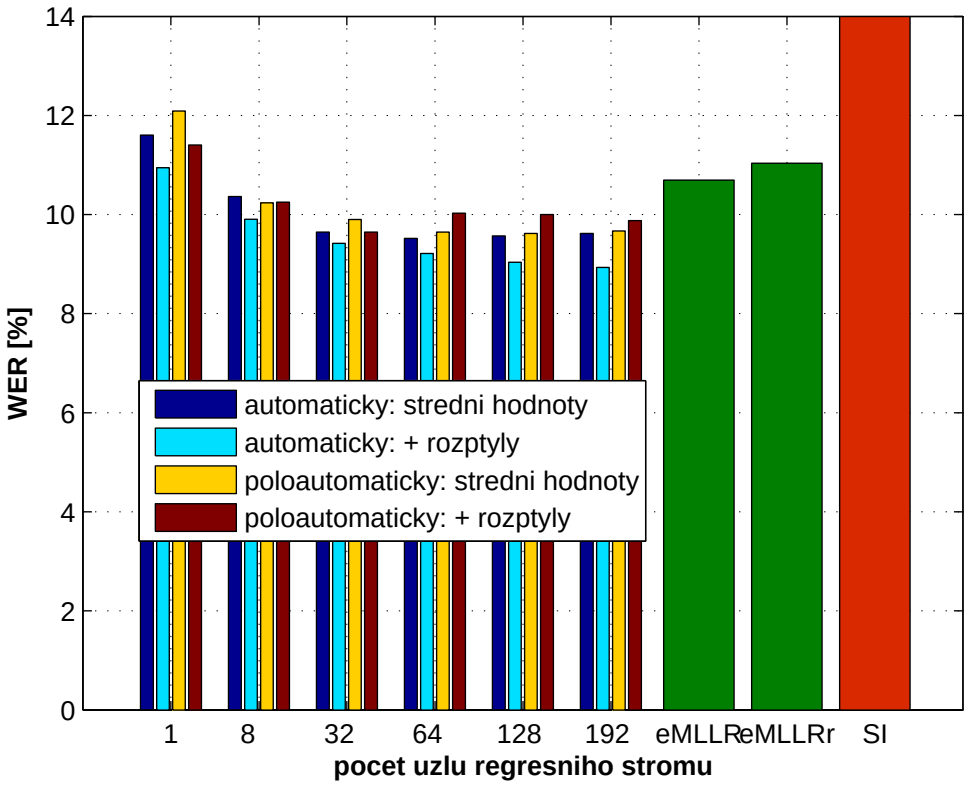
V pořadí druhý experiment byl zaměřen na adaptaci metodou MLLR. Adaptace byla prováděna s použitím regresního stromu vytvořeného třemi různými způsoby tak, jak to umožňuje vytvořený adaptační software:

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. 4.1.

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 4.1: 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 4.2: 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. 4.2 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.

4.3.4 Vliv použité sady adaptačních slov

Ve všech předchozích experimentech byla pro adaptaci použita sada 300 speciálně vybraných adaptačních slov (viz kap. 4.3.1). 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. 4.2.

	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

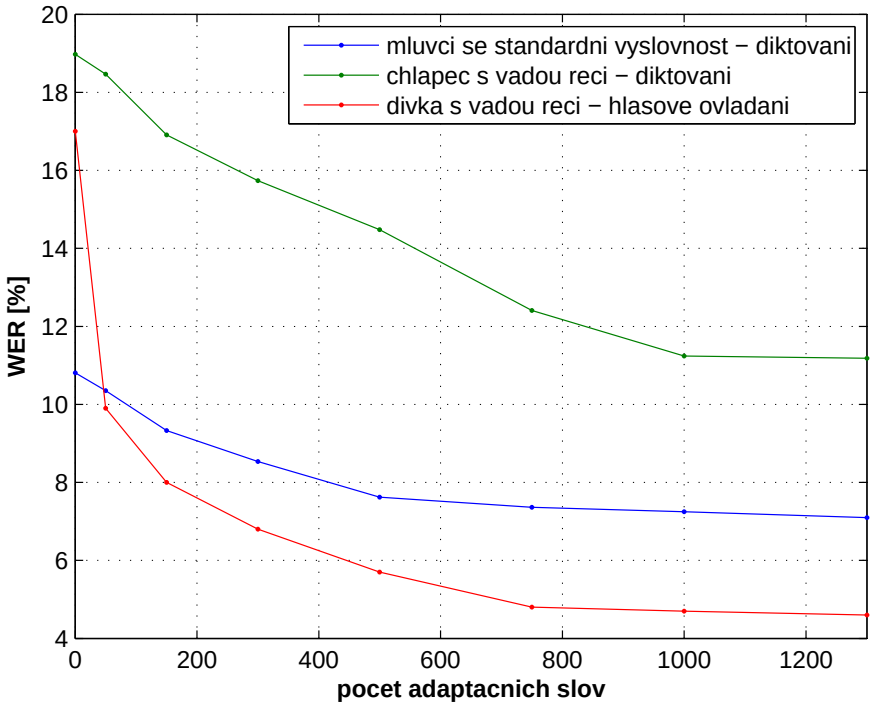
Tabulka 4.2: IWSR - porovnání hodnot WER [%] po adaptaci založené na použití běžného textu a speciálně připravené sady adaptačních slov.

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.

4.3.5 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. 4.3 proto ukazuje, jakých vý-

sledků lze pomocí adaptace dosáhnout právě u osob s motorickým handicapem doprovázeným mírnou vadou řeči.



Obrázek 4.3: 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 cha-

rakteru 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.

4.4 Úloha rozpoznávání plynulé řeči

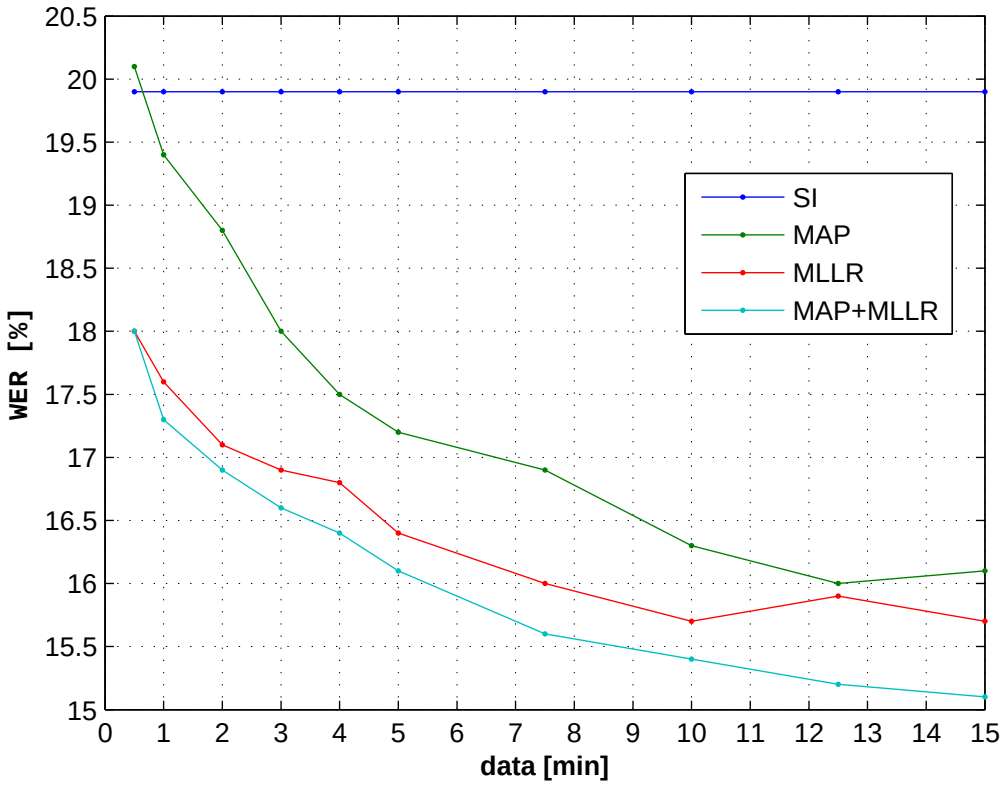
Po úloze rozpoznávání izolovaných slov byla další fáze experimentálního vývoje a vyhodnocování zaměřena na složitější úlohu rozpoznávání plynulé řeči. Cílem bylo najít sadu nejvhodnějších technik pro adaptaci na mluvčího v systému pro plynulé rozpoznávání řeči (Continuous Speech Recognition - CSR), který 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.

4.4.1 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. 5).

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 4.4: 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. 4.4) 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.

4.4.2 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. 5).

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. 4.3) byl vyhodnocen za pomoci stejného rozpoznávacího systému a na stejné testovací databázi 1127 vět jako v kapitole 4.4.1. Pro adaptaci byla použita kombinace metody MAP a MLLR. Adaptovány byly střední hodnoty.

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

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.

KAPITOLA 5

NAVRŽENÁ METODA ADAPTACE NA MLUVČÍHO S NEZNÁMOU IDENTITOU

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).

5.1 Navržená metoda dvoufázové neřízené adaptace

Adaptační schéma navržené pro výše zmíněný účel (viz obr. 5.1 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.

5.1.1 Postup tvorby modelů referenčních mluvčích

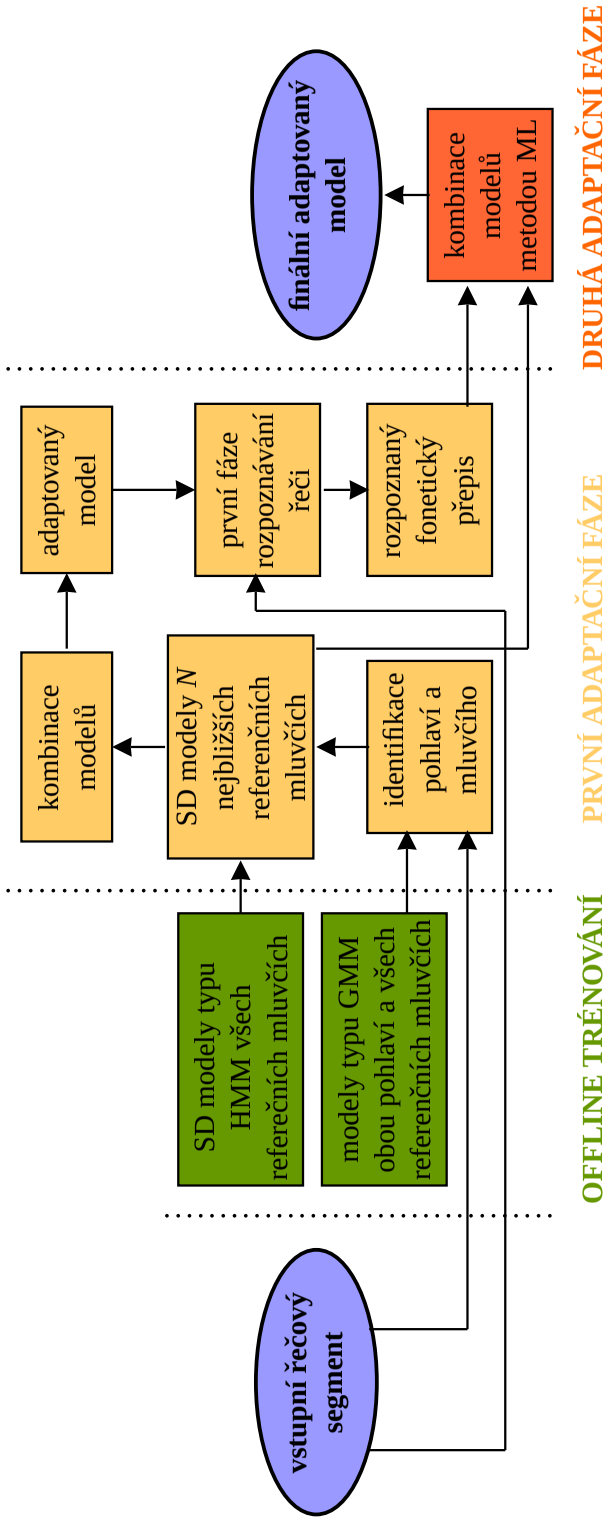
Akustické modely jsou vytvořeny pro všechny referenční mluvčí ve fázi trénování systému. Kromě modelů typu HMM jsou natrénovány i modely typu GMM sloužící pro automatickou identifikaci mluvčích a pohlaví. Zatímco GMM modely jsou natrénovány metodou ML, pro trénování Markovových modelů v našem případě 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.

5.1.2 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 5.1: 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.

5.1.3 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, která vychází 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.

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

$$\lambda_n = \frac{n}{\sum_{j=1}^{N} j} \tag{5.1}$$

Uvedená rovnice 5.1 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.

5.1.4 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 maximální věrohodnosti 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.

5.2 Experimentální vyhodnocení

5.2.1 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.

V rámci experimentu byl parametr N nastaven na hodnoty, které se ukázaly jako nejlepší během předchozích experimentů provedených na vývojové databázi (nejsou součástí autoreferátu). To znamená, že během první adaptační fáze bylo N rovno 25 a ve druhé 190.

Výsledky experimentu jsou uvedeny v tab. 5.1. 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 5.1: 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.

5.2.2 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. 5.2. 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 %.

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.

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

Značně lepší výsledky pak přináší lineární kombinace založená na znalosti fonetického 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 5.3: BNT - úspěšnost neřízené dvoufázové adaptace v závislosti na velikosti slovníku během první fáze rozpoznávání řeči.

Poslední provedený experiment (viz tab. 5.3) 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 6

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

[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.

[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.

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

[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.

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

[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.

[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.

[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>.

[Nouza97-1] 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.

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

[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.

Ing. Petr Červa

Řízená a neřízená adaptace na mluvčího v systémech rozpoznávání řeči

Autoreferát disertační práce

Technická univerzita v Liberci
Fakulta mechatroniky a mezioborových inženýrských studií

Náklad 20 výtisků

září 2007