Fakulta aplikovaných věd

Katedra kybernetiky

ODBORNÁ PRÁCE KE STÁTNÍ DOKTORSKÉ ZKOUŠCE

Ing. Zbyněk Zajíc

Automatická adaptace akustického modelu

školitel: Doc. Dr. Ing. Vlasta Radová

Plzeň, 2008

Poděkování

V prvé řadě bych rád poděkoval svému vedoucímu práce, Doc. Vlastě Radové a dále pak Doc. Luďkovy Müllerovi za připomínky a návrhy k zpracování této práce. Velké poděkování patří celé mé rodině, včetně přítelkyně, kteří mě v mém úsilí plně podporovali.

Obsah

1 Úvod **1**

1.1 Současný stav a struktura práce 2

2 Akustické modelování **3**

2.1 Struktura akustického modelu 3

2.2 Výpočet pravděpodobnosti promluvy 5

2.2.1 Rekurzivní výpočet forward-backward algoritmem 5

2.2.2 Iterativní Viterbiho algoritmus 6

2.3 Trénování parametrů akustického modelu 6

2.3.1 Metoda maximální věrohodnosti (ML) 6

2.3.2 Metoda maximální aposteriorní pravděpodobnosti (MAP) 8

2.3.3 Diskriminativní trénování (DT) 9

3 Metody adaptace **11**

3.1 Obecné dělení adaptačních metod 12

3.2 Metoda maximální aposteriorní pravděpodobnosti (MAP) 13

3.2.1 Diskriminativní MAP (DMAP) 14

3.3 Metody adaptace založené na lineární transformaci 15

3.3.1 Metoda maximální věrohodné lineární regrese (MLLR) 15

3.3.2 Metoda MLLR pro transformace vektorů pozorování (fMLLR) 18

3.3.3 Diskriminativní lineární transformace (DLT) 19

3.3.4 Shlukování podobných parametrů modelu 20

3.4 Kombinace přístupu MAP a MLLR 23

3.4.1 Dvoukroková adaptace 23

3.4.2 Regresní predikce modelu (RMP) 23

3.4.3 Regrese vážených sousedů (WMR) 24

3.4.4 Strukturální MAP 24

3.5 Shlukování mluvčích (SC) 25

3.6 Dekompozice vlastních hlasů (ED) 26

4 Adaptační techniky pro trénování **27**
4.1 Trénování s adaptací na mluvčího (SAT) 28
4.1.1 SAT pro MLLR 28
4.1.2 SAT pro fMLLR 29
4.1.3 Diskriminativní adaptace pro trénování (DAT) 30
4.2 Trénování s adaptací pomocí shlukování mluvčích (CAT) 31
4.2.1 Hledání parametrů modelu a transformací 31
4.2.2 Reprezentace shluků 32
4.2.3 Diskriminativní adaptace pro trénování pomocí shlukování (DCAT) . 32
4.3 Normalizace délky hlasového traktu (VTLN) 32
4.3.1 Transformační funkce 33
4.3.2 Odhad warpovacího faktoru 34
4.3.3 Normalizovaný akustický model 34

5 Experimenty **35**
5.1 Data a akustický model 35
5.2 Hodnocení úspěšnosti rozpoznávání 35
5.3 Výsledky 36
5.3.1 Klasické metody adaptace 36
5.3.2 Kombinace adaptačních metod 37
5.3.3 Adaptační trénování 37
5.3.4 Množství dat pro adaptaci 37
5.4 Zhodnocení experimentů 38

6 Závěr **39**
6.1 Dílčí cíle disertační práce 39

Seznam tabulek

5.1 ACC[%] vybraných adaptačních metod. 36
5.2 ACC[%] kombinace adaptačních metod. 37
5.3 Porovnání ACC[%] dané SI a SAT modelem po adaptaci. 37
5.4 ACC[%] při různém počtu adaptačních vět. 38

Seznam obrázků

2.1 Příklad třístavového skrytého Markovova modelu pro trifóny 4

3.1 Schématické znázornění adaptace. 11
3.2 Ilustrativní příklad adaptace složek modelu SI ve směru adaptačních dat. . . 12
3.3 Příklad binárního regresního stromu. 21
3.4 Příklad fonetického stromu. 22
3.5 Blokový diagram WNR adapatace převzatý z [HWFW00]. 24

4.1 Ilustrativní příklad rozdílné variability složek modelu SI a kanonického modelu. 28
4.2 Metoda SAT založená na MLLR transformacích 29
4.3 Metoda SAT založená na fMLLR transformacích 30
4.4 Warpovací funkce a) po částech lineární, b) bilineární. 33

Seznam zkratek

ACC	Accuracy
ASR	Automatic Speech Recognition
BIC	Bayes Information Criterion
CAT	Cluster Adaptive Training
CM	Certainty Measure
CMLLR	Constrained Maximum Linear Regression
CMN	Cepstrum Mean Normalization
CORR	Correctness
DAT	Discriminative Adaptation Training
DCAT	Discriminative Cluster Adaptive Training
DLT	Discriminative Linear Transformation
DMAP	Discriminative Maximum A-Posteriori
DT	Discriminative Training
ED	Eigenvoices Decomposition
EM	Expectation-Maximization
fMLLR	feature Maximum Likelihood Linear Regression
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
LD	Linear Discriminant
LSE	Least Square Error
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori
MCE	Minimum Classification Error
MFCC	Mel Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MLLRcov	Maximum Likelihood Linear Regression for covariance
MLLRmean	Maximum Likelihood Linear Regression for mean
MMI	Maximum Mutual Information
MMI-FD	Maximum Mutual Information Frame Discrimination
MPE	Minimum Phone Error
MWE	Minimum Word Error
OOV	Out Of Vocabulary

PCA	Principal Component Analysis
PLP	Perceptual Linear Predictive
RMP	Regression-based Model Prediction
SA	Speaker Adaptive
SAT	Speaker Adaptive Training
SC	Speaker Clustering
SD	Speaker Dependent
SI	Speaker Independent
SMAP	Structural Maximum A Posteriori
SMAPLR	Structural Maximum A Posteriori Linear Regression
VTLN	Vocal Tract Length Normalization
WER	Word Error Rate
WMR	Weighted Neighbor Regression
WSMAP	Weighted Structural Maximum A Posteriori

Anotace

Tato práce se zabývá problematikou automatické adaptace akustického modelu na aktuální testovací data od konkrétního řečníka. Pro natrénování modelu je potřeba velkého množství dat, které je z praktického hlediska nemožné získat od jednoho řečníka. Řešením je konstrukce akustického modelu na datech od více řečníků, vzniká pak na řečníku nezávislý model, který je schopný rozpoznat data od kteréhokoliv uživatele.

Pokud je však cílový řečník znám, lze snížit chybovost rozpoznávání užitím modelu natrénovaného na jeho datech. Obvykle je však v praxi nemožné získat dostatečné množství potřebných dat, proto byly navrženy adaptační metody, které mají za úkol normalizovat testovací data nebo posunout parametry akustického modelu směrem k testovacím datům. Zároveň s řečníkem jsou adaptovány i akustické podmínky při nahrávaní, jako jsou typické ruchy prostředí, použité nahrávací zařízení atd.

Práce si klade za cíl vysvětlit principy používaných adaptačních metod a postupy adaptačního trénování. Provedené experimenty dokazují účinnost vybraných metod a jejich možné kombinace.

Anotation

This work is concerned with the automatic speaker adaptation of an acoustic model in the automatic speech recognition system. For the model training, it is necessary to have large amount of data from many speakers. The final model, speaker independent, is then able to recognize speech from any speaker.

When speaker's identity is known, we could lower the error rate by using a model trained on the data from the particular speaker. Such a model is called speaker dependent model. The main problem with the construction of speaker dependent model is the need for large database of utterances from one speaker. This problem is often non-solvable in real conditions, however, it can be overcome by adaptation techniques. The model is adapted on the specific speaker as well as on acoustic conditions (e.g. additive noise, channel distortion) in the test utterance.

The aim of this work is to discuss the methods for adaptation and the procedures of adaptation training. Some of these methods were tested and the experiments shown, that adaptation has a significant benefit for automatic speech recognition systems.

Kapitola 1

Úvod

Řeč, jako jeden z nejpoužívanějších způsobů předávání informací mezi lidmi, je v popředí zájmu oboru umělé inteligence již několik desítek let. Mezi problémy zpracování řeči počítačem patří, mimo jiné, úloha **automatického rozpoznávání řeči** (ASR – Automatic Speech Recognition), tedy úloha přepisu mluveného slova na text pomocí stroje. První automatické rozpoznávače se objevily v šedesátých letech minulého století, avšak jejich úspěšnost byla značně omezena tehdejšími možnostmi výpočetní techniky. První rozpoznávače se soustředily pouze na přepis izolovaných slov. Teprve v sedmdesátých letech, s příchodem myšlenky **skrytých Markovových modelů** (HMM – Hidden Markov Model) a prudkým rozvojem výpočetní techniky, došlo k nastartování vývoje systémů ASR a jejich směřování k rozpoznávání řeči spojité.

Se zdokonalováním ASR začal také růst počet slov obsažených v rozpoznávacím slovníku, z několika stovek v osmdesátých letech na několik tisíc v letech devadesátých. Systémy využívající slovník s velkým počtem slov se odborně označují **LVCSR** (Large Vocabulary Continuous Speech Recognition) systémy. Také kvalita rozpoznávané řeči přešla z čistých laboratorních dat k spontánním hovorům v rušném prostředí.

V současné době, kdy je obvyklé rozpoznávat spontánní hovory ve špatné akustické kvalitě, čelíme mnoha problémům. Jedním z nich jsou právě různé akustické podmínky v nahraných datech způsobené rozdílným nahrávacím prostředím, různým kanálem a odlišným řečníkem. To vše přidává nežádoucí varianci v nahraných datech. Při rozpoznávání testovacích dat s jinými akustickými vlastnostmi, než měla trénovací data použitá pro vytvoření akustického modelu, dochází k degradaci úspěšnosti rozpoznávání. Řešením by bylo použít model natrénovaný na datech se stejnými akustickými podmínkami jako v testovaných datech, to však v principu není zcela možné. Například získání dostatečného množství dat od jednoho řečníka pro natrénování akustického modelu je v praxi nereálné.

Z toho důvodu jsou již dvacet let vyvíjeny **adaptační techniky** normalizující testovací data nebo posouvající parametry akustického modelu směrem k testovacím datům. Úspěšnost rozpoznávání může být díky adaptaci výrazně zlepšena, a to již při použití několika málo promluv od cílového řečníka. Zároveň s řečníkem jsou adaptovány i akustické podmínky při nahrávaní, jako jsou typické ruchy prostředí, použité nahrávací zařízení atd.

Úkolem trénování je vytvořit model dobře odpovídající testovaným datům. V praxi však obecně máme nehomogenní data, která obsahují směs různých akustických zdrojů. Natréno-

vaný model se pak nazývá **multi-style model**. Tento model je možno použít pro testování nebo jej dále adaptovat na testované akustické podmínky, čím se zvýší jeho efektivita pro testovaná data. Problém velké variability v trénovacích datech tím ale není úplně odstraněn. Řešením je **adaptační trénování**, jehož úkolem je snížit variabilitu z trénovacích dat a vytvořit tzv. **kanonický model**, z něhož je vyloučena jakákoliv informace o prostředí či řečníkovi. Kanonický model je následně adaptován na testovací podmínky.

Tato práce si klade za cíl vysvětlit principy používaných adaptačních metod a postupy adaptačního trénování. Adaptace je zde popisována jako přizpůsobení se cílovému řečníku, ale z principu věci jde vlastně o obecnou adaptaci na akustické podmínky, protože cílový řečník není nic jiného než jiný akustický kanál pro přenos hlasu.

Automatickou adaptací rozumíme adaptaci bez vyššího zásahu člověka, tedy tzv. **adaptaci bez učitele** (*unsupervized adaptation*). Ta nepracuje s předem danými přesnými přepisy nahrávek, ale s textem z rozpoznávače s ohodnocením přesnosti jednotlivých přepisů (více v podkapitole 3.1).

1.1 Současný stav a struktura práce

Postupy adaptace jsou již dlouhou dobu v popředí zájmu převážně světových ale i tuzemských vědeckých pracovišť zabývajících se zpracováním řeči. Myšlenka adaptace vychází z metod pro trénování akustického modelu a je založena na předpokladu malého množství trénovacích (adaptačních) dat. Nové přístupy k trénování akustického modelu, využívající diskriminativní metody, jsou dále rozvíjeny v diskriminativních přístupech k adaptaci. Akustický model a postupy trénování jsou popsány v kapitole 2.

Bylo navrženo a dále rozvíjeno několik přístupů k adaptaci, ze kterých se v dnešní době nejvíce využívají metody lineární transformace, metoda maximalizace aposteriorní pravděpodobnosti (MAP), případně jejich kombinace. Tyto a další metody jsou popsány v kapitole 3.

Pro odstranění nežádoucí variability v akustickém modelu, a tím usnadnění adaptace takového modelu, slouží metody adaptačního trénování. Jim je věnována kapitola 4.

V kapitole 5 jsou uvedeny vlastní srovnávací experimenty jednotlivých nejpoužívanějších adaptačních metod a adaptačního trénování.

Adaptační metody obvykle zahrnují adaptaci složek jednotlivých stavů akustického modelu, tedy středních hodnot a kovariančních matic (obvykle diagonálních), výjimečně i jejich váhových koeficientů. Předpokládáme-li však, že původní model je multi-style modelem, tedy jeho trénovací data měla značnou varianci, bude tento model obsahovat nadbytečnou informaci i po zadaptování na cílového řečníka. Lze očekávat, že počet složek bude u zadaptovaného modelu nadbytečný, tedy některé složky budou pro cílového řečníka "nedostupné" a tedy zbytečné. Dalším problémem je časová náročnost adaptace, často nevhodná pro on-line aplikace. Závěrečná kapitola 6 se podrobněji zabývá právě těmito dvěma klíčovými problémy a nastiňuje další směr výzkumu na disertační práci.

Kapitola 2

Akustické modelování

Tato kapitola si klade za cíl přiblížit čtenáři základní principy modelování řeči pomocí akustického modelu reprezentovaného **skrytými Markovovými modely** (HMM – Hidden Markov Model). Je zde popsána struktura modelu a postupy při rozpoznávání posloupnosti řeči. Hlavní důraz je kladen na metody konstrukce HMM, neboť ty jsou základem adaptačních technik, jimiž se tato práce zabývá. Detailní popis trénování i využití skrytého Markovova modelu je možno nalézt v [PMMR06], [YEG+06] nebo [Rab89].

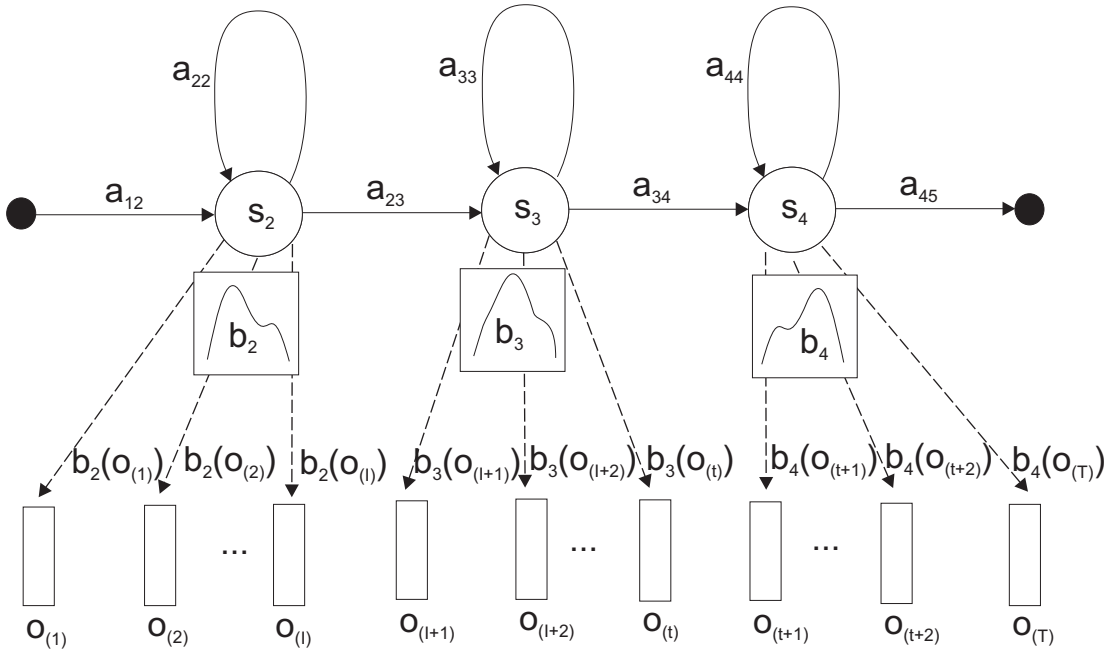
2.1 Struktura akustického modelu

Při rozpoznávání souvislé řeči jsou v dnešní době nejvíce dominantní klasifikátory pracující se statistickými metodami, kdy jsou slova (častěji subslovní jednotky jako slabiky, fonémy, trifóny a pod.) modelovány pomocí HMM. Vyslovená posloupnost slov W je nejprve rozčleněna na krátkodobé úseky, tzv. mikrosegmenty, po jejichž dobu předpokládáme, že parametry hlasového ústrojí jsou stacionární. Pro každý mikrosegment je vypočítán vektor příznaků $\boldsymbol{o}(t)$, který tvoří parametrizovaný přepis vyřčené promluvy $\boldsymbol{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$. Celá promluva je modelována zřetězením subslovních modelů HMM sériově za sebou. Cílem rozpoznávání je pak nalézt posloupnost slov W^*, která maximalizuje podmíněnou pravděpodobnost $P(\boldsymbol{O}|W)$ pro danou akustickou informaci $\boldsymbol{O}$.

Jako akustický model je uvažován skrytý Markovův model, patřící do množiny pravděpodobnostních konečných automatů, které mají tzv. Markovovu vlastnost, tedy současný stav modelu je závislý pouze na n stavech předcházejících. Skrytý se nazývá proto, že pozorovatel vidí jen výstup, ale posloupnost stavů modelu je mu skryta. Používají se zejména tzv. levo-pravé Markovovy modely, které jsou zvláště vhodné pro modelování procesů jako je spojitá řeč, jejichž vývoj je spojen s postupujícím časem.

Na skrytý Markovův model (příklad na obrázku 2.1) lze pohlížet jako na pravděpodobnostní konečný automat, který přechází ze stavu s_i do stavu s_j přes předem dané pravděpodobnostní přechody a_{ij} a tím generuje náhodnou posloupnost pozorování $\boldsymbol{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$. Stav s_j, do kterého model přejde, generuje příznakový vektor $\boldsymbol{o}(t)$ podle rozdělení výstupní pravděpodobnosti $b_j(\boldsymbol{o}(t))$.

Podmíněná pravděpodobnost přechodu a_{ij} určuje, s jakou pravděpodobností přechází mo-



Obrázek 2.1: Příklad třístavového skrytého Markovova modelu používaného pro modelování trifónů, převzatý z [PMMR06].

del ze stavu s_i v čase t do stavu s_j v čase $t+1$

$$a_{ij} = P(s(t+1) = s_j | s(t) = s_i). \tag{2.1}$$

Pravděpodobnost přechodu je v čase generování akustické informace pro všechny stavy s_i konstantní a pro $i = 1, 2, \ldots, N-1$ platí:

$$\sum_{j=2}^{N} a_{ij} = 1. \tag{2.2}$$

Výstupní pravděpodobnost $b_j(\boldsymbol{o}(t))$ popisuje rozdělení pravděpodobnosti pozorování $\boldsymbol{o}(t)$ produkovaného stavem s_j v čase t

$$b_j(\boldsymbol{o}(t)) = p(\boldsymbol{o}(t) | s(t) = s_j). \tag{2.3}$$

Ve stavech akustického modelu pro rozpoznávání plynulé řeči se v současné době nejvíce využívá rozdělení výstupní pravděpodobnosti jako **směsi M spojitých normálních hustotních funkcí pravděpodobnosti** (GMM – Gaussian Mixture Model)

$$b_j(\boldsymbol{o}(t)) = \sum_{m=1}^{M} \omega_{jm} b_{jm}(\boldsymbol{o}(t)), \tag{2.4}$$

$$\text{kde} \quad b_{jm}(\boldsymbol{o}(t)) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{C}_{jm}|}} \exp\left(-\frac{1}{2}(\boldsymbol{o}(t) - \boldsymbol{\mu}_{jm})^T \boldsymbol{C}_{jm}^{-1} (\boldsymbol{o}(t) - \boldsymbol{\mu}_{jm})\right), \tag{2.5}$$

$$\text{platí také} \quad \int_{\boldsymbol{o}} b_j(\boldsymbol{o}) d\boldsymbol{o} = 1, \tag{2.6}$$

kde M značí počet složek hustotní směsi, n je dimenze kovarianční matice, ω_{jm} vyjadřuje váhu m-té složky j-tého stavu, $\boldsymbol{\mu}_{jm}$ a $\boldsymbol{C}_{jm}$ značí střední hodnotu a kovarianční matici normálního pravděpodobnostního rozložení.

2.2 Výpočet pravděpodobnosti promluvy

Určení podmíněné pravděpodobnosti $P(\boldsymbol{O}|W)$ lze nahradit výpočtem $P(\boldsymbol{O}|\lambda)$ kde λ je skrytým Markovým modelem promluvy W. Výpočet pravděpodobnosti generování pozorované posloupnosti $\boldsymbol{O} = \{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$ modelem λ, u něhož není známa posloupnost stavů $S = s(0), s(1)...s(T+1)$, kterými posloupnost pozorování prošla, lze počítat jako součet pravděpodobností všech možných posloupností stavů:

$$P(\boldsymbol{O}|\lambda) = \sum_S P(\boldsymbol{O}, S|\lambda) = \sum_S \left(a_{s(0)s(1)} \left[\prod_{t=1}^{T} b_{s(t)}(\boldsymbol{o}(t)) a_{s(t)s(t+1)} \right] \right), \tag{2.7}$$

kde $s(0)$ je vstupní neemitující stav a $s(T+1)$ výstupní neemitující stav modelu λ. Neemitující stavy jsou takové, které negenerují žádná pozorování a nemají tedy žádné k nim příslušné rozdělení pravděpodobnosti. Skryté modely modelují jednotlivé řečové jednotky, neemitující stavy slouží k pospojování těchto jednotek v jakoukoliv řečovou posloupnost.

Přímý výpočet $P(\boldsymbol{O}|\lambda)$ je výpočetně náročný, proto byl navrhnut iterační **forward-backward algoritmus**, který snižuje složitost výpočtu průběžným ukládáním mezivýsledků, které jsou poté použity pro všechny posloupnosti stavů z S se stejnou počáteční sekvencí stavů. Alternativou k výpočtu $P(\boldsymbol{O}|\lambda)$ jako součtu přes všechny možné cesty délky T modelem λ je aproximovat tuto sumu pouze jednou nejpravděpodobnější posloupností stavů, se kterou projde posloupnost $\boldsymbol{O}$ modelem λ

$$P_S(\boldsymbol{O}|\lambda) = \max_S P(\boldsymbol{O}, S|\lambda) = \max_S \left(a_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\boldsymbol{o}(t)) a_{s(t)s(t+1)} \right). \tag{2.8}$$

Pro nalezení optimální posloupnosti stavů a vypočtení pravděpodobnosti $P_S(\boldsymbol{O}|\lambda)$ se využívá tzv. **Vitterbiův algoritmus** [Vit67] pracující na principu dynamického programování.

2.2.1 Rekurzivní výpočet forward-backward algoritmem

Při výpočtu odpředu (forward) definujeme sdruženou pravděpodobnost $\alpha_j(t)$ pozorování posloupnosti prvních t akustických vektorů $\{\boldsymbol{o}(1), \ldots, \boldsymbol{o}(t)\}$ končící v aktuálním stavu s_j v čase t za podmínky modelu λ

$$\alpha_j(t) = P(\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(t), s(t) = s_j|\lambda). \tag{2.9}$$

Pro výpočet odzadu (backward) definujeme podmíněnou pravděpodobnost $\beta_j(t)$ pozorování posloupnosti posledních $T-t$ akustických vektorů $\{\boldsymbol{o}(t+1), \boldsymbol{o}(t+2), \ldots, \boldsymbol{o}(T)\}$ za podmínky, že model λ je v čase t ve stavu s_j

$$\beta_j(t) = P(\boldsymbol{o}(t+1), \boldsymbol{o}(t+2), \ldots, \boldsymbol{o}(T)|s(t) = s_j, \lambda). \tag{2.10}$$

Konkrétní algoritmy výpočtu pravděpodobnosti $P(\boldsymbol{O}|\lambda)$ lze nalézt např. v [PMMR06]. Hledaná pravděpodobnost $P(\boldsymbol{O}|\lambda)$ může být snadno vyčíslena kombinací proměnných $\alpha_j(t)$ a $\beta_j(t)$

$$P(\boldsymbol{O}|\lambda) = \sum_{i=2}^{N-1} \alpha_i(t)\beta_i(t). \tag{2.11}$$

2.2.2 Iterativní Viterbiho algoritmus

Při procházení modelu si algoritmus uchovává proměnnou $\varphi_j(t)$ určující pravděpodobnost maximálně pravděpodobné posloupnosti stavů $s(1), s(2), \ldots, s(t) = s_j$ pro částečnou posloupnost pozorování $\{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(t)\}$

$$\varphi_j(t) = \max_{s(1),\ldots,s(t-1)} P(\boldsymbol{o}(1), \ldots, \boldsymbol{o}(t), s(1), \ldots, s(t) = s_j|\lambda). \tag{2.12}$$

Algoritmus postupuje odpředu, ale pro určení maximálně pravděpodobné posloupnosti stavů je potřeba si při jeho výpočtu ještě pamatovat v každém časovém kroku t, z kterého stavu v předchozím kroku byla vybrána maximální hodnota. K tomuto účelu je v algoritmu zavedena proměnná $\Psi_j(t)$, která se využívá při zpětném trasování k nalezení maximálně pravděpodobné cesty modelem λ pro posloupnost $\{\boldsymbol{o}(1), \boldsymbol{o}(2), \ldots, \boldsymbol{o}(T)\}$. Kompletní algoritmus lze nalézt např. v [PMMR06].

2.3 Trénování parametrů akustického modelu

Stanovení topologie skrytého Markovova modelu je úlohou expertního návrhu, vycházejícího z vlastností spojité řeči. Naopak ke stanovení parametrů modelu dochází na základě statistických metod aplikovaných na trénovací data, která jsou předem zanotována [PMMR06]. Parametry skrytého Markovova modelu jsou pravděpodobnosti přechodů a_{ij} a výstupní pravděpodobnosti $b_j(.)$ vyjádřené pomocí hustotní směsi normálního rozdělení s vahami ω_{jm}, středními hodnotami $\boldsymbol{\mu}_{jm}$ a kovariančními maticemi $\boldsymbol{C}_{jm}$

$$\lambda = \{a_{ij}, \omega_{jm}, \boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}\}, \quad \text{kde } 1 \leq i, j \leq N \text{ a } 1 \leq m \leq M. \tag{2.13}$$

2.3.1 Metoda maximální věrohodnosti (ML)

Jako metoda odhadu parametrů bývá pro svou efektivitu často využívána **metoda maximální věrohodnosti** (ML – Maximum Likelihood), která maximalizuje výpočet pravděpodobnosti modelu

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda) \tag{2.14}$$

pro soubor E známých trénovacích promluv $\{\boldsymbol{O}^e\}_{e=1}^E$, kde $\boldsymbol{O^e} = \{\boldsymbol{o}^e(1), \boldsymbol{o}^e(2), \ldots, \boldsymbol{o}^e(T_e)\}$. Využívá se **Fisherova funkce věrohodnosti**

$$F(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda) = P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda) = \prod_{e=1}^{E} P(\boldsymbol{O}^e|\lambda), \tag{2.15}$$

která je maximalizována přes neznámé parametry modelu λ (v praxi se spíše pracuje s logaritmem věrohodnostní funkce)

$$\hat{\lambda} = \arg\max_{\lambda} \log \prod_{e=1}^{E} P(\boldsymbol{O}^e|\lambda) = \arg\max_{\lambda} \sum_{e=1}^{E} \log P(\boldsymbol{O}^e|\lambda). \tag{2.16}$$

Pro stanovení optimálních parametrů modelu λ, tedy nalezení globálního maxima věrohodnostní funkce, v podstatě neexistuje žádná explicitní metoda. Efektivně se však k výpočtu využívá iterativního **Baum-Welchova algoritmu** [BPSW70], který je speciálním případem **EM algoritmu** (EM – expectation-maximization) [DLR77]. Ten nalezne parametry modelu, které zabezpečí pouze lokální maximum funkce $P(\boldsymbol{O}|\lambda)$, výsledek tedy závisí na počáteční volbě parametrů.

EM algoritmus

Nejprve zavedeme skrytou proměnnou y^e, která ponese informaci o indexech stavů $s^e(t)$ a indexech složek hustotní směsi $m^e(t)$, tedy y^e je časová posloupnost dvojic $[s^e(t), m^e(t)], t = 1, \ldots, T_e$. Pak lze odvodit reestimační vztahy pro EM algoritmus ze vztahu:

$$P(\boldsymbol{O}^e|\lambda) = \sum_{y_e} P(\boldsymbol{O}^e, y^e|\lambda) = \sum_{y_e} P(y^e|\lambda) P(\boldsymbol{O}^e|y^e, \lambda) = \sum_{y_e} P(\boldsymbol{O}^e|\lambda) P(y^e|\boldsymbol{O}^e, \lambda). \tag{2.17}$$

Pokud uvažujeme rozdíl logaritmů věrohodnostních funkcí dvou modelů λ a $\bar{\lambda}$, platí po úpravě [PMMR06]:

$$\sum_{e=1}^{E} \log \frac{P(\boldsymbol{O}^e|\bar{\lambda})}{P(\boldsymbol{O}^e|\lambda)} = \sum_{e=1}^{E} \sum_{y^e} P(y^e|\boldsymbol{O}^e, \lambda) \log \left[\frac{P(\boldsymbol{O}^e, y^e|\bar{\lambda})}{P(\boldsymbol{O}^e, y^e|\lambda)} \frac{P(y^e|\boldsymbol{O}^e, \lambda)}{P(y^e|\boldsymbol{O}^e, \bar{\lambda})} \right]. \tag{2.18}$$

Vhodnou úpravou a aplikací nerovnosti $z \leq z - 1 (z \geq 0)$ dostáváme základní nerovnost EM algoritmu

$$\sum_{e=1}^{E} \log \frac{P(\boldsymbol{O}^e|\bar{\lambda})}{P(\boldsymbol{O}^e|\lambda)} \geq \sum_{e=1}^{E} \sum_{y_e} P(y^e|\boldsymbol{O}^e, \lambda) \log \frac{P(\boldsymbol{O}^e, y^e|\bar{\lambda})}{P(\boldsymbol{O}^e, y^e|\lambda)} = Q(\lambda, \bar{\lambda}) - Q(\lambda, \lambda), \tag{2.19}$$

$$\text{kde} \quad Q(\lambda, \bar{\lambda}) = \sum_{e=1}^{E} \sum_{y_e} P(y^e|\boldsymbol{O}^e, \lambda) \log P(\boldsymbol{O}^e, y^e|\bar{\lambda}). \tag{2.20}$$

Tato nerovnost říká, že pokud vybereme model $\bar{\lambda}$ tak, abychom dosáhli přírůstku funkce $Q(\lambda, \bar{\lambda})$ oproti funkci $Q(\lambda, \lambda)$, pak vzroste i logaritmus věrohodnostní funkce $\sum_{e=1}^{E} \log P(\boldsymbol{O}^e|\bar{\lambda})$. Výpočet EM algoritmu probíhá iterativně ve dvou krocích, nejprve vypočteme očekávání (expectation) funkce $Q(\lambda, \bar{\lambda})$ a následně vybereme takový model $\bar{\lambda}$, který maximalizuje (maximalization) funkci $Q(\lambda, \bar{\lambda})$. Odvození algoritmu lze nalézt mimo jiné v [DRN95].

Rozepsáním pravděpodobnostní funkce pro jednotlivé parametry hustotních směsí modelu $\bar{\lambda}$ a dosazením do vztahu (2.20) dostáváme vztah pro přírůstkovou funkci $Q(\lambda, \bar{\lambda})$ s vyjádřenými parametry hustotních směsí

$$Q(\lambda, \bar{\lambda}) = \sum_{e=1}^{E} \sum_{y^e} P(y^e|\boldsymbol{O}^e, \lambda) \log P(\boldsymbol{O}^e, y^e|\bar{\lambda}) =$$

$$= \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \sum_{y^e} P(\boldsymbol{O}^e, y^e|\lambda) \log \left[\prod_{t=1}^{T_e} (\bar{a}_{s^e(t-1)s^e(t)} + \bar{c}_{s^e(t)m_t^e} + \bar{b}_{s^e(t)m_t^e}(\boldsymbol{o}^e(t))) + \bar{a}_{s^e(T_e)s^e(T_e+1)} \right]. \tag{2.21}$$

Tuto rovnici použijeme k odvození vztahů pro trénování parametrů modelu.

Reestimační Baum-Welchův algoritmus

Jde o speciální případ EM algoritmu, platí pro něj tedy stejné vztahy, které byly odvozeny v předchozí sekci. Nově odhadnutý model $\bar{\lambda}$ v každém kroku (pomocí maximalizace funkce $Q(\lambda, \bar{\lambda})$) zvyšuje pravděpodobnost modelu $P(\boldsymbol{O}^e|\bar{\lambda}) \geq P(\boldsymbol{O}^e|\lambda)$ až do posledního kroku, kdy $P(\boldsymbol{O}^e|\bar{\lambda}) = P(\boldsymbol{O}^e|\lambda)$. Popis algoritmu lze nalézt například v [PMMR06].

2.3.2 Metoda maximální aposteriorní pravděpodobnosti (MAP)

Metoda maximální aposteriorní pravděpodobnosti (MAP – Maximum A Posteriory) [PMMR06] staví také na ML kritériu (viz 2.3.1), rozdíl však je v uvažování λ jako náhodného vektoru a ne jako pevnou hodnotu (jak je tomu v metodě ML) . MAP kombinuje informaci získanou apriorním modelem λ s informací z trénovacích dat. Výhodou metody MAP je potřeba menšího množství trénovacích dat oproti metodě ML.

Úlohu nalezení parametrů λ lze formulovat na základě maximální pravděpodobnosti následovně:

$$\lambda^* = \arg\max_{\lambda} P(\lambda|\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E). \tag{2.22}$$

Využitím Bayesova pravidla dostáváme vztah:

$$\lambda^* = \arg\max_{\lambda} \frac{P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda)P(\lambda)}{P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E)}. \tag{2.23}$$

Jmenovatel $P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E)$ je pro všechny hodnoty λ konstantní, tady vztah (2.23) lze zjednodušit na tvar

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda)P(\lambda), \tag{2.24}$$

kde $P(\lambda)$ je apriorní informace rozdělení vektoru parametrů, což je jediná odlišnost od metody maximální věrohodnosti (2.14). Opět se využije Fisherova funkce věrohodnosti (2.15) jako při odvozování metodou ML.

Pro parametry diskrétních rozdělení, jako je případ pravděpodobností přechodu a_{ij} a vah hustotní směsi ω_{ij}, se jako apriorní hustota volí Dirichletovo rozdělení. Pro mnoharozměrné normální rozdělení s vektorem středních hodnot $\boldsymbol{\mu}$ a plnou kovarianční maticí $\boldsymbol{C}$ se volí apriorní hustota ve tvaru normálního-Wishartova rozdělení. Odvozené vztahy pro nové parametry modelu $\bar{\lambda}$ metodou MAP mají následující tvar:

$$\bar{a}_{1j} = \frac{(\eta_{1j} - 1) + \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \alpha_j^e(1)\beta_j^e(1)}{\sum_{i=2}^{N-1}(\eta_{1i} - 1) + E}, \tag{2.25}$$

$$\bar{a}_{ij} = \frac{(\eta_{ij} - 1) + \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \sum_{t=1}^{T_e - 1} \alpha_i^e(t) a_{ij} b_j(\boldsymbol{o}^e(t)) \beta_j^e(t+1)}{\sum_{i=2}^{N-1}(\eta_{1i} - 1) + \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \sum_{t=1}^{T_e} \alpha_i^e(t) \beta_i^e(t)}, \tag{2.26}$$

$$\bar{a}_{iN} = \frac{(\eta_{1N} - 1) + \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \alpha_i^e(T_e) \beta_i^e(T_e)}{\sum_{i=2}^{N-1}(\eta_{1i} - 1) + \sum_{e=1}^{E} \frac{1}{P(\boldsymbol{O}^e|\lambda)} \sum_{t=1}^{T_e} \alpha_i^e(t) \beta_i^e(t)}, \tag{2.27}$$

$$\bar{\omega}_{jm} = \frac{(\upsilon_{jm} - 1) + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t)}{\sum_{\acute{m}=1}^{M} \left[(\upsilon_{j\acute{m}} - 1) + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{j\acute{m}}^e(t) \right]}, \tag{2.28}$$

$$\bar{\boldsymbol{\mu}}_{jm} = \frac{\tau_{jm} \boldsymbol{\zeta}_{jm} + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t) \boldsymbol{o}^e(t)}{\tau_{jm} + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t)}, \tag{2.29}$$

$$\bar{\boldsymbol{C}}_{jm} = \frac{\boldsymbol{u}_{jm} + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t) (\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})(\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})^T - \tau_{jm} (\bar{\boldsymbol{\mu}}_{jm} - \boldsymbol{\zeta}_{jm})(\bar{\boldsymbol{\mu}}_{jm} - \boldsymbol{\zeta}_{jm})^T}{\alpha_{jm} - n + \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t)}, \tag{2.30}$$

kde

$$\gamma_{jm}^e(t) = \gamma_j^e(t) \frac{\mathcal{N}(\boldsymbol{o}^e(t)|\boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm})}{\sum_{\acute{m}=1}^{M} c_{j\acute{m}} \mathcal{N}(\boldsymbol{o}^e(t)|\boldsymbol{\mu}_{j\acute{m}}, \boldsymbol{C}_{j\acute{m}})}. \tag{2.31}$$

Matice $\boldsymbol{u}_{jm}$ řádu n, vektor $\boldsymbol{\zeta}_{jm}$ a skaláry τ_{jm}, α_{jm} jsou parametry normálního-Wishartova apriorního rozdělení m-té komponenty j-tého stavu a η_{ij}, υ_{jm} jsou složky vektorů parametrů Dirichletových apriorních hustot pravděpodobností přechodů z i-tého do j-tého stavu HMM a vah m-té komponenty hustotní směsi j-tého stavu HMM[1]. Souhrnně se tyto parametry nazývají "hyperparametry" a reprezentují parametry apriorního modelu. Nalezení hyperparametrů je složitý problém, jednou z možností je odhadovaní přímo z trénovacích dat [PMMR06].

2.3.3 Diskriminativní trénování (DT)

Nejpoužívanější přístup k trénování (ML kritérium) je vhodný pro rychlé vytvoření dobrého modelu využitím Baum-Welchova algoritmu. Tento přístup vykazuje nejlepší vlastnosti za určitých předpokladů, které je však často velmi obtížné splnit. Jedním z nich je stacionarita řečového ústrojí v mikrosegmentech řeči, tedy že řeč je generována diskrétně. Druhým nesplnitelnou podmínkou je předpoklad nekonečného množství dat pro trénování [Yu06]. Pro překonání těchto problémů byla navržena alternativní **diskriminativní kritéria pro trénování** (DT – Discriminative Training) HMM modelu. V mnoha odborných pracích bylo dokázáno [Pov03], že diskriminativní trénování muže zlepšit úspěšnost rozpoznávání vytvořeného modelu formulováním funkce, která penalizuje parametry snižující správnost rozpoznávání. Diskriminativní trénování se snaží nastavit parametry modelu tak, aby jednotlivé stavy odpovídaly svým pozorováním s největší pravděpodobností a zároveň minimalizuje pravděpodobnost pozorování patřících jiným stavům modelu.

[1] Pro hodnoty $\eta_{ij} = 1, \upsilon_{jm} = 1, \boldsymbol{u}_{jm} = 0$ a $\alpha_{jm} = n$ nabývají vztahy (2.25) až (2.30) pro metodu MAP stejného tvaru jako rovnice pro metodu ML, tedy apriorní rozložení nenese žádnou informaci a odhad nových parametrů je proveden jen na základě trénovacích dat.

Jednotlivá diskriminativní kritéria:

- **Maximalizace vzájemné informace** (MMI – Maximum Mutual Information)[Cho90] umožňuje vybrat sekvenci slov s minimální nejistotou správné hypotézy. Tento přístup využívá informaci o správném přepisu promluvy $\boldsymbol{O}$ (tzv. referenční přepis $\boldsymbol{W_{ref}}$) a informaci o všech možných přepisech $\boldsymbol{W}$ (včetně toho správného). V praxi jsou uvažovány pouze N-nejlepší přepisy získané z rozpoznávače nebo N-nejpravděpodobnějších cest ze slovní mřížky. Toto kritérium lze napsat ve formě

$$\mathcal{F}_{MMI}(\lambda) = \frac{p^{\kappa}(\boldsymbol{O}|W_{ref},\lambda)P(W_{ref})}{\sum_W p^{\kappa}(\boldsymbol{O}|W,\lambda)P(W)}, \tag{2.32}$$

  kde W_{ref} je přepis nahrávky $\boldsymbol{O}$, zatímco W značí všechny možné přepisy, včetně toho správného. λ je HMM model. κ je empiricky volený faktor, kterým lze měnit poměr mezi pravděpodobností správného přepisu a pravděpodobností ostatních přepisů, tedy lze jím regulovat míru diskriminativnosti výsledného modelu. Podobné kritérium **Maximalizace vzájemné informace pomocí diskriminace pozorování** (MMI-FD – Maximum Mutual Information Frame Discrimination) [PW99] pracuje přímo s vektory pozorování a jejich příslušností ke stavům modelu namísto informací ze slovní mřížky.

- **Minimalizace chyby klasifikace** (MCE – Minimum Classification Error)[MHSN05] minimalizuje chybu očekávání přidáním ztrátové funkce $l(W, W_{ref})$ k disktriminativnímu kritériu

$$\mathcal{F}_{MCE}(\lambda) = \frac{p^{\kappa}(\boldsymbol{O}|W_{ref},\lambda)P(W_{ref})}{\sum_W p^{\kappa}(\boldsymbol{O}|W,\lambda)P(W)} l(W, W_{ref}), \tag{2.33}$$

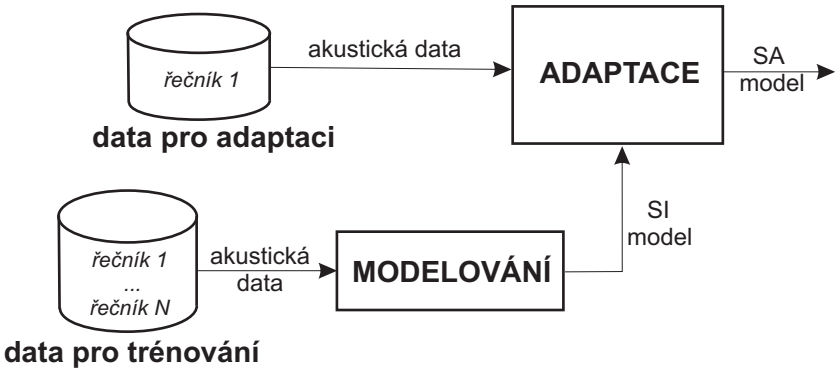
  kde opět W_{ref} je referenční přepis nahrávky $\boldsymbol{O}$, W značí všechny možné přepisy a κ je empiricky volený faktor. Možností jak vypočítat $l(W, W_{ref})$ je uvažovat **minimalizaci chyby fonému** (MPE – Minimum Phone Error) [Pov03] nebo **minimalizaci chyby slova** (MWE – Minimum Word Error) [SM98].

Výše vyjmenované přístupy jsou vzájemně kombinovatelné, což přináší další zlepšení účinnosti akustického modelu [ZS05]. Nevýhodou diskriminativního přístupu je potřeba většího množství dat pro trénování než pro klasické ML kritérium.

Kapitola 3

Metody adaptace

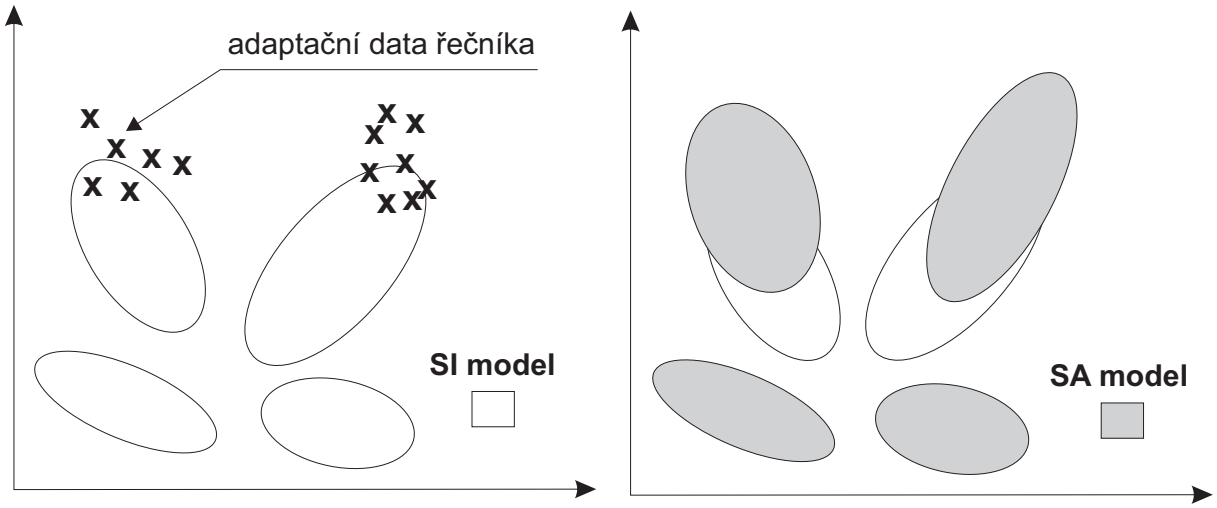
Skrytý Markovovův model (HMM – Hidden Markov Model) v kombinaci s **modelem Gaussovských směsí** (GMM – Gaussian Mixture Model) se v poslední době stal účinným nástrojem pro modelování akustických příznaků v úloze rozpoznávaní řeči [PMMR06]. Pro natrénování modelu je potřeba velkého množství dat od velkého počtu řečníků. Výsledný model, **na řečníku nezávislý** (SI – Speaker Independent), pak dovede rozpoznávat řeč libovolného řečníka (viz obr. 3.1). Trénovací data pro SI model jsou v jisté míře průměrná, pokud



Obrázek 3.1: Schématické znázornění adaptace.

je však totožnost řečníka při rozpoznávání známá, bylo by možné dosáhnout větší úspěšnosti natrénováním modelu jenom z dat konkrétního řečníka, kterého budeme chtít testovat. Takovému modelu se pak říká **na řečníku závislý** (SD – Speaker Dependent). Problémem při tvorbě SD modelu je nutnost mít k dispozici velký počet trénovacích promluv od jednoho řečníka, což je v praxi často nemožné získat. Řešení poskytuje adaptace SI modelu na data konkrétního řečníka, vzniklý model je **na řečníka adaptovaný** (SA – Speaker Adaptive). Jde vlastně o transformaci SI modelu ve smyslu dosažení maximální pravděpodobnosti pro nová data (viz obrázek 3.2).

Na rozdíl od vlastního trénování akustického modelu využívá adaptace apriorní znalost o rozložení parametrů akustického modelu. Tato znalost je obvykle odvozována z předem natrénovaného SI modelu. Adaptace přizpůsobuje SI model tak, aby byla maximalizována



Obrázek 3.2: Ilustrativní příklad adaptace modelu. Hustoty rozložení složek SI modelu (zde reprezentovány elipsou) se "posunou" ve směru adaptačních dat tak, aby SA model tato data lépe modeloval.

pravděpodobnost adaptačních dat:

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}^1, \ldots, \boldsymbol{O}^E|\lambda)P(\lambda), \tag{3.1}$$

kde $P(\lambda)$ představuje apriorní informaci o rozdělení vektoru parametrů modelu λ (dána obvykle SI modelem), $\boldsymbol{O}^e = \{\boldsymbol{o}^e(1), \boldsymbol{o}^e(2), \ldots, \boldsymbol{o}^e(T_e)\}, e = 1, \ldots, E$, je posloupnost vektorů příznaků přidružených jednomu řečníkovi, λ^* je nejlepším odhadem parametrů SA modelu.

3.1 Obecné dělení adaptačních metod

Adaptačních přístupů a z nich vyplývajících různých metod k adaptaci je velké množství. Obecně je možné dělit tyto metody dle několika kritérií podle jejich vlastností.

- Adaptace může probíhat buď **za chodu aplikace** (on-line), nebo může být provedena **před vlastním testováním** (off-line).
- Pokud máme při adaptaci k dispozici přesný fonetický přepis adaptačních dat, značíme úlohu za **adaptaci s učitelem** (supervised adaptation). Pokud však přesný přepis nemáme, **adaptace bez učitele** (unsupervised adaptation), lze jej nahradit automatickým přepisem pomocí SI modelu. Výsledný přepis obvykle obsahuje nepřesnosti a chyby, které lze odstranit například využitím adaptovaného modelu v další iteraci (zpřesňujeme přepis a tím i SA model), popřípadě uvažováním **míry jistoty** (CM – Certainty Measure) [WSMN01] přepsaných slov jako výstupu z jazykového modelu (bereme jen slova, která se rozpoznala s dostatečně velkou jistotou).
- Adaptační metody lze dělit podle toho, zda **transformují parametry modelu** (model transform) nebo **transformují vektory pozorování** (feature transform). Druhá možnost má výhodu v paměťových nárocích, protože si není třeba pamatovat pro

každého řečníka celý model ale jen transformaci, která upraví jeho testované nahrávky na lepší rozpoznání SI modelem.

- Pokud jsou při adaptaci použita všechna data najednou, jedná se o **dávkovou adaptaci** (batch adaptation). Pokud se však systém adaptuje postupně, jak přicházejí nová adaptační data, jde o **inkrementální adaptaci** (incremental adaptation), která se nejčastěji používá v on-line systémech.

- Pro vygenerování SA modelu lze použít přístup **generativní adaptace** (generative adaptation), kdy složky modelu nejlépe reprezentují svá data. Jiným přístupem je **diskriminativní adaptace** (discriminative adaptation), kdy složky SA modelu nejlépe reprezentují svá data, ale navíc se co nejméně vzájemně překrývají.

- Při určování efektivity adaptačních metod lze uvažovat také **množství dat**, které jsou pro adaptaci k dispozici. Adaptovaný SA model konverguje k modelu SD konkrétního řečníka při dostatečném množství adaptačních dat (množství, které by bylo potřebné pro vlastní natrénování SD modelu). Na druhou stranu, pro menší počet adaptačních dat je adaptace rychlá a přitom je dobrou aproximací SD modelu.

Parametry, které nesou nejdůležitější informaci o řečníkovi, jsou střední hodnoty a kovarianční matice výstupních pravděpodobností stavů HMM tvořených GMM. Následující vzorce jsou pro jednotlivé adaptační techniky společné a v dalším textu na ně bude odkazováno.

Nechť

$$\gamma_{jm}^{e}(t) = \frac{\omega_{jm} p(\boldsymbol{o}^{e}(t)|jm)}{\sum_{m=1}^{M} \omega_{jm} p(\boldsymbol{o}^{e}(t)|jm)} \tag{3.2}$$

je aposteriorní pravděpodobnost, že pozorování $\boldsymbol{o}(t)$ je generováno m-tou složkou Gaussovské směsi j-tého stavu HMM. ω_{jm}, $\boldsymbol{\mu}_{jm}$ a $\boldsymbol{C}_{jm}$ je váha, střední hodnota a kovarianční matice m-té složky v j-tém stavu HMM, pak lze definovat

$$c_{jm} = \sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma_{jm}^{e}(t) \tag{3.3}$$

jako obsazení m-té složky v j-tém stavu HMM přes všechny časy t a vektor

$$\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}) = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^{e}(t) \boldsymbol{o}^{e}(t)}{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^{e}(t)} \tag{3.4}$$

$$\text{resp.} \quad \boldsymbol{\varepsilon}_{jm}(\boldsymbol{o} \cdot \boldsymbol{o}^{\boldsymbol{T}}) = \frac{\sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^{e}(t) \boldsymbol{o}^{e}(t) \boldsymbol{o}^{eT}(t)}{\sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma_{jm}^{e}(t)} \tag{3.5}$$

jako průměrnou hodnotu počtu příznaků ve vektoru příznaků patřících m-té složce GMM v j-tém stavu HMM.

3.2 Metoda maximální aposteriorní pravděpodobnosti (MAP)

Metoda **maximální aposteriorní pravděpodobnosti** (MAP – Maximum A-Posteriori) je založena na Bayesově metodě odhadu parametrů akustického modelu s jednotkovou ztrátovou funkcí [GL94]. Nástin odvození metody byl uveden v podkapitole 2.3.2, kde byly také

uvedeny vztahy pro přepočet nových parametrů modelu $\bar{\lambda}$, které maximalizují funkci $Q(\bar{\lambda}, \lambda)$. V případě adaptace odpadá problém hledání apriorních parametrů (tzv. hyperparametrů). Jako apriorní model je brán v úvahu právě námi adaptovaný SI model. Zbylé hyperparametry mají význam experimentálně určené adaptační konstanty τ. V praxi je výhodné adaptovat především vektory středních hodnot $\boldsymbol{\mu}_{jm}$, popřípadě i kovarianční matice $\boldsymbol{C}_{jm}$ a váhy ω_{jm} jednotlivých složek hustotních směsí modelu, zbylé parametry zůstávají totožné s apriorním modelem.

Z (2.25) až (2.27) lze odvodit následující vztahy pro MAP adaptaci:

$$\bar{\omega}_{jm} = \left[\frac{\alpha_{jm} c_{jm}}{T} + (1 - \alpha_{jm})\omega_{jm}\right] \chi \quad , \tag{3.6}$$

$$\bar{\boldsymbol{\mu}}_{jm} = \alpha_{jm} \boldsymbol{\varepsilon}_{jm}(\boldsymbol{o}) + (1 - \alpha_{jm})\boldsymbol{\mu}_{jm} \quad , \tag{3.7}$$

$$\bar{\boldsymbol{C}}_{jm} = \alpha_{jm} \boldsymbol{\varepsilon}_{jm}(\boldsymbol{o} \cdot \boldsymbol{o}^T) + (1 - \alpha_{jm})(C_{jm} + \boldsymbol{\mu}_{jm}\boldsymbol{\mu}_{jm}^T) - \bar{\boldsymbol{\mu}}_{jm}\bar{\boldsymbol{\mu}}_{jm}^T \quad , \tag{3.8}$$

$$\alpha_{jm} = \frac{c_{jm}}{c_{jm} + \tau} \quad , \tag{3.9}$$

kde c_{jm} a $\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o})$ jsou definovány vztahy (3.3), respektive (3.4). χ je normalizační parametr, který garantuje, že všechny adaptované váhy každého GMM budou v součtu rovny jedné. α_{jm} je adaptační koeficient, který kontroluje rovnováhu mezi starými a novými parametry. K tomu je využívána empiricky určená konstanta τ, která nám říká, jak moc se mají staré parametry posunout ve směru nových parametrů určených z adaptačních dat. Čím více dat k danému parametru máme, tím méně se původní hodnota projeví na výsledku. Při malém počtu adaptačních dat pro konkrétní parametr se adaptace metodou MAP pro tento parametr neprojeví. Adaptovaný model metodou MAP konverguje k výsledku získanému klasickým trénováním pro dostatečné množství dat.

3.2.1 Diskriminativní MAP (DMAP)

Klasická metoda MAP je založena na kritériu ML, viz (2.14). Takto adaptovaný model trpí stejnými problémy, které byly zmíněny v úvodu do diskriminativního trénování 2.3.3. Metoda **diskriminativní MAP** (DMAP – Discriminative MAP) naproti tomu staví na některých z kritérií definovaných pro diskriminativní trénování 2.3.3, jako je například v [GRP00] kritérium MMI (2.32). Maximalizováním MMI kritéria zabezpečíme rostoucí pravděpodobnost pro správné přepisy, zatímco pravděpodobnost pro ostatní přepisy se bude snižovat, což vede k diskriminativnímu charakteru adaptace.

Pomocí MMI kritéria lze odvodit vztahy pro DMAP adaptaci. Na rozdíl od klasického MAP se nasčítávají statistiky $\gamma_{jm}(t)$ a $\gamma_{jm}^{den}(t)$, kde $\gamma_{jm}(t)$ se počítá dle čitatele MMI kritéria, tedy stejně jak je definováno v (3.2), zatímco horní index *den* označuje jmenovatel MMI kritéria, tedy $\gamma_{jm}^{den}(t)$ je počítáno právě s využitím všech možných přepisů.

Pro DMAP je pak nutno pravděpodobnostní momenty $c_{jm}(t)$, $\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o})$ a $\boldsymbol{\varepsilon}_{jm}(\boldsymbol{o} \cdot \boldsymbol{o}^T)$ nahradit rozdílem původní hodnoty, vypočtené dle vzorců (3.3),(3.4), respektive (3.5), a hodnoty spočtené stejnými vztahy, pouze $\gamma_{jm}(t)$ je nahrazeno $\gamma_{jm}^{den}(t)$.

Podrobnější odvození DMAP pro diskriminativní kritérium MMI i MPE lze nalézt např. v [PGKW03].

3.3 Metody adaptace založené na lineární transformaci

Základním nedostatkem adaptační metody MAP je potřeba dostatečného množství dat pro každý parametr akustického modelu. Jelikož je adaptovaných parametrů v modelu velmi mnoho, metoda vyžaduje nemalé množství adaptačních nahrávek, kterých se nám často nedostává. Metody založené na **lineárních transformacích** [Gal97] omezují počet volných parametrů modelu shlukováním akusticky podobných složek stavů do tříd C, které pak adaptují stejným způsobem. Díky shlukování složek poskytují tyto metody dobré výsledky i s relativně malým počtem adaptačních dat (v porovnání s MAP) a samotná adaptace pak může být mnohem rychlejší. Metody se snaží pro každý shluk nalézt takovou lineární transformaci, kdy by adaptované parametry akustického modelu lépe odpovídaly hlasu konkrétního řečníka. Všechny parametry v jednom shluku se pak adaptují stejnou lineární transformací. Pro výpočet transformace je pak dostatek dat a adaptačními daty nepokryté parametry jsou také zadaptovány. Více o shlukování parametrů v podkapitole 3.3.4.

V této práci rozlišujeme dva způsoby lineárních transformací modelu, a to **neomezenou** (unconstrained) a **omezenou** (constrained) transformaci. První z nich používá jiné transformační vztahy pro střední hodnoty a jiné pro kovarianční matice, na rozdíl od druhého způsobu, kde jsou tyto parametry transformovány stejnou transformační maticí. Dále lze u každé metody rozlišit, zda je adaptace zaměřena na **transformaci parametrů modelu** nebo na **transformaci příznaků pozorování**.

3.3.1 Metoda maximální věrohodné lineární regrese (MLLR)

Nejčastěji používaná adaptační technika ze skupiny lineárních transformací je metoda **maximální věrohodné lineární regrese** (MLLR – Maximum Likelihood Linear Regression) [PS06]. Metoda je založena na neomezené transformaci, tedy střední hodnoty a kovarianční matice jsou transformovány různými transformacemi. Předpokládejme opět adaptační data ve formě $\boldsymbol{O}^e = \{\boldsymbol{o}^e(1), \boldsymbol{o}^e(2), \ldots, \boldsymbol{o}^e(T_e)\}, e = 1, \ldots, E$.

Lineární transformace střední hodnoty je dána:

$$\bar{\boldsymbol{\mu}}_{jm} = \boldsymbol{A}_{(n)}\boldsymbol{\mu}_{jm} + \boldsymbol{b}_{(n)} = \boldsymbol{W}_{(n)}\boldsymbol{\xi}_{jm}, \tag{3.10}$$

kde $\boldsymbol{\mu}_{jm}$ je stará (původní) střední hodnota m-té složky GMM v j-tém stavu HMM, $\bar{\boldsymbol{\mu}}_{jm}$ je nová (adaptovaná) střední hodnota, $\boldsymbol{\xi}_{jm}^T = [\boldsymbol{\mu}_{jm}^T, 1]$ je původní střední hodnota rozšířená o 1, $\boldsymbol{A}_{(n)}$ je regresní matice a $\boldsymbol{b}_{(n)}$ je aditivní vektor, $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$ je transformační matice pro třídu C_n.

Transformace kovarianční matice je vyjádřena vztahem:

$$\bar{\boldsymbol{C}}_{jm} = \boldsymbol{L}\boldsymbol{H}_{(n)}\boldsymbol{L}^T, \tag{3.11}$$

kde $\boldsymbol{H}_{(n)}$ je transformační matice pro třídu C_n a $\boldsymbol{L}$ je Choleskiho faktor matice $\boldsymbol{C}_{jm}$. Ekvivalentně lze vztah (3.11) zapsat ve tvaru

$$\bar{\boldsymbol{C}}_{jm} = \boldsymbol{H}_{(n)}\boldsymbol{C}_{jm}\boldsymbol{H}_{(n)}^T. \tag{3.12}$$

Úloha nalezení lineárních transformačních matic je vázána na nalezením optima následu-

jící funkce:

$$Q(\lambda, \bar{\lambda}) = const - \frac{1}{2} \sum_{b_{jm} \in \lambda} \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}(t)(c_{jm} + \log|\bar{\boldsymbol{C}}_{jm}| + (\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})^T \bar{\boldsymbol{C}}_{jm}^{-1}(\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})). \tag{3.13}$$

Implementačně lze rozdělit úlohu na dvě části, nalezení transformací pro střední hodnoty (3.10) a pro kovarianční matice zvlášť, viz (3.11) nebo (3.12).

Metoda MLLR pro střední hodnoty (MLLRmean)

Naším úkolem je nalézt matici $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$, která transformuje střední hodnoty všech gaussovských složek b_{jm} patřících do shluku C_n, tedy maximalizovat optimalizační funkci (3.13) [Gan05]. Provedením derivace a vhodnou úpravou lze dostat vztah

$$\sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t) \boldsymbol{C}_{jm}^{-1} \boldsymbol{o}^e(t) \boldsymbol{\xi}_{jm}^T = \sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t) \boldsymbol{C}_{jm}^{-1} \boldsymbol{W}_{(n)} \boldsymbol{\xi}_{jm} \boldsymbol{\xi}_{jm}^T. \tag{3.14}$$

Výraz (3.14) lze pro zjednodušení intuitivně přepsat zavedením substituční matice $\boldsymbol{Z}_{(n)}$ za celou levou část rovnice a $\boldsymbol{V}_{jm}\boldsymbol{D}_{jm}$ za pravou část uvnitř sumy přes složky shluku C_n. Zredukovaný tvar rovnice je pak

$$\boldsymbol{Z}_{(n)} = \sum_{b_{jm} \in C_n} \boldsymbol{V}_{jm} \boldsymbol{W}_{(n)} \boldsymbol{D}_{jm}. \tag{3.15}$$

Řešení rovnice (3.15) je výpočetně náročné, proto se v praxi více využívá výpočet přes řádky matice $\boldsymbol{W}_{(n)}$, který předpokládá model s diagonálními kovariančními maticemi. Je-li matice $\boldsymbol{C}_{jm}$ diagonální (tedy vektor $\boldsymbol{\sigma}_{jm}^2 = diag(\boldsymbol{C}_{jm})$), pak je diagonální i matice $\boldsymbol{V}_{jm}$. i-tý řádek matice $\boldsymbol{W}_{(n)}$ lze pak spočítat pro všechna $i = 1, \ldots, I$ ze vztahu

$$\boldsymbol{w}_i^T = z_i^T G_i^{-1} \tag{3.16}$$

kde

$$\boldsymbol{G}_i = \sum_{b_{jm} \in C_n} \frac{1}{\sigma_{jm}(i)^2} \boldsymbol{\xi}_{jm} \boldsymbol{\xi}_{jm}^T \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_{jm}^e(t). \tag{3.17}$$

Nepatrně odlišné odvození výpočtu transformační matice $\boldsymbol{W}_{(n)}$ za předpokladu diagonálních kovariančních matic lze nalézt v [PS06]. Uvedeno je zde pro lepší návaznost na odvození vztahů pro omezenou transformaci fMLLR (Feature MLLR) v podkapitole 3.3.2. Odvození využívá vztahy (3.2), (3.3), (3.4) definované na začátku této kapitoly.

Část optimalizační funkce (3.13), která je závislá na $\boldsymbol{W}_{(n)}$, je:

$$Q_{\boldsymbol{W}_{(n)}} = const - \sum_{b_{jm} \in C_n} c_{jm} \sum_{i=1}^{I} \frac{(\boldsymbol{w}_{(n)i}^T \boldsymbol{\xi}_{jm})^2 - 2(\boldsymbol{w}_{(n)i}^T \boldsymbol{\xi}_{jm}) \boldsymbol{\varepsilon}_{jm}(o)(i)}{\sigma_{jm}^2(i)}. \tag{3.18}$$

Rovnice (3.18) může být dále přepsána na tvar:

$$Q_{\boldsymbol{W}_{(n)}} = \boldsymbol{w}_{(n)i}^T \boldsymbol{k}_{(n)i} - 0.5 \boldsymbol{w}_{(n)i}^T \boldsymbol{G}_{(n)i} \boldsymbol{w}_{(n)i}, \tag{3.19}$$

kde

$$\boldsymbol{k}_{(n)i} = \sum_{b_{jm} \in C_n} \frac{c_{jm} \boldsymbol{\xi}_{jm} \boldsymbol{\varepsilon_{jm}(o)}(i)}{\sigma^2_{jm}(i)} \tag{3.20}$$

a

$$\boldsymbol{G}_{(n)i} = \sum_{b_{jm} \in C_n} \frac{c_{jm} \boldsymbol{\xi}_{jm} \boldsymbol{\xi}^T_{jm}}{\sigma^2_{jm}(i)}, \tag{3.21}$$

Pak maximalizováním rovnice (3.19) dostáváme:

$$\boldsymbol{w}_{(n)i} = \boldsymbol{G}^{-1}_{(n)i} \boldsymbol{k}_{(n)i}. \tag{3.22}$$

Metoda MLLR pro kovarianční matice (MLLRcov)

Tato metoda [Gan05] se počítá ve dvou krocích, nejprve transformujeme střední hodnoty (stejný postup jako u metody MLLRmean), poté kovarianční matice. Postupně získáváme modely: $\lambda = \{\boldsymbol{\mu}, \boldsymbol{C}\}, \bar{\lambda} = \{\bar{\boldsymbol{\mu}}, \boldsymbol{C}\}, \tilde{\lambda} = \{\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{C}}\}$ a platí pro ně: $p(\boldsymbol{O}|\lambda) \leq p(\boldsymbol{O}|\bar{\lambda}) \leq p(\boldsymbol{O}|\tilde{\lambda})$. Jak již bylo zmíněno, lze transformaci kovarianční matice spočítat dvěma způsoby. První vychází z rovnice (3.11), kde $\boldsymbol{L}$ je získáno Choleskiho rozkladem matice $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^T$. Pak nejlepší odhad transformační matice $\boldsymbol{H}_{(n)}$ lze získat [Gal97]

$$\boldsymbol{H}_{(n)} = \frac{\sum_{b_{jm} \in C_n} \left((\boldsymbol{L}^{-1}_{jm})^T \left[\sum_{e=1}^{E} \sum_{t=1}^{T_e} y^e_{jm}(t) (\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})(\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm}) \right] \boldsymbol{L}^{-1}_{jm} \right)}{\sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} y^e_{jm}(t)}. \tag{3.23}$$

Rozpoznávání s takto adaptovaným modelem je značně výpočetně náročné (pokud uvažujeme plné kovarianční matice), protože logaritmus věrohodnosti $\mathcal{L}$ vektoru pozorování $\boldsymbol{o}(t)$ daný transformovaným modelem $\tilde{\lambda}$ je počítán jako logaritmus normálního rozdělení $\mathcal{N}$

$$\log \mathcal{L}(\boldsymbol{o}^e(t), \boldsymbol{\mu}, \boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H}) = \log \mathcal{N}(\boldsymbol{o}^e(t), \bar{\boldsymbol{\mu}}, \bar{\boldsymbol{C}}), \tag{3.24}$$

kde $\bar{\boldsymbol{C}}$ bude nadále plná kovarianční matice, W a H jsou transformační funkce získané při adaptaci MLLRmean resp. MLLRcov.

Pokud však předpokládáme původní kovarianční matice modelu diagonální, je efektivnější vycházet ze vztahu (3.12) a počítat transformační matici po řádcích. Vektor $\boldsymbol{\sigma}^2_{jm} = diag(\boldsymbol{C}_{jm})$ je diagonála kovarianční matice $\boldsymbol{C}_{jm}$. i-tý řádek transformační matice $\boldsymbol{H}(n)$, tedy $\boldsymbol{h}_{(n)i}$, lze iterativně vypočítat jako:

$$\boldsymbol{h}^{-1}_{(n)i} = \boldsymbol{v}_{(n)i} \boldsymbol{G}^{-1}_{(n)i} \sqrt{\frac{\sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma^e_{jm}(t)}{\boldsymbol{v}_{(n)i} \boldsymbol{G}^{-1}_{(n)i} \boldsymbol{C}^T_{jm}(i)}}, \tag{3.25}$$

kde

$$\boldsymbol{G}_{(n)i} = \sum_{b_{jm} \in C_n} \frac{1}{\sigma^2_{jm}(i)} \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma^e_{jm}(t) (\boldsymbol{o}^e(t) - \boldsymbol{\mu}_{jm})(\boldsymbol{o}^e(t) - \boldsymbol{\mu}_{jm})^T \tag{3.26}$$

a $\boldsymbol{v}_{(n)i}$ je kofaktor matice $\boldsymbol{H}^{-1}_{(n)}$.

Alternativní výpočet výsledného logaritmus věrohodnosti $\mathcal{L}$ pro konkrétní Gausian ze třídy C_n může být nyní počítán

$$\log \mathcal{L}(\boldsymbol{o}^e(t)|\boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}, \boldsymbol{W}_{(n)}, \boldsymbol{H}_{(n)}) = \log \mathcal{N}(\boldsymbol{H}_{(n)}^{-1}\boldsymbol{o}^e(t); \boldsymbol{H}_{(n)}^{-1}\bar{\boldsymbol{\mu}}_{jm}, \boldsymbol{C}_{jm}) - 0.5 \log(|\boldsymbol{H}_{(n)}|^2), \tag{3.27}$$

Druhý z uvedených vztahů je méně výpočetně náročný, protože není zapotřebí inverze matice $\boldsymbol{H}$ ani dvojitého násobení kovarianční matice $\boldsymbol{C}$ transformační maticí $\boldsymbol{H}$, viz (3.11) .

3.3.2 Metoda MLLR pro transformace vektorů pozorování (fMLLR)

Metoda **maximální věrohodné lineární regrese vektorů pozorování** (fMLLR – feature Maximum Likelihood Linear Regression) [Gal97] je zaměřena na lineární transformaci vektoru příznaků $\boldsymbol{O}$, spíše než na transformaci samotného akustického modelu. To přináší výhody převážně v rychlosti adaptace (není potřeba transformovat rozsáhlý model s tisícem příznaků) a v paměťové náročnosti (pamatujeme si pouze transformaci, nikoliv celý nový model pro každého z řečníků). Transformace modelu metodou fMLLR je však v zásadě možná pouhým přepisem transformačních vztahů do jiné formy (viz níže), pak je metoda nazývána **omezenou MLLR** (CMLLR – Constrained Maximum Linear Regression). Z principu metody je zřejmé, že jde o omezenou transformaci, tedy střední hodnoty a kovarianční matice jsou transformovány stejnou transformací

$$\bar{\boldsymbol{o}}^e(t) = \boldsymbol{A}_{(n)}\boldsymbol{o}^e(t) + \boldsymbol{b}_{(n)} = \boldsymbol{A}_{(n)c}^{-1}\boldsymbol{o}^e(t) + \boldsymbol{A}_{(n)c}^{-1}\boldsymbol{b}_{(n)c} = \boldsymbol{W}_{(n)}\boldsymbol{\xi}^e(t), \tag{3.28}$$

kde $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$ je transformační matice, $\boldsymbol{\xi}^{eT}(t) = [\boldsymbol{o}^{eT}(t), 1]$ je rozšířený vektor příznaků a $\boldsymbol{A}_{(n)c}, \boldsymbol{b}_{(n)c}$ jsou matice pro ekvivalentní transformaci parametrů akustického modelu

$$\bar{\boldsymbol{\mu}}_{jm} = \boldsymbol{A}_{(n)c}\boldsymbol{\mu}_{jm} - \boldsymbol{b}_{(n)c}, \tag{3.29}$$

a

$$\bar{\boldsymbol{C}}_{jm} = \boldsymbol{A}_{(n)c}\boldsymbol{C}_{jm}\boldsymbol{A}_{(n)c}^T, \tag{3.30}$$

Optimalizační funkce pro odhad transformací nabývá tvaru:

$$Q(\lambda, \bar{\lambda}) = const - \frac{1}{2}\sum_{b_{jm}\in\lambda}\sum_{t,e=1}^{T_E}\gamma_{jm}^e(t)(c_{jm} + \log|\boldsymbol{C}_{jm}| - \log(|\boldsymbol{A}_{(n)}|^2) + (\bar{\boldsymbol{o}}^e(t) - \boldsymbol{\mu}_{jm})^T C_{jm}^{-1}(\bar{\boldsymbol{o}}^e(t) - \boldsymbol{\mu}_{jm})). \tag{3.31}$$

Analogicky jako v odvození pro metodu MLLRmean 3.3.1 lze optimalizační funkci (3.31) upravit na tvar [PS06]

$$Q_{\boldsymbol{W}_{(n)}}(\lambda, \bar{\lambda}) = \log(|\boldsymbol{A}_{(n)}|) - \sum_{i=1}^{I}\boldsymbol{w}_{(n)i}^T\boldsymbol{k}_i - 0.5\boldsymbol{w}_{(n)i}^T\boldsymbol{G}_{(n)i}\boldsymbol{w}_{(n)i}, \tag{3.32}$$

kde

$$\boldsymbol{k}_{(n)i} = \sum_{jm\in C_n}\frac{c_{jm}\boldsymbol{\mu}_{jm}(i)\boldsymbol{\varepsilon}(\boldsymbol{\xi})_{jm}}{\sigma_{jm}^2(i)}, \tag{3.33}$$

$$\boldsymbol{G}_{(n)i} = \sum_{jm\in C_n}\frac{c_{jm}\boldsymbol{\varepsilon}(\boldsymbol{\xi}\boldsymbol{\xi}^T)_{jm}}{\sigma_{jm}^2(i)}, \tag{3.34}$$

$$\boldsymbol{\varepsilon}(\boldsymbol{\xi})_{jm} = [\boldsymbol{\varepsilon}(\boldsymbol{o})_{jm}; 1], \tag{3.35}$$

a

$$\boldsymbol{\varepsilon}(\boldsymbol{\xi}\boldsymbol{\xi}^T)_{jm} = \begin{bmatrix} \boldsymbol{\varepsilon}(\boldsymbol{o}\boldsymbol{o}^T)_{jm} & \boldsymbol{\varepsilon}(\boldsymbol{o})_{jm} \\ \boldsymbol{\varepsilon}(\boldsymbol{o})_{jm}^T & 1 \end{bmatrix}. \tag{3.36}$$

Pro nalezení řešení rovnice (3.32) musíme vyjádřit matici $\boldsymbol{A}_{(n)}$ ve tvaru $\boldsymbol{W}_{(n)}$. Je možné matematicky dokázat, že $\log(|\boldsymbol{A}|) = \log(|\boldsymbol{w}_i^T \boldsymbol{v}_i|)$, kde $\boldsymbol{v}_i$ je kofaktor matice $\boldsymbol{A}_{(n)}$ rozšířený o nulu v poslední dimenzi. Maximalizováním funkce (3.32) dostáváme řešení:

$$\boldsymbol{w}_{(n)i} = \boldsymbol{G}_{(n)i}^{-1}\left(\frac{\boldsymbol{v}_{(n)i}}{f} + \boldsymbol{k}_{(n)i}\right). \tag{3.37}$$

Lze dokázat, že $f_{1,2}$ je řešením kvadratické rovnice, jejíž koeficienty jsou

$$[a, b, c] = [\beta_{(n)}, -\boldsymbol{c}_{(n)i}^T \boldsymbol{G}_{(n)i}^{-1} \boldsymbol{k}_{(n)i}, -\boldsymbol{v}_{(n)i}^T \boldsymbol{G}_{(n)i}^{-1} \boldsymbol{v}_{(n)i}], \tag{3.38}$$

$$\beta_{(n)} = \sum_{jm \in C_n} \sum_t \gamma_{jm}^e(t). \tag{3.39}$$

Po dosazení vypočteného $f_{1,2}$ do rovnice (3.37) dostáváme dvě řešení $\boldsymbol{w}_{(n)i}^{1,2}$. Vybíráme takové, které maximalizuje pomocnou funkci (3.32).

Následně můžeme spočítat logaritmus pravděpodobnosti pro metodou **CMLLR** jako:

$$\log \mathcal{L}(\boldsymbol{o}^e(t)|\boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}, \boldsymbol{A}_{(n)c}, \boldsymbol{b}_{(n)c}) = \log \mathcal{N}(\boldsymbol{o}^e(t); \boldsymbol{A}_{(n)c}\boldsymbol{\mu}_{jm} - \boldsymbol{b}_{(n)c}, \boldsymbol{A}_{(n)c}\boldsymbol{C}_{jm}\boldsymbol{A}_{(n)c}^T), \tag{3.40}$$

nebo pro metodu **fMLLR** jako:

$$\log \mathcal{L}(\boldsymbol{o}^e(t)|\boldsymbol{\mu}_{\boldsymbol{jm}}, \boldsymbol{C}_{jm}, \boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}) = \log \mathcal{N}(\boldsymbol{A}_{(n)}\boldsymbol{o}^e(t) + \boldsymbol{b}_{(n)}; \boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}) + 0.5 \log(|\boldsymbol{A}_{(n)}|^2). \tag{3.41}$$

Odhad matice $\boldsymbol{W}_{(n)} = [\boldsymbol{A}_{(n)}, \boldsymbol{b}_{(n)}]$ je iterativní procedura, naším úkolem je tedy na začátku inicializovat matice $\boldsymbol{A}_{(n)}$ a $\boldsymbol{b}_{(n)}$. Matice $\boldsymbol{A}_{(n)}$ je inicializována jako diagonální matice s jednotkovou diagonálou a vektor $\boldsymbol{b}_{(n)}$ je volen jako nulový. Iterace skončí tehdy, když změna v parametrech transformační matice $\boldsymbol{W}_{(n)}$ je zanedbatelná (obvykle po 20 iteracích).

3.3.3 Diskriminativní lineární transformace (DLT)

U metody **diskriminativní lineární transformace** (DLT – Discriminative Linear Transformation) je, stejně jako v metodě DMAP 3.2.1, ML kritérium (2.14) nahrazeno některým z diskriminativních kritérií pro trénování (viz 2.3.3). Například v práci [UW01] je využito MMI kritérium (2.32) a tzv. **H-kriteriální funkce** (H-Criterion)

$$(\alpha - 1)\mathcal{F}_{ML}(\lambda) - \mathcal{F}_{MMI}(\lambda), \tag{3.42}$$

kde uživatelsky volitelný parametr $\alpha \geq 1$ zajistí kombinaci kritérií MMI a ML. Kriteriální funkce (3.42) lze dle [UW01] přepsat do tvaru

$$\begin{aligned} &\sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} (\alpha\gamma_{jm}^{num}(t) - \gamma_{jm}^{den}(t)) \boldsymbol{C}_{jm}^{-1} \boldsymbol{o}^e(t) \boldsymbol{\xi}_{jm}^T = \\ = &\sum_{b_{jm} \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{T_e} (\alpha\gamma_{jm}^{num}(t) - \gamma_{jm}^{den}(t)) \boldsymbol{C}_{jm}^{-1} \boldsymbol{W}(n) \boldsymbol{\xi}_{jm} \boldsymbol{\xi}_{jm}^T, \end{aligned} \tag{3.43}$$

kde $\boldsymbol{\xi}_{jm}$ je rozšířený vektor střední hodnoty j-tého stavu HMM m-té složky GMM, $\boldsymbol{o}^e(t)$ je vektor pozorování a $\gamma^e_{jm}(t)$ je aposteriorní pravděpodobnost, že pozorování $\boldsymbol{o}^e(t)$ je generováno m-tou složkou j-tého stavu HMM. Horní index *num* nebo *den* označuje čitatel nebo jmenovatel zlomku (2.32), který je použit k vypočítání hodnoty $\gamma^e_{jm}(t)$. Rovnice (3.43) je formálně shodná s rovnicí (3.14) pro výpočet MLLR transformací, jen $\gamma^e_{jm}(t)$ je zde nahrazeno $(\alpha\gamma^{num}_{jm}(t) - \gamma^{den}_{jm}(t))$. Stejný postup může být aplikován i pro odvození transformací pro kovarianční matice.

Stejná myšlenka je rozvíjena i v práci [TDB05] zabývající se **diskriminativní lineární transformací pro vektory pozorování**. Opět je zde nahrazeno ML MMI kritériem. Podobný přístup k DLT založený na MPE kritérui (2.33) je popsán v [WW04].

3.3.4 Shlukování podobných parametrů modelu

Výhodou metod založených na lineární regresi (jako je např. metoda MLLR nebo fMLLR) je možnost nashlukování podobných parametrů modelu (jednotlivé směsi GMM definované střední hodnotou a kovarianční maticí) dle potřeby a množství adaptačních dat. Všechny parametry patřící do jednoho shluku jsou transformovány stejnou transformací. Počet shluků záleží na množství adaptačních dat. Před výběrem shlukovací metody je třeba si položit dvě otázky:

- jak vhodně nashlukovat parametry do jedné třídy, aby pro ně mohla být použita stejná transformace a
- kolik transformací je potřeba pro dané množství adaptačních dat.

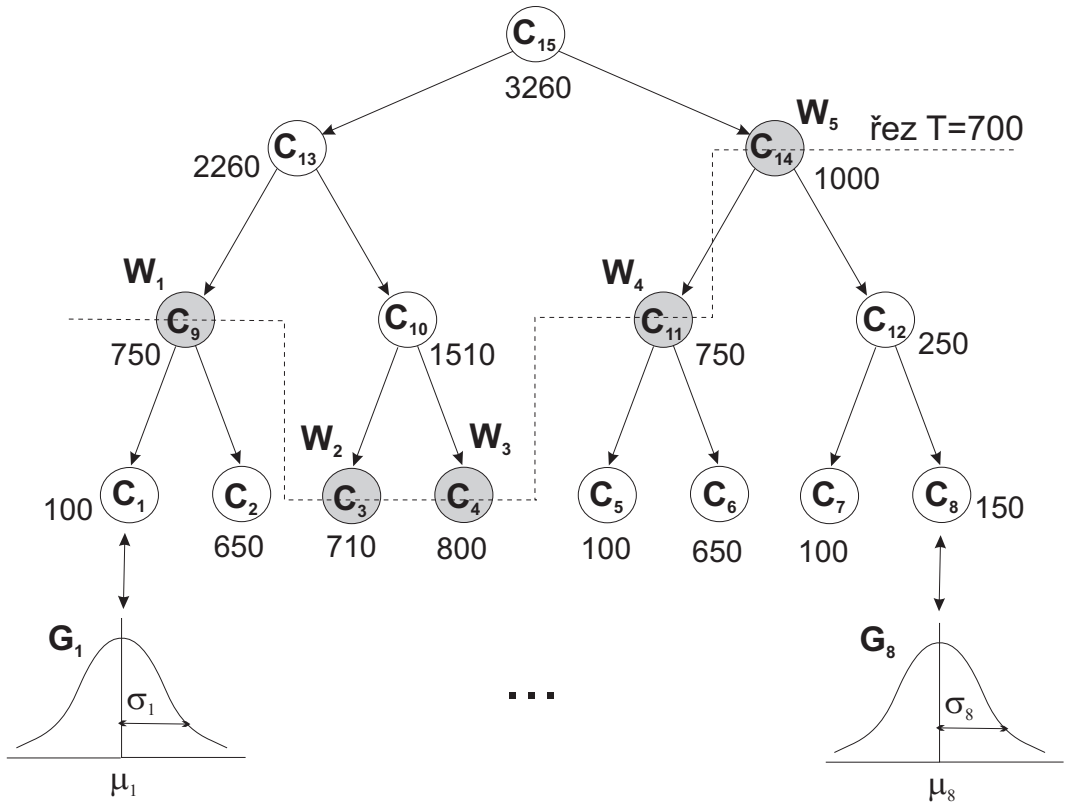
Vlastní shluky mohou být vytvořeny a zafixovány před adaptací, pak se jedná o **zafixované regresní třídy**. Pro zajištění flexibility a robustnosti shlukování bylo v článku [LW95] navrženo použití **regresního stromu** pro hierarchické shlukování parametrů modelu do regresních tříd.

Regresní strom je obvykle binárním stromem, kde každý uzel stromu reprezentuje jeden shluk $C_i, i = 1, \ldots, I$, parametrů modelu. Ke každé třídě může být přiřazena transformace $\boldsymbol{W}_{(n)}, n = 1, \ldots, N$, (obvykle je $N < I$, protože se budou počítat pouze ty transformace, pro které je dostatečný počet aktuálních adaptačních dat). Kořenový uzel obsahuje všechny parametry (složky GMM) celého modelu a každý finální list regresního stromu obsahuje pouze jednu konkrétní složku $G_m, m = 1, \ldots, M$, kde v tomto případě M určuje počet všech komponent všech stavů akustického modelu.

Regresní strom je využit jako apriorní informace o všech možných variantách shlukování v prostoru parametrů modelu. Podle množství a typu adaptačních dat je vybráno vhodné rozdělení prostoru parametrů podle "řezu" (viz příklad na obrázku 3.3) v regresním stromě.

Během adaptačního procesu jsou adaptační data rozdělena příslušným Gausovským komponentám (parametrům) modelu a je akumulována tzv. "okupace" (obsazení daty) jednotlivých tříd regresního stromu

$$\beta_{(n)} = \sum_{jm \in C_n} c_{jm} = \sum_{jm \in C_n} \sum_{e=1}^{E} \sum_{t=1}^{Te} \gamma^e_{jm}(t), \tag{3.44}$$



Obrázek 3.3: Příklad binárního regresního stromu. C_1 až C_{15} označují jednotlivé uzly, resp. třídy parametrů k nim náležící. Čísla u uzlů značí jejich aktuální obsazení adaptačními daty, šedivě jsou podbarveny ty uzly, které tvoří takzvaný řez stromu, hladinu s dostatečně velkou okupací (větší než práh $T = 700$). Pro tyto uzly jsou vypočítány transformace $\boldsymbol{W}_1$ až $\boldsymbol{W}_5$. Např. pro třídu C_{12} neexistuje dostatečné množství dat (její okupace adaptačními daty je 250), naopak její rodičovská třída C_{14} má již dostatek pozorování (okupace = 1000) na to, aby pro ni mohla být vypočtena transformace $\boldsymbol{W}_5$. Všechny parametry, které obsahuje třída $C_{14} = C_{11} \cup C_{12}$ jsou použity pro výpočet transformace $\boldsymbol{W}_5$, avšak pouze parametry z třídy C_{12} budou touto transformací adaptovány, protože třída C_{11} má dostatek pozorování pro výpočet své vlastní transformace $\boldsymbol{W}_4$.

kde $\gamma^e_{jm}(t)$ je aposteriorní pravděpodobnost, že pozorování $\boldsymbol{o}^e(t)$ je generováno m-tou komponentou j-tého stavu HMM viz rovnice (3.2). Strom je procházen ze zdola nahoru a jsou generovány transformace pouze pro ty uzly stromu, které dosáhnou předem definované úrovně okupace (jejich obsazení daty je větší jak předem definovaný práh T).

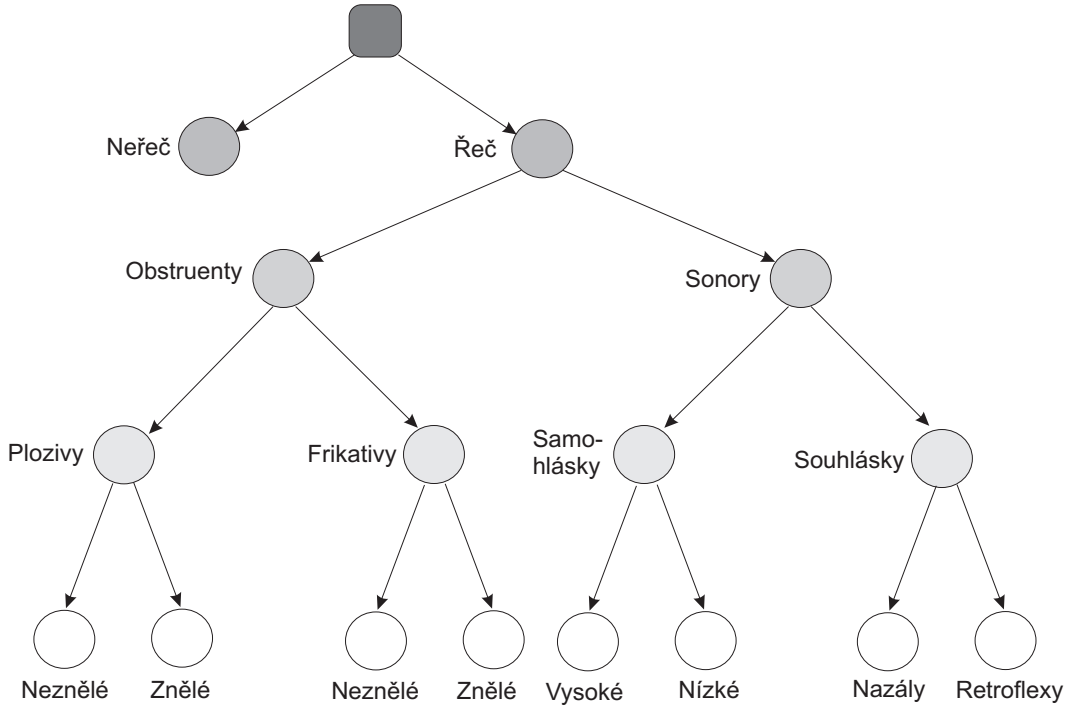
Vytváření regresního stromu

Obvykle dělíme regresní stromy do dvou kategorií, podle informace kterou využívají pro shlukování parametrů.

- **Fonetická znalost**. Existují určité expertní znalosti o podobnosti jednotlivých akustických elementů (fonémů), které jsou využity při tvorbě regresního stromu. Příkladem může být fonetický strom na obrázku 3.4.

- **Akustický prostor**. Parametry modelu jsou shlukovány podle vzájemné blízkosti v akustickém prostoru. Tato metoda využívá výhod "data-driven" přístupu, tím nevyžaduje

expertní znalost (viz příklad na obrázku 3.3). Dále se v textu budeme věnovat právě tomuto přístupu.



Obrázek 3.4: Příklad fonetického stromu. Na obrázku je fonetické dělení navrhnuto pro angličtinu, převzato z [SKFS07].

Optimální rozdělení akustického prostoru na shluky podle [Gal96] vychází z kritéria

$$\hat{Tree} = \arg\max_{T} ree \sum_{s=1}^{S} Q(M, \bar{M}|Tree), \tag{3.45}$$

$$\begin{aligned} Q(M, \bar{M}|Tree) = cost - \frac{1}{2}\log \mathcal{L}(O|M) \sum_{b_{jm} \in M} \sum_{e=1}^{E} \sum_{t=1}^{T_e} K_{jm}(t)[const_{jm} + \\ + log(|\bar{\boldsymbol{C}}_{jm}|) + (\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})^T ree \bar{\boldsymbol{C}}_{jm}^{-1}(\boldsymbol{o}^e(t) - \bar{\boldsymbol{\mu}}_{jm})], \end{aligned} \tag{3.46}$$

kde $Tree$ značí regresní strom, M je původní SI model a $\bar{M}$ je nový SA model s parametry $\bar{\boldsymbol{\mu}}_{jm}$ a $\bar{\boldsymbol{C}}_{jm}$. Není však možné garantovat dosažení globálního optima, pouze každé dělení stromu nalezne lokální optimum.

Podle [YEG+06] je shlukování prováděno dle středních hodnot a jejich blízkost je dána Eukleidovskou mírou. Konstrukce regresního stromu je prováděna rozdělováním shluků, obvykle se končí v předem definované úrovni stromu. Nepokračuje se tedy až do konečného rozdělení, kde by každému listu odpovídala jedna komponenta GMM.

Jiný přístup shlukování přináší [CXWF06], kde tvorba stromu je rozdělena do dvou kroků. V prvním kroku jsou parametry modelu iterativně rozdělovány od vrcholu dolů použitím divizní hierarchické strategie založené na **Bayesově informačním kritériu** (BIC – Bayes Information Criterion) [FR98], které automaticky odvodí optimální počet finálních tříd.

Ve druhém kroku jsou pak finální třídy z prvního kroku iterativně spojovány ze zdola nahoru k vytvoření regresního stromu aglomerativní strategií (blízkost shluků je opět dána BIC kritériem). Výhodou tohoto přístupu je jeho plná automatizace, tedy není při něm potřeba žádné vnější informace (jako je znalost finálního počtu listů).

3.4 Kombinace přístupu MAP a MLLR

Výhodou metody MAP je fakt, že při dostatečném množství dat SA model konverguje k SD modelu. Naopak výhodou metody MLLR je její dobrá účinnost i při malém počtu adaptačních dat díky shlukování podobných složek hustotních směsí a tím snižování počtu volných parametrů modelu. Nabízí se možnost výše zmíněné metody zkombinovat dohromady.

3.4.1 Dvoukroková adaptace

Jednou z možností, která se intuitivně jeví jako nejjednodušší, je adaptace modelu ve dvou krocích [MZ07]. S ohledem na princip metody MLLR a MAP je výhodný následující postup:

- 1.krok - Adaptovat SI model pomocí MAP adaptace, získáme SA_1 model. Metoda MAP provede adaptaci jednotlivých složek, pro které máme dostatečné množství dat.
- 2.krok - Adaptovat SA_1 model pomoci MLLR adaptace, získáme SA model. Pokud není dostatek dat pro adaptaci každého parametru, budou se pomocí MLLR adaptovat společně parametry nashlukované regresním stromem.

3.4.2 Regresní predikce modelu (RMP)

Při malém množství dat je metoda MAP neúčinná, protože dochází k adaptaci pouze těch parametrů, pro které se vyskytují adaptační data. Z toho důvodu byla do této metody zakomponována myšlenka shlukování podobných parametrů modelu, převzatá z metody MLLR. Výsledná metoda se nazývá **regresní predikce modelu** (RMP – Regression-based Model Prediction) [AW97]. Tato metoda používá malé množství tzv. **zdrojových parametrů**, pro které je dostatečné množství dat, a ty pak využívá k predikci adaptované hodnoty tzv. **cílových parametrů**, které jsou adaptačními daty špatně podmíněné. Pokud předpokládáme lineární vztah mezi zdrojovým parametrem a jeho cílovou skupinou parametrů, lze pak použít lineární regresi k odvození vztahů mezi těmito parametry. Například pro dva parametry x a y lze zapsat lineární vztah:

$$y = b_1 x + b_0 + \epsilon, \tag{3.47}$$

kde ϵ označuje chybu aproximace a b_1, b_0 jsou regresní parametry, které lze nalézt např. aplikováním **metody nejmenších čtverců** (LSE – Least Square Error)

$$\arg\min_{b_1,b_0} \sum_{k=1}^{K} \epsilon_k^2 = \arg\min_{b_1,b_0} \sum_{k=1}^{K} (y - b_1 x_k - b_0)^2, \tag{3.48}$$

kde K je konečný počet regresních bodů.

3.4.3 Regrese vážených sousedů (WMR)

Metoda **regrese vážených sousedů** (WMR – Weighted Neighbor Regression) [HWFW00] je založena na výše zmíněné technice RMP. Pokud uvažujeme adaptaci pouze středních hodnot $\boldsymbol{\mu}_{jm}$ na novou hodnotu $\bar{\boldsymbol{\mu}}_{jm}$, lze napsat regresní model ve tvaru:

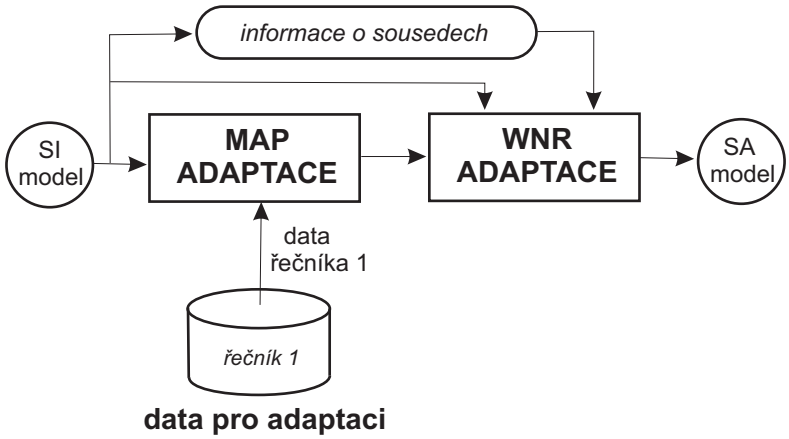
$$\bar{\boldsymbol{\mu}}_{jm} = \boldsymbol{B}\boldsymbol{\mu}_{jm} + \boldsymbol{b}_0 + \boldsymbol{\epsilon}_{jm}. \tag{3.49}$$

Metodou vážených nejmenších čtverců lze nalézt hodnoty regresních transformací minimalizováním vztahu

$$\sum_{k=1}^{K} w_k \boldsymbol{\epsilon}_k^2 = \sum_{k=1}^{K} w_k (\bar{\boldsymbol{\mu}}_k - \boldsymbol{B}\boldsymbol{\mu}_k - \boldsymbol{b}_0)^T (\bar{\boldsymbol{\mu}}_k - \boldsymbol{B}\boldsymbol{\mu}_k - \boldsymbol{b}_0), \tag{3.50}$$

kde všech K středních hodnot $\boldsymbol{\mu}_k$ (parametrů modelu) patří do jedné množiny vzájemně nejbližších (sousedních) středních hodnot $\boldsymbol{\mu}_{jm}$. w_k je váha k-tého parametru v dané množině, která je nepřímo úměrná Mahalanobisově vzdálenosti k-tého parametru od středu množiny.

Postup metody je následovný [Čer07]: Pro každý parametr modelu SI je pomocí Mahalanobisovy vzdálenosti nalezeno K nejbližších parametrů. Po MAP adaptaci jsou všechny komponenty rozděleny podle množství adaptačních dat k nim přidruženým na zdrojové (pro ně bylo k dispozici dostatečné množství dat) a cílové (s malým množstvím adaptačních dat). Pro každý zdrojový parametr a jeho přidruženou množinu sousedů je vypočítána regresní přímka. S její pomocí jsou adaptovány cílové parametry sousedící s daným zdrojovým parametrem (viz obrázek 3.5).



Obrázek 3.5: Blokový diagram WNR adapatace převzatý z [HWFW00].

3.4.4 Strukturální MAP

Metoda **strukturální MAP** (SMAP – Structural Maximum A Posteriori) [SL97] využívá hierarchickou strukturu v prostoru parametrů modelu (jako je regresní strom v podkapitole 3.3.4). Metoda odvozuje transformaci pro každou úroveň této hierarchické struktury. Parametry v konkrétní úrovni jsou použité i pro další své podúrovně. Výsledná transformace

parametrů je tedy kombinací transformací vyšších úrovní. Pomocí metody ML lze odhadnout transformace $\boldsymbol{A}_{jm}$ a $\boldsymbol{B}_{jm}$ pro každý uzel binárního dělícího stromu, pak lze střední hodnoty a kovarianční matice transformovat vztahy

$$\bar{\boldsymbol{\mu}}_{jm} = \boldsymbol{\mu}_{jm} + \boldsymbol{B}_{jm}, \tag{3.51}$$

$$\bar{\boldsymbol{C}}_{jm} = \boldsymbol{A}_{jm}\boldsymbol{C}_{jm}. \tag{3.52}$$

Ekvivalentní metody **strukturální MAP s lineární regresí** (SMAPLR – Structural Maximum A Posteriori Linear Regression) lze nalézt v [MSC00] nebo **vážené strukturální MAP** (WSMAP – Weighted Structural Maximum A Posteriori) v [JWJY03].

3.5 Shlukování mluvčích (SC)

Adaptační strategie popsaná v této části, metoda **shlukování mluvčích** (SC – Speaker Clustering) [PBNP98], je založena na hledání podmnožiny řečníků z trénovací množiny, kteří jsou akusticky blízko k testovanému řečníku. K přepočítání parametrů modelu jsou s výhodou použita data od nejbližších řečníků (apriorní znalost), než celá kompletní trénovací databáze obsahující promluvy od velkého množství řečníků. Nový model má pak mnohem blíže k testovacím datům než původní SI model. Jednou z možných implementací tohoto přístupu je na pohlaví závislý model (GD – Gender Dependent).

Adaptace na testovaného řečníka probíhá v těchto krocích:

- 1. Vytvoření akustického modelu z celé trénovací databáze (SI model).
- 2. Vytvoření akustických modelů pro všechny řečníky vyskytující se v trénovací databázi. Pokud nemáme dostatečné množství dat pro natrénování SD modelu, lze použít některou z adaptačních metod ke konstrukci SA modelu, popřípadě natrénovat pouze **jednostavový gaussovský model** (GMM – Gaussian mixture model) a použít některou z metod identifikace řečníka.
- 3. Pro adaptační data od testovaného řečníka nalézt N nejbližších řečníků (výběr nejpravděpodobnějších modelů pro adaptační data). K rychlejšímu výběru nejlepších řečníků může být použit i regresní strom viz [SBD95].
- 4. Z trénovacích dat od nejbližších řečníků vytvořit adaptovaný model. Obvykle se adaptují jen střední hodnoty, přičemž zbylé parametry zůstávají shodné s SI modelem. Jednou z možností vytvoření nového modelu je kombinace vektorů středních hodnot nejbližších řečníků metodou MAP či ML [HCC02].

Metoda SC si vystačí s malým množstvím adaptačních dat, jejím cílem je najít si podobná data v trénovací množině (data od nejbližších řečníků) k testovanému řečníkovi. Výhodou také je, že adaptace modifikuje všechny parametry SI modelu, nejen ty, které byly obsazeny adaptačními daty testovaného řečníka.

3.6 Dekompozice vlastních hlasů (ED)

Další možností jak reprezentovat apriorní znalost pro adaptaci je **analýzou hlavních komponent** (PCA – Principal Component Analysis) pro získání **vlastních hlasů** (eigenvoices) [Wes99]. Chceme-li adaptovat určitou množinu parametrů v modelu (např. střední hodnoty), pak lze tyto parametry zformovat do tzv. supervektoru dimenze D. Z T supervektorů modelů trénovacích řečníků vytvoříme PCA transformaci z D dimenzionálního prostoru do prostoru s nižší dimenzí K popsaného vlastními hlasy $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{K-1}$, kde $K < T << D$. Tím omezíme prostor, ve které je hledán model nového řečníka.

Adaptovaný vektor středních hodnot $\bar{\boldsymbol{\mu}}_{jm}$ j-tého stavu m-té složky modelu je počítán jako lineární kombinace vlastních hlasů $\boldsymbol{e}_0, \ldots, \boldsymbol{e}_{K-1}$

$$\bar{\boldsymbol{\mu}} = [\bar{\boldsymbol{\mu}}_1, \ldots, \bar{\boldsymbol{\mu}}_{jm}, \ldots, \bar{\boldsymbol{\mu}}_{JM}]^T = \sum_{i=0}^{K-1} w_i \boldsymbol{e}_i. \tag{3.53}$$

Množinu vah $w_0, \ldots, w_{K-1}$ odvodíme maximalizací pomocné funkce

$$Q(\lambda, \bar{\lambda}) = -\frac{1}{2} P(\boldsymbol{O}|\lambda) \sum_{j}^{J} \sum_{m}^{M} \sum_{t}^{T} \gamma_{jm}(t)(nlog(2\pi) + log|\boldsymbol{C}_{jm}| + (o(t) - \bar{\boldsymbol{\mu}}_{jm})^T \boldsymbol{C}_{jm}^{-1}(o(t) - \bar{\boldsymbol{\mu}}_{jm})) \tag{3.54}$$

Tento proces adaptace uvedený v [KNJ+98] se nazývá **dekompozice vlastních hlasů** (ED – Eigenvoices Decomposition). Výhodou této metody adaptace je její dobrá účinnost při malém množství adaptačních dat, hodí se tedy především pro rychlou on-line adaptaci.

Kapitola 4

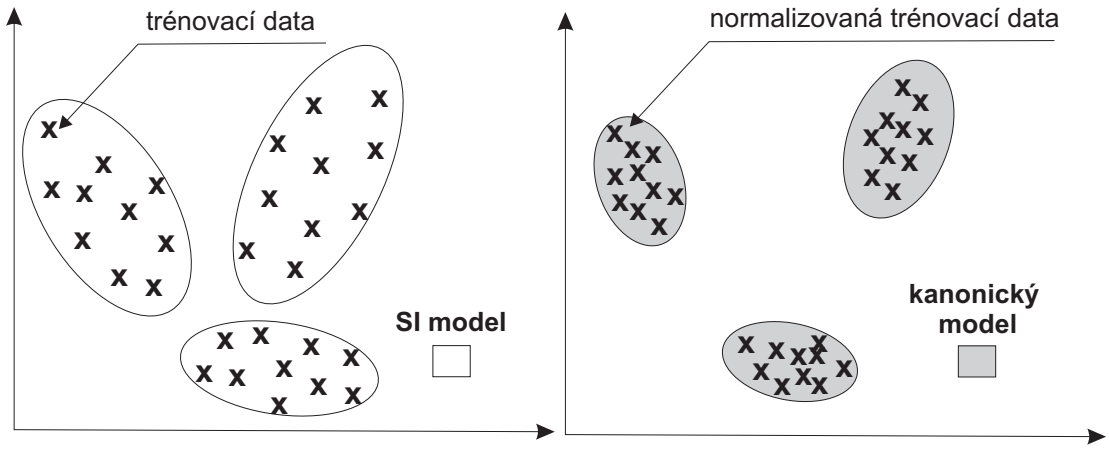
Adaptační techniky pro trénování

Tato kapitola popisuje adaptační techniky pro využití v trénovací fázi. Namísto adaptace SI modelu pomocí transformací vypočtených z dat dostupných v adaptační fázi, jsou tyto adaptační metody aplikovány na trénovací data, z kterých je pak vytvořen model bez rušivé informace o řečníkovi.

Postup je založen na hypotéze, že variabilita v akustickém modelu SI je způsobena jak fonetickou odlišností jednotlivých subslovních elementů (tuto informaci využíváme), tak rozdílem hlasových charakteristik mluvčích z trénovací databáze (které jsou pro rozpoznávání řeči rušivé) a vlivů prostředí, ve kterém byla trénovací data nahrána (aditivní či konvolutorní šum). Výsledkem je větší variabilita v trénovacích datech u SI modelu, než u SD modelu. Cílem adaptačních technik používaných při trénování modelu je odstranění právě této na řečníkovi a prostředí závislé nežádoucí informace. Metody se snaží snížit rozptyl trénovacích dat pro konkrétní subslovní jednotku a tím zajistit její lepší separovatelnost od ostatních jednotek. Na rozdíl od předchozí kapitoly 3, kdy byly charakteristiky modelu přizpůsobeny konkrétnímu řečníku, je zde vytvářen tzv. **kanonický model**, z něhož je informace o řečníkovi pomocí adaptačních technik odstraněna (viz obrázek 4.1).

Kanonický model reprezentuje veškerou požadovanou řečovou variabilitu celé trénovací databáze, ale je nezávislý na akustických podmínkách. Takovýto model je vytvářen jen tehdy, když máme k dispozici množinu transformací pro odstranění neřečové variability v datech. Tvar kanonického modelu závisí na formách adaptačních transformací. Pro lineární transformace jde o standardní HMM. Kanonický model je mnohem více kompaktní, je tedy nutné jej při vlastním rozpoznávání dále adaptovat na konkrétní testované akustické podmínky.

Možností, jak snížit variabilitu v trénovacích datech, je hned několik. Za zmínění stojí například **kepstrální normalizace** (CMN – Cepstrum Mean Normalization) [MIT97], která je jednoduchou a často používanou metodou k odstranění vlivu kanálu. Další z metod je **gaussionalizace** (Gaussianisation) viz [SDP04], normalizující kumulativní hustotní funkci vektorů pozorování na standardní Gaussovské rozložení. Sofistikovanější přístupy [Gal97] jsou **trénování s adaptací na mluvčího** (SAT – Speaker Adaptive Training) [AMSM96], [MSJN97], **trénování s adaptací pomocí shlukování mluvčích** (CAT – Cluster Adaptive Training) [Gal00] a **normalizace délky hlasového traktu** (VTLN – Vocal Tract Length Normalization) [ZW97], [LR96]. Výhodou zmíněných metod je, že se dají snadno a úspěšně kombinovat dohromady.



Obrázek 4.1: Ilustrativní příklad rozdílné variability složek modelu SI a kanonického modelu.

4.1 Trénování s adaptací na mluvčího (SAT)

Metoda **trénování s adaptací na mluvčího** (SAT – Speaker Adaptive Training) využívá lineárních transformací popsaných v podkapitole 3.3. Metoda se snaží odstranit variabilitu řečníků z fonetické informace a vytvořit kompaktní kanonický model λ_C, který informaci o řečníkovi neobsahuje. Zatímco klasická adaptace hledá model $\hat{\lambda}$, který by maximalizoval pravděpodobnost adaptačních dat od všech řečníků R

$$\hat{\lambda} = \arg\max_{\lambda} \prod_{r=1}^{R} P(\boldsymbol{O}^r|\lambda), \tag{4.1}$$

SAT počítá na řečníku r závislou transformaci H^r ke kanonickému modelu λ_C tak, aby se maximalizovala pravděpodobnost [Gan05]

$$(\hat{\lambda}_C, \hat{\boldsymbol{H}}) = \arg\max_{(\lambda_C, \boldsymbol{H})} \prod_{r=1}^{R} P(\boldsymbol{O}^r|\boldsymbol{H}^r(\lambda_C)), \tag{4.2}$$

tedy hledáme kanonický model $\hat{\lambda}_C$ a jeho transformaci $\hat{\boldsymbol{H}}^r$ závislou na řečníkovi r, které budou maximalizovat pravděpodobnost pro každého řečníka r zvlášť. Fonetická informace je uložena v kanonickém modelu $\hat{\lambda}_C$, informace od řečníka pak v transformaci $\hat{\boldsymbol{H}}^r$. Kanonický model, spolu s některou z adaptačních metod (viz předešlá kapitola 3) použitou při fázi rozpoznávání, zajistí výsledky lepší, než lze získat s původním SI modelem.

4.1.1 SAT pro MLLR

Při klasickém trénování akustického modelu se využívá EM algoritmus (viz podkapitola 2.3.1), který se snaží maximalizovat pomocnou funkci (2.20) vedoucí k odvození parametrů modelu, které zvyšují pravděpodobnost pro trénovací data.

V SAT přístupu [AMSM96] je naší snahou maximalizovat pomocnou funkci

$$Q(\rho, \hat{\rho}) = \sum_{r} \sum_{t} \sum_{jm} \gamma_{jm}^{r}(t) \mathcal{N}(\boldsymbol{o}^r(t); \hat{\boldsymbol{A}}^r \hat{\mu}_{jm} + \hat{\boldsymbol{b}}^r, \hat{\boldsymbol{C}}_{jm}), \tag{4.3}$$

kde parametr $\rho = (\boldsymbol{H}^r, \lambda_C) = ((\boldsymbol{A}_r, \boldsymbol{b}_r), (\hat{\boldsymbol{\mu}}_j m, \hat{\boldsymbol{C}}_j m))$ se skládá z transformace na řečníka a z kanonického modelu.

Pro zjednodušení výpočtu je maximalizace rozdělena iterativně na tři části. V každé z nich se snažíme optimalizovat pouze jeden z parametrů, zatímco zbylé dva zůstávají fixovány. V každé části optimalizačního procesu musí hodnota pomocné funkce Q růst:

$$Q(\boldsymbol{H}^r, \lambda_C) \leq Q(\hat{\boldsymbol{H}}^r, \lambda_C) \leq Q(\hat{\boldsymbol{H}}^r, (\hat{\boldsymbol{\mu}}, \boldsymbol{C})) \leq Q(\hat{\boldsymbol{H}}^r, \hat{\lambda}_C). \tag{4.4}$$

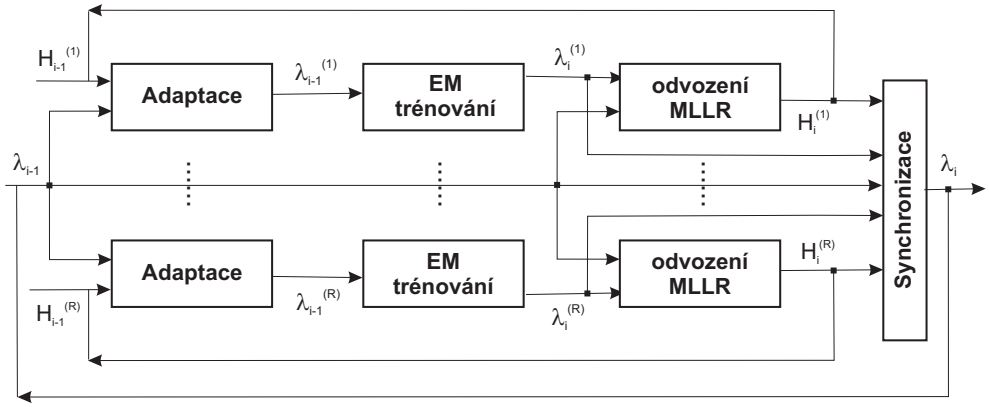
Konkrétně rovnice pro střední hodnoty a kovarianční matice kompaktního modelu lze zapsat ve formě

$$\hat{\boldsymbol{\mu}}_{jm} = \left(\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t) \hat{\boldsymbol{A}}^{rT} \hat{\boldsymbol{C}}_{jm}^{-1} \hat{\boldsymbol{A}}^r \right)^{-1} \sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t) \hat{\boldsymbol{A}}^{rT} \hat{\boldsymbol{C}}_{jm}^{-1} (\boldsymbol{o}^r(t) - \boldsymbol{b}^r), \tag{4.5}$$

$$\hat{\boldsymbol{C}}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t) (\boldsymbol{o}^r(t) - \hat{\boldsymbol{\mu}}_{jm}^r)(\boldsymbol{o}^r(t) - \hat{\boldsymbol{\mu}}_{jm}^r)^T}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t)}, \tag{4.6}$$

kde odhad transformace $\boldsymbol{H}^r$ je proveden pomocí standardní metody MLLRmean (viz podkapitola 3.3.1) a $\hat{\boldsymbol{\mu}}_{jm}^r = \boldsymbol{A}^r \hat{\boldsymbol{\mu}}_{jm} + \boldsymbol{b}^r$ je transformovaná střední hodnota kanonického modelu.

Re-estimační SAT proces je zobrazen na obrázku 4.2, kde celková zpětná vazba značí, že proces může být opakován, dokud model nedokonverguje do svého optima.



Obrázek 4.2: Blokový diagram pro metodu SAT založenou na MLLR transformacích. První blok zadaptuje model pomocí transformací $\boldsymbol{H_{i-1}}$, druhý blok odvodí nové parametry modelu λ_i (viz rovnice (4.5) a (4.6)), třetí blok pak spočte nové transformační matice $\boldsymbol{H_i}$ pomocí klasických adaptačních metod. Celý proces lze iterativně opakovat. Obrázek je převzat z práce [MSJN97].

Nevýhodou tohoto přístupu k trénování je značná paměťová náročnost [MSJN97], protože je potřeba uchovávat v paměti každou střední hodnotu a kovarianční matici kanonického modelu spolu s transformací, a to pro každého řečníka r zvlášť. S tím je též spojena časová náročnost díky I/O operací při práci s pamětí. Redukování náročnosti metody je navrženo v [MSJN97].

4.1.2 SAT pro fMLLR

Druhý přístup k SAT navržený v [Gal97] je založen na metodě fMLLR (viz podkapitola 3.3.2). Jeho výhodou oproti předchozí metodě je, že adaptační transformace jsou počítány

pro trénovací vektory pozorování. Tím je značně ušetřen čas a paměť pro ukládání mezivýsledků, protože přepočet středních hodnot a kovariančních matic probíhá v jednom optimalizačním kroku právě z výsledných transformovaných vektorů pozorování.

Pomocná funkce má tvar

$$Q(\rho, \hat{\rho}) = \sum_r \sum_t \sum_{jm} \gamma_{jm}^r(t) \mathcal{N}(\hat{\boldsymbol{o}}^r(t); \boldsymbol{\mu}_{jm}, \boldsymbol{C}_{jm}). \tag{4.7}$$

Dosadíme-li do této rovnice (4.7) vztah pro transformovaný vektor pozorování $\hat{\boldsymbol{o}}^r(t) = \boldsymbol{A}^r \boldsymbol{o}^r(t) + \boldsymbol{b}^r$ a s využitím rovnice (3.41) dostáváme

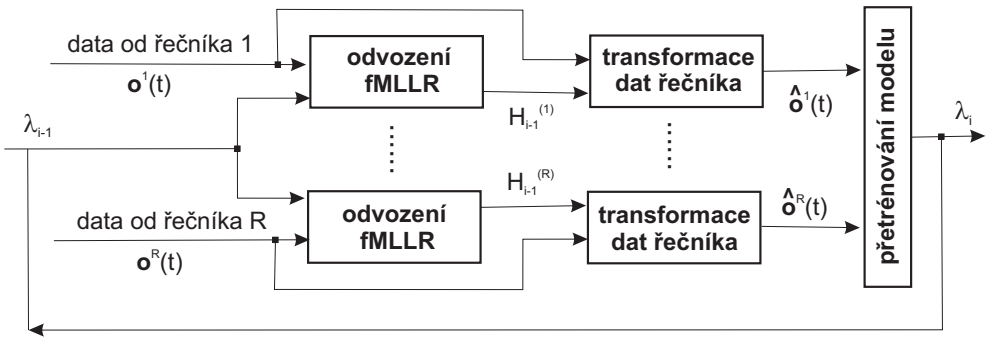
$$\begin{aligned} Q(\rho, \hat{\rho}) = c - \frac{1}{2} \sum_r \sum_t \sum_{jm} \gamma_{jm}^r(t)(c_{jm} + \\ + \log(|\hat{\boldsymbol{C}}_{jm}|) - log(|\boldsymbol{A}^{r2}|) + (\hat{\boldsymbol{o}}^r(t) - \hat{\boldsymbol{\mu}}_{jm})^T \hat{\boldsymbol{C}}_{jm}^{-1} (\hat{\boldsymbol{o}}^r(t) - \hat{\boldsymbol{\mu}}_{jm})). \end{aligned} \tag{4.8}$$

Transformační matice $\boldsymbol{H}^r = (\boldsymbol{A}^r, \boldsymbol{b}^r)$ jsou odvozeny adaptační metodou fMLLR (viz podkapitola 3.3.2) pro dané trénovací vektory pozorování od konkrétního řečníka. Střední hodnoty a kovarianční matice kanonického modelu lze poté přepočítat s využitím znalosti o transformacích vektorů pozorování v jednom kroku

$$\hat{\boldsymbol{\mu}}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t) \hat{\boldsymbol{o}}^r(t)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t)}, \tag{4.9}$$

$$\hat{\boldsymbol{C}}_{jm} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t) (\boldsymbol{o}^r(t) - \hat{\boldsymbol{\mu}}_{jm}^r)(\boldsymbol{o}^r(t) - \hat{\boldsymbol{\mu}}_{jm}^r)^T}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \gamma_{jm}^r(t)}, \tag{4.10}$$

Stejně jako u předchozí uvedené metody 4.1.1, i tento postup lze iterativně opakovat (viz obrázek 4.3).



Obrázek 4.3: Blokový diagram pro metodu SAT založenou na fMLLR transformacích. Nejprve se odvodí transformační matice $\boldsymbol{H_{i-1}}$ metodou fMLLR, kterými se zadaptuje vektor příznaků $\boldsymbol{o(t)}$. Z nových příznaků $\hat{\boldsymbol{o(t)}}$ se přetrénuje model λ_{i-1} (pomocí vztahů 4.9 a 4.10). Postup lze iterativně opakovat.

4.1.3 Diskriminativní adaptace pro trénování (DAT)

Metoda **diskriminativní adaptace pro trénování** (DAT – Discriminative Adaptation Training) je diskriminativní verzí SAT. Například v [TDB05] je odvozena metoda adaptačního

trénování vycházející z **diskriminativní lineární transformace pro vektory pozorování** 3.3.3 a MMI kritéria (2.32). Jiný přístup, uvedený v [WW03], používá kritérium MPE (2.33). V DAT, stejně jako v metodě SAT, je každá iterace rozdělena do dvou kroků, nejprve se odhadnou lineární transformace a poté parametry kanonického modelu. Diskriminativní kritérium (MMI popř. MPE) je používáno v obou krocích. Opět je s výhodou využívána omezená transformace vektorů pozorování, spíše než neomezená transformace vyžadující značné paměťové nároky.

4.2 Trénování s adaptací pomocí shlukování mluvčích (CAT)

Metoda **trénování s adaptací pomocí shlukování mluvčích** (CAT – cluster Adaptive Training) je jednoduchým rozšířením metody **shlukování mluvčích** (viz podkapitola 3.5). Poznamenejme, že máme trénovací data všech řečníků rozdělená do P shluků dle akustické blízkosti. Nad každým shlukem je vytvořen model (ať již trénováním nebo adaptací SI modelu). Množina $\mathcal{M}$ (4.14) těchto shlukových modelů nahrazuje jeden kanonický model používaný v metodě SAT (viz podkalpitola 4.1).

K vytvoření akustického modelu pomocí metody CAT [Yu06] je využit vektor interpolačních vah $\boldsymbol{\nu}^r$ pro kombinaci všech středních hodnot P shlukových modelů, obvykle sdružených do matice

$$\boldsymbol{M}_{jm} = [\mu^1_{jm}, \ldots, \mu^P_{jm}] \text{ pro } jm = 1, \ldots, M \ , \tag{4.11}$$

kde M je celkový počet všech složek všech stavů modelu, P je počet shluků. Metoda CAT se zaměřuje na adaptaci pouze středních hodnot akustického modelu $\boldsymbol{\mu}$, zbylé parametry shlukových kanonických modelů (pravděpodobnosti přechodů a a kovarianční matice $\boldsymbol{C}$) zůstávají nezměněny. Vektor vah

$$\boldsymbol{\nu}^r = [\nu^{1r}, \ldots, \nu^{Pr}], \tag{4.12}$$

hrající úlohu transformace, je počítán pro každé odlišné akustické podmínky $r = 1, \ldots, R$ (různý řečník či různé prostředí). Adaptovaná střední hodnota jm-té komponenty pro jednotlivé akustické podmínky r je dána vztahem

$$\boldsymbol{\mu}^r_{jm} = \boldsymbol{M}_{jm}\boldsymbol{\nu}^r. \tag{4.13}$$

4.2.1 Hledání parametrů modelu a transformací

Stejně jako v metodě SAT, i zde se pro trénování používá ML kritérium [Gal00]. Změna je pouze v kanonickém modelu, který je zde tvořen množinou středních hodnot jednotlivých shlukových modelů a kovariančních matic, které mají všechny shlukové modely stejné

$$\mathcal{M} = \{\{\boldsymbol{M}_1, \ldots, \boldsymbol{M}_M\}, \{\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M\}\}. \tag{4.14}$$

Pro odvození parametrů kanonického modelu $\mathcal{M}$ a váhových vektorů (transformací) $\boldsymbol{\Upsilon} = \boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^R$ se s výhodou používá algoritmus EM (viz podkapitola 2.3.1). Pomocná funkce pak má tvar

$$Q(\mathcal{M}, \boldsymbol{\Upsilon}, \hat{\mathcal{M}}, \hat{\boldsymbol{\Upsilon}}) = -\frac{1}{2}\sum_r \sum_{jm} \sum_t \gamma_{jm}(t) \left((\boldsymbol{o}^r(t) - \boldsymbol{M}_{jm}\nu^r)^T C^{-1}_{jm} (\boldsymbol{o}^r(t) - \boldsymbol{M}_{jm}\nu^r) \right), \tag{4.15}$$

kde $\mathcal{M}$ je starý kanonický model a $\hat{\mathcal{M}}$ je nově odvozený model (analogicky i pro $\boldsymbol{\Upsilon}$). Je nesnadné odvozovat kanonický model $\mathcal{M}$ a transformace $\boldsymbol{\Upsilon}$ společně, proto se odhad provádí ve dvou krocích, nejprve $\hat{\boldsymbol{\Upsilon}}$ a pak $\hat{\mathcal{M}}$. Obvykle se postup iterativně opakuje dokud kritérium nezačne konvergovat.

4.2.2 Reprezentace shluků

Existují dvě možnosti jak reprezentovat střední hodnoty jednotlivých shluků, **CAT založené na modelu** a **CAT založené na transformacích**. Jejich kompletní popis je uveden v [Gal00]. V prvním zmíněném způsobu je každý shluk přímo reprezentován akustickým modelem, druhý způsob popisuje shluk adaptační maticí, která transformuje globální model na model daného shluku (např. metoda MLLRmean viz podkapitola 3.3.1). Pro inicializaci kanonických modelů [Yu06] se s výhodou využívají **dekompozice vlastních hlasů** (ED – Eigenvoices Decomposition) (viz podkapitola 3.6).

Výhodou metody CAT je rychlá adaptace pro malý objem adaptačních dat. V porovnání s jinými adaptačními metodami, jako je např. SAT, metoda CAT vyžaduje znatelně menší počet adaptačních parametrů (dimenze P váhového vektoru je obvykle v jednotkách). Čím méně parametrů je pro metodu CAT použito, tím menší je její efektivita v porovnání s metodou SAT. Metoda může být také snadno a efektivně rozšířena jinou adaptací viz [Gal01].

4.2.3 Diskriminativní adaptace pro trénování pomocí shlukování (DCAT)

Rozšířením přístupu CAT je **diskriminativní adaptace pro trénování pomocí shlukování mluvčích** (DCAT – Discriminative CAT). Na rozdíl od klasické metody CAT, jednotlivé shlukové modely jsou zde trénovány pomocí disktiminativních metod 2.3.3. Ty však vyžadují mnohem více dat, než je třeba pro klasické trénování založené na ML kritériu (2.14). Navržené metody využívající diskriminativní kritéria MMI (2.32) či MPE (2.33) jsou uvedeny v [YMJFG06].

4.3 Normalizace délky hlasového traktu (VTLN)

Důvodů řečové variability mezi řečníky je velké množství, např. lingvistické odlišnosti, způsob artikulace, zdravotní a psychický stav řečníka a jiné. Převládajícím faktorem však je rozlišná fyziologická stavba hlasového ústrojí. Jedním z hlavních zdrojů odlišnosti řečníků je rozdílná délka hlasové trubice, která se může pohybovat od 13 cm pro ženy do 18 cm pro muže. Délka hlasového traktu zásadně ovlivňuje polohu formantových frekvencí (a to s nepřímou úměrou), které jsou detekovány převážně u znělých hlásek.

Metoda **normalizace délky hlasového traktu** (VTLN – Vocal Tract Length Normalization) [ZW97] se snaží kompenzovat projevy různé délky hlasové trubice v řeči transformováním frekvenční osy řečníka tak, aby se jeho pozice formantů blížily pozicím průměrného řečníka.

4.3.1 Transformační funkce

Transformace frekvenční osy spočívá v jejím nelineárním natažení (popř. smrštění), odborně nazývaném **borcení** (anglicky **warping**). Warpovacích funkcí $\tilde{\omega} = \mathcal{F}_\alpha(\omega)$ je celá řada, nejpoužívanější z nich jsou podle [PKT03] tyto dvě:

- 1. Po částech lineární funkce

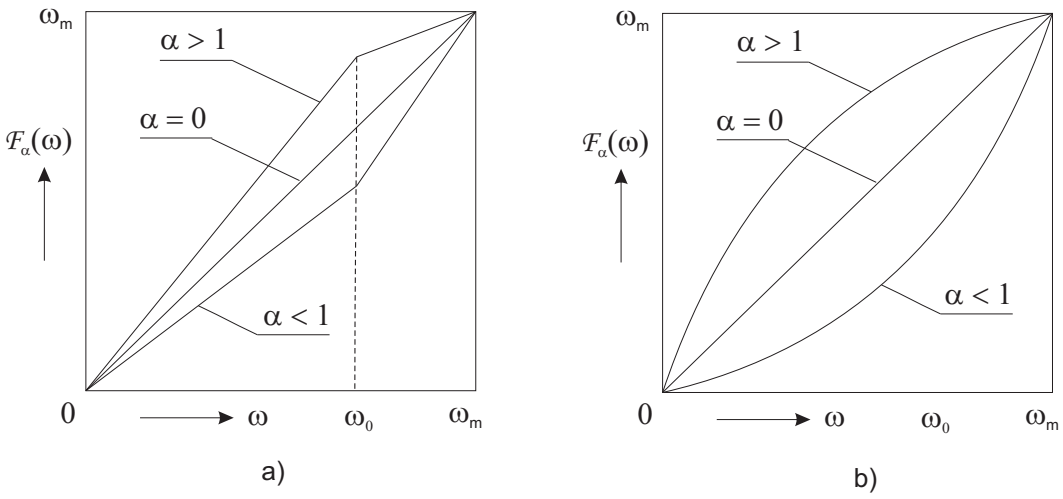
$$\mathcal{F}_\alpha(\omega) = \begin{cases} \alpha\omega & \text{pro } 0 \leq \omega < \omega_0 \\ \alpha\omega + \frac{\pi - \alpha\omega_0}{\pi - \omega_0}(\omega - \omega_0) & \text{pro } \omega_0 \leq \omega \leq \omega_m, \end{cases} \tag{4.16}$$

  kde frekvence ω_0 je rovna nebo větší jak průměrná frekvence třetího formantu a je α warpovací faktor. Průběh funkce je zobrazen v levé části obrázku 4.4.

- 2. Bilineární funkce

$$\mathcal{F}_\alpha(\omega) = \omega + 2\arctan\left(\frac{(1-\alpha)\sin\omega}{1-(1-\alpha)\sin\omega}\right), \tag{4.17}$$

  Průběh funkce je zobrazen v pravé části obrázku 4.4.



Obrázek 4.4: Warpovací funkce a) po částech lineární, b) bilineární.

Pro obě funkce platí, že původní frekvenční osa je transformována na stejný interval $\tilde{\omega} \in \langle 0, \omega_m \rangle$. Naším úkolem je nalézt pro každého řečníka jeho warpovací faktor α tak, aby byl nejlépe znormalizován jeho hlasový trakt. α je obvykle hledáno z intervalu $\langle 0,88; 1,12 \rangle$ [PMMR06]. Jinou warpovací funkci s více než jedním proměnným parametrem lze nalézt např. v [PA06].

Zatímco metoda SAT je zaměřena na transformaci parametrů modelu, metoda VTLN pracuje s vektory příznaků. Předpokládáme-li, že vektor pozorování je parametrizován pomocí metody **melovských frekvenčních kepstrálních koeficientů** (MFCC – Mel Frequency Cepstral Coefficient), popř. metodou **perceptivní lineární prediktivní analýzy** (PLP – Perceptual Linear Predictive), lze warpování frekvenční osy provádět buď přímo přes spektrální vzorky nebo transformovat meze jednotlivých pásem v bance filtrů. Druhý ze zmíněných způsobů je výpočetně méně náročný.

4.3.2 Odhad warpovacího faktoru

Nalezení optimálního warpovacího faktoru r-tého řečníka α_r^* jde ruku v ruce s optimalizačním kritériem pro rozpoznávání. Označme soubor trénovacích promluv r-tého řečníka $\boldsymbol{O}_r = \{\boldsymbol{O}_r^1, \ldots, \boldsymbol{O}_r^E\}$ a k němu odpovídající soubor přepisů $\boldsymbol{W}_r = \{\boldsymbol{W}_r^1, \ldots, \boldsymbol{W}_r^E\}$, pak soubor těchto promluv warpovaných faktorem α můžeme označit $\boldsymbol{O}_r^\alpha = \{\boldsymbol{O}_r^{1\alpha}, \ldots, \boldsymbol{O}_r^{E\alpha}\}$.

Optimální warpovací faktor pro daného řečníka lze nalézt maximalizací věrohodnosti warpovaných promluv za předpokladu SI modelu λ a daného přepisu promluv $\boldsymbol{W}_r$

$$\alpha_r^* = \arg\max_\alpha P(\boldsymbol{O}_r^\alpha|\lambda, \boldsymbol{W}_r). \tag{4.18}$$

Pro zjednodušené hledání warpovacího faktoru α_r^* byl navržen vhodný interval doporučených hodnot α, uvedený v této kapitole v části 4.3.1.

Jiné, než výše uvedené ML kritérium pro výběr optimálního warpovacího faktoru, tzv. **lineární diskriminativní kritérium** (LD – Linear Discriminant), lze nalézt v [WSW98]. Je založeno na kovariančních maticích daných akustických vzorků. Předpokládáme, že každý vzorek je přidružen do některé z akustických tříd. Pak LD kritérium má formu

$$LD = \frac{|T|}{|W|}, \tag{4.19}$$

kde T je kovarianční matice všech vzorků a W je průměrná kovarianční matice vzorků patřících do konkrétních tříd c_i

$$W = \sum_i p(c_i) W_i. \tag{4.20}$$

Hledáme takový parametr α_i^*, který maximalizuje kritérium (4.19). Tehdy budou různé třídy vzorků od sebe vzájemně daleko, ale mají v průměru malý rozptyl mezi svými vzorky. Tato metoda je také použita pro rychlou transformaci při on-line použití v [PKT03].

4.3.3 Normalizovaný akustický model

S pomocí warpovaných promluv $\boldsymbol{O}_r^\alpha$ lze natrénovat kompaktní model λ_c, který je "na míru ušitý" na řečníka s průměrnou délkou vokálního traktu. Při procesu rozpoznávání je pak nutné testované promluvy normalizovat příslušným warpovacím faktorem.

Kapitola 5

Experimenty

V této kapitole se nachází experimentální výsledky adaptačních metod, které byly doposavad programově realizovány. Navržené experimenty byly zaměřeny na adaptaci řečníka. Systém ASR neobsahoval žádný jazykový model, v promluvách se nenacházela žádná **slova mimo slovník** (OOV – Out Of Vocabulary). Výsledky zde uvedené jsou pouze ke srovnání jednotlivých adaptačních metod, nikoliv celého systému.

5.1 Data a akustický model

Všechny experimenty byly provedeny na českém telefonním korpusu obsahujícím čtenou řeč. Digitalizace analogového telefonního signálu byla provedena pomocí DIALOGIC D/21D vzorkovací frekvencí 8 kHz a následnou konverzí do μ-law 8 bitového rozlišení.

Z korpusu bylo vybráno 100 trénovacích řečníků a kolem 40 vět (30-40 tisíc vzorků) pro každého z nich, z těchto dat byl natrénován SI model. Jiných 100 řečníků bylo použito na experimenty, kde adaptační část obsahovala po 20-30 větách (průměrně 20 tisíc vzorků pro řečníka) ke každému z řečníků, testovací část pak jiných 5 vět. Slovník pro přepis obsahoval 475 různých slov, kde několik z nich mělo více různých fonetických přepisů, tedy finální počet položek ve slovníku byl 528. V testovacích větách se nenacházela žádná OOV slova.

Nahrávky byly zpracovány MFCC parametrizací s 12 filtry (použity delta i delta-delta koeficienty), 25 ms okénko s posuvem 10 ms. Na závěr byla použita metoda kepstrální normalizace CMN. Akustický model byl trifónový HMM s 8 složkami pro každý stav s diagonálními kovarianční maticemi.

5.2 Hodnocení úspěšnosti rozpoznávání

V úlohách rozpoznávání řeči je výsledný přepis porovnán s referenčním textem dané promluvy pomocí algoritmu dynamického borcení času (DTW – dinamic Time Warping) [Psu95]. Úspěšnost přepisu se dá hodnotit [Čer07] například pomocí **procenta chybně rozpoznaných slov** (WER – Word Error Rate), **přesností** (ACC – Accuracy) a **správností** (CORR – Correctness) výsledného přepisu. Slovo správně rozpoznáno je označeno jako H

(Hit), špatně rozpoznáno S (Substituce), slova která v přepisu chybí jsou D (Delete) a ta, která přebývají I (Inzerce). Jednotlivé míry úspěšnosti lze psát ve tvaru

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

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

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

5.3 Výsledky

5.3.1 Klasické metody adaptace

V prvním řádku tabulky 5.1 jsou uvedené hodnoty ACC pro experiment s SI modelem v prvním sloupci a modelem adaptovaným metodou MAP, MLLRmean, MLLRvar, fMLLR v sloupcích následujících. Veškeré výsledky jsou pouze po jedné adaptační iteraci, vyjímkou je metoda MLLRcov, která je z principu dvouiterační (viz podkapitola 3.3.1). Obecně lze předpokládat další zlepšování pro více adaptačních cyklů. Druhý řádek tabulky uvádí průměrný čas adaptace na jednoho řečníka[1].

V metodě MAP byly adaptovány střední hodnoty, kovarianční matice i váhy složek najednou. Konstanta τ byla experimentálně nastavena na hodnotu 16.

Regresní strom v metodě MLLR (resp. fMLLR) byl konstruován pomocí HTK verze 3.4 [YEG+06]. Ke konstrukci byla využita pouze blízkost středních hodnot. Strom měl 32 listových uzlů, tedy 32 základních shluků, které se pak podle aktuálního množství adaptačních dat spojovaly do sebe dle navrženého stromu.

Tabulka 5.1: ACC[%] vybraných adaptačních metod.

	SI model	MAP	MLLR mean	MLLR cov	fMLLR
ACC	66.18%	75.00%	74.84%	**77.93**%	76.99%
čas adaptace		**0.69**s	2.39s	17.03s	14.91s

Z výsledků je vidět podstatné zlepšení přesnosti rozpoznávání při použití adaptačních metod a to až o 17% relativně. Nejlepší výsledky dává metoda MLLRcov, která adaptuje jak střední hodnoty, tak kovarianční matice modelu a to různými transformacemi. Tato metoda v přesnosti předčí i fMLLR (ta adaptuje střední hodnoty a kovarianční matice stejnou transformací), ale je ze všech testovaných metod nejpomalejší.

Nejrychlejší metodou je MAP, pro kterou jsme v testu měli dostatečné množství dat, proto i ona má dobré výsledky. Malá rychlost adaptací založených na lineárních transformacích lze přičítat hlavně velkému regresnímu stromu. Pro takové množství adaptačních dat obsažených

[1] Výpočet prováděn na domácí stanici s procesorem Core2duo a vnitřní pamětí 2MB.

v našem testu bylo vytvořeno v průměru 10 transformačních matic pro každého řečníka. Rapidní zpomalení metod MLLRcov a fMLLR (oproti MLLR) je také vinnou nutného iterativní výpočtu matic uvnitř vlastní adaptace pro adaptaci vektorů příznaků.

5.3.2 Kombinace adaptačních metod

Tabulka 5.2 zobrazuje výsledné ACC získané po kombinaci vybraných metod adaptace. Kombinace spočívala v postupné adaptaci pomocí dvou nebo třech metod aplikovaných ve dvou resp. třech adaptačních iteracích. Byly použity adaptační metody se stejným nastavením popsaným výše.

Tabulka 5.2: ACC[%] kombinace adaptačních metod.

MAP -MLLRmean	MAP -MLLRcov	MAP -fMLLR	MLLRmean -MAP	MLLRmean -MAP-fMLLR	MLLRmean -MAP-MLLRcov
77.03%	78.14%	78.37%	77.22%	78.37%	78.46%

Kombinace metod vykazují další zlepšení adaptace. Optimálním z hlediska účinnosti a časové náročnosti (viz 5.1) se jeví kombinace metody MAP, která je značně rychlá, s adaptací vektoru pozorování (fMLLR).

5.3.3 Adaptační trénování

Tabulka 5.3 obsahuje výsledné ACC samotného SI modelu a přetrénovaného metodou SAT. V dalších řádcích je porovnání těchto modelů po adaptaci. Opět měly adaptační metody stejné nastavení jako v prvním experimentu popsaném výše. Metoda SAT adaptovala pouze střední hodnoty modelu.

Tabulka 5.3: Porovnání ACC[%] dané SI a SAT modelem po adaptaci.

typ adaptace	SI	SAT
žádná	66.18%	68.18%
MAP	75.00%	75.28%
MLLR	74.75%	76.82%
fMLLR	76.99%	78.21%

Metoda SAT odstraňuje z modelu informaci o řečníkovi, model se pak stává vhodnější pro adaptaci a adaptační metody na něm vykazují lepší účinnost v porovnání s SI modelem.

5.3.4 Množství dat pro adaptaci

Potřebný počet dat pro adaptaci metodou MAP a fMLLR je uvedeno v tabulce 5.4. Výsledky ACC jsou průměrem ze sta řečníků (stejných jako v experimentech výše) pro různý

počet adaptačních vět. Průměrná věta se skládá přibližně z 1000 vzorků (vektorů pozorování).

Tabulka 5.4: ACC[%] při různém počtu adaptačních vět.

	Adaptace	
počet vět	MAP	fMLLR
1	67.57%	72.10%
2	68.14%	74.82%
3	68.43%	75.76%
4	69.15%	75.60%
5	69.14%	75.36%
6	69.97%	75.75%
8	70.63%	75.59%
10	70.89%	76.14%
12	72.17%	76.20%

Metoda fMLLR dokázala (oproti MAP) adaptovat model již při malém počtu adaptačních dat, což jí činní ideální pro postupnou adaptaci v aplikacích pracujících v reálném čase.

5.4 Zhodnocení experimentů

Experimenty provedené na českém korpuse s dostatečným množstvím dat a velkým počtem různých řečníků dokázaly opodstatnění adaptace na řečníka. Z výsledků jsou také patrné výhody a nevýhody jednotlivých metod, jejich rychlost a účinnost. Například metoda MAP se ukázala být dobrou volbou pro první iterační krok. Pomocí ní se změní adaptačními daty dobře podmíněné složky modelu, ostatní složky jsou pak v druhé iteraci zpřesněny jinou metodou, např. fMLLR s výhodou adaptace vektoru pozorování.

Nejlepší výsledky byly získány aplikací metody MLLRcov, která adaptuje jak střední hodnoty, tak i diagonální kovarianční matice (odlišnými transformačními maticemi oproti metodě fMLLR). Kovarianční matice jsou adaptovány prostřednictvím transformací vektorů pozorování, tím je snížena jak časová tak i paměťová náročnost metody MLLRcov (resp. fMLLR).

Model předpřipravený pomocí adaptačního trénování (v experimentech testováno pouze SAT na středních hodnotách) dokázal zvýšit účinnost jednotlivých metod v porovnání s klasicky natrénovaným modelem.

Kapitola 6

Závěr

Problém adaptace akustického modelu v úloze rozpoznávání spojité řeči je již dlouhou dobu řešen množstvím vědeckých pracovišť po celém světě. Existuje velké množství metod a přístupů v různých oblastech zpracování jak modelu tak i signálu. Přesto jde stále o otevřený problém.

Jak dochází k zrychlování výpočtů a tím k zpřesňování samotného akustického modelu, objevují se nové přístupy k adaptaci (např. diskriminativní metody), které mají větší účinnost. Aktuálním problémem je také rychlost adaptace pro použití v reálném čase, kdy je akustický model adaptován za běhu řečového rozpoznávače. Tyto dva problémy (rychlost a přesnost) jsou si navzájem v protikladu. Některé metody jsou cíleně vyvíjeny pro malý počet adaptačních dat, právě pro práci v on-line režimu. Naopak jiné úlohy vyžadují co největší přesnost modelu na konkrétního řečníka, mají k dispozici velké množství dat i dostatečný čas na adaptaci. Těmito problémy se bude zabývat má disertační práce.

6.1 Dílčí cíle disertační práce

- Zaměřit se na zlepšení účinnosti metod adaptace, převážně pak metod založených na lineárních transformacích, které vykazují dobré vlastnosti i pro malý počet adaptačních dat. V těchto metodách realizovat diskriminativní přístup.
- Dále se zabývat zrychlením metod a jejich implementaci v úloze on-line adaptace.
- Popsané metody mění z množiny parametrů modelu pouze ty, které popisují pravděpodobnost stavu, pro Gaussovské rozložení jde o střední hodnoty, kovarianční matice a váhy. Lze předpokládat, že SI model trénovaný na datech od velkého počtu řečníků bude mít vyšší složitost struktury, než SA model vázaný pouze na jednoho konkrétního řečníka. Strukturou modelu se v tomto případě myslí počet složek GMM v jednotlivých stavech modelu. Protože rozptyl v datech od jednoho konkrétního řečníka bude podstatně menší než v datech od více řečníků, lze u SA modelu uvažovat o změně počtu složek GMM. Disertační práce by se tedy měla zaměřit na adaptaci právě těchto parametrů, na snížení složitosti výsledného modelu.

Literatura

[AMSM96] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul. A compact model for speaker-adaptive training. In *IEEE International Conference on Spoken Language Processing*, pages 1137–1140, 1996.

[AW97] S. M. Ahadi and P. C. Woodland. Combined bayesian and predictive techniques next term for previous termrapid speaker adaptationnext term of continuous density hidden markov models. *Computer Speech & Language*, 11(3):187–206, 1997.

[BPSW70] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[Cho90] Y.L. Chow. Maximum mutual information estimation of HMM parameters forcontinuous speech recognition using the n-best algorithm. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 701–704, Albuquerque, USA, 1990.

[CXWF06] Shih-Sian Cheng, Yeong-Yuh Xu, Hsin-Min Wang, and Hsin-Chia Fu. *Lecture Notes in Computer Science*, chapter Automatic Construction of Regression Class Tree for MLLR Via Model-Based Hierarchical Clustering, pages 390–398. Springer Berlin / Heidelberg, 2006.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[DRN95] V. Digalakis, D. Rtischev, and L. Neumeyer. Speaker adaptation using constrained estimation of gaussian mixtures. *IEEE Transactions On Speech and Audio Processing*, 3(3):357–366, 1995.

[Čer07] Petr Červa. *Řízená a neřízená adaptace na mluvčího v systémech rozpoznávání řeči*. PhD thesis, Technická univerzita v Liberci, 2007.

[FR98] C. Fraley and A. E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Jurnal*, 41:578–588, 1998.

[Gal96] M.J.F. Gales. The generation and use of regression class trees for MLLR adaptation. Technical report, Cambridge University Engineering Department, 1996.

[Gal97] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical report, Cambridge University Engineering Department, 1997.

[Gal00] M.J.F. Gales. Cluster adaptive training of hidden Markov models. *IEEE Transactions on Speech and Audio Processing*, 8:417–428, 2000.

[Gal01] M. J. F. Gales. Multiple-cluster adaptive training schemes. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[Gan05] Juri Ganitkevitch. Speaker adaptation using maximum likelihood linear regression. Technical report, Rheinisch-Westfälische Technische Hochschule Aachen, 2005.

[GL94] Jean-Luc Gauvain and Chin-Hui Lee. Maximum a-posteriori estimation for multivariate gaussian mixture observations of Markov chains. *IEEE Transactions On Speech and Audio Processing*, 2(2):291–298, 1994.

[GRP00] Yuqing Gao, Bhuvana Ramabhadran, and Michael Picheny. New adaptation techniques for large vocabulary continuous speech recognition. In *ICSA ITRW ASR*, Paris, France, 2000.

[HCC02] Chao Huang, Tao Chen, and Eric Chang. Adaptive model combination for dynamic speaker selection training. In *IEEE International Conference on Spoken Language Processing*, volume 1, pages 774 – 777, 2002.

[HWFW00] Lei He, Jian Wu, Ditang Fang, and Wenhu Wu. Speaker adaptation based on combination of map estimation and weighted neighbor regression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II 981 – II 984, Istanbul, Turkey, 2000.

[JWJY03] Gyucheol Jang, Sooyoung Woo, Minho Jin, and Chang D. Yoo. Improvements in speaker adaptation using weighted training. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I 548 – I 551, 2003.

[KNJ+98] R. Kuhn, P. Nguyen, J. C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, and M. Contolini. Eigenvoices for speaker adaptation. In *IEEE International Conference on Spoken Language Processing*, pages 1771–1774, 1998.

[LR96] Li Lee and R.C. Rose. Speaker normalization using efficient frequency warping procedures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 353 – 356, 1996.

[LW95] C. J. Leggeter and P. C. Woodland. Maximum likelihood linear regression for speaker adaption of continuous density hidden markov models. *Computer Speech and Language*, 9:171–185, 1995.

[MHSN05] Wolfgang Macherey, Lars Haferkamp, Ralf Schlüter, and Hermann Ney. Investigations on error minimizing training criteria for discriminative training in automatic speech recognition. In *Interspeech*, Lisbon, Portugal, 2005.

[MIT97] M. Morishima, T. Isobe, and J. Takahashi. Phonetically adaptive cepstrum mean normalization for acousticmismatch compensation. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 436 – 441, 1997.

[MSC00] Tor Andre Myrvoll, Olivier Siohan, and Chin-Hui Leeand Wu Chou. Structural maximum a posteriori linear regression for unsupervised speaker adaptation. In *IEEE International Conference on Speech and Language Processing*, Beijing, China, 2000.

[MSJN97] Spyros Matsoukas, Rich Schwartz, Hubert Jin, and Long Nguyen. Practical implementations of speaker-adaptive training. In *DARPA Speech Recognition Workshop*, 1997.

[MZ07] Lukáš Machlica and Zbyněk Zajíc. The speaker adaptation of an acoustic model. In *The 1st Young Researchers Conference on Applied Sciences*, pages 212–217, Plzeň, 2007. ZČU.

[PA06] S. Panchapagesan and A. Alwan. Multi-parameter frequency warping for VTLN by gradient search. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006.

[PBNP98] M. Padmanabhan, L.R. Bahl, D. Nahamoo, and M.A. Picheny. Speaker clustering and transformation for speaker adaptation in speech recognition systems. *IEEE Transactions on Speech and Audio Processing*, 6:71 – 77, 1998.

[PGKW03] D. Povey, M.J.F. Gales, D.Y. Kim, and P.C. Woodland. Mmi-map and mpe-map for acoustic model adaptation. In *Eurospeech*, 2003.

[PKT03] Dénes Paczolay, András Kocsor, and László Tóth. *Lecture Notes in Computer Science*, volume 2807/2003, chapter Real-Time Vocal Tract Length Normalization in a Phonological Awareness Teaching System, pages 309–314. Springer Berlin / Heidelberg, 2003.

[PMMR06] Josef Psutka, Luděk Müller, Jindřich Matoušek, and Vlasta Radová. *Mluvíme s počítačem česky*. ACADEMIA Praha, 2006.

[Pov03] Daniel Povey. *Discriminative Training for Large Vocabulary Speech Recognition*. PhD thesis, University of Cambridge, 2003.

[PS06] Daniel Povey and George Saon. Feature and model space speaker adaptation with full covariance gaussians. In *Interspeech*, pages 1145 – 1148, Pittsburgh, PA, USA, 2006.

[Psu95] Josef Psutka. *Komunikace s počítačem mluvenou řečí*. Academia, 1995.

[PW99] D. Povey and P.C. Woodland. Frame discrimination training of HMMs for large vocabulary speech recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 333–336, 1999.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE*, 77(2):257–286, 1989.

[SBD95] A. Sankar, F. Beaufays, and V. Digalakis. Training data clustering for improved speech recognition. In *Eurospeech*, pages 502–505, 1995.

[SDP04] G. Saon, A. Dharanipragada, and D. Povey. Feature space gaussianization. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 329–332, 2004.

[SKFS07] Andreas Stolcke, Sachin S. Kajarekar, Luciana Ferrer, and Elizabeth Shriberg. Speaker recognition with session variability normalization based on MLLR adaptation transforms. *IEEE International Conference on Spoken Language Processing*, 15:1987–1998, 2007.

[SL97] Koichi Shinoda and Chin-Hui Lee. Structural map speaker adaptation using hierarchical priors. In *IEEE Automatic Speech Recognition and Understanding*, pages 381 – 388, Santa Barbara, USA, 1997.

[SM98] Ralf Schlüter and Wolfgang Macherey. Comparison of discriminative training criteria. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[TDB05] Stavros Tsakalidis, Vlasios Doumpiotis, and William Byrne. Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation. *IEEE Transactions on Speech and Audio Processing*, 13:367 – 376, 2005.

[UW01] L.F. Uebel and P.C. Woodland. Discriminative linear transforms for speaker adaptation. In *ISCA ITRW on Adaptation Methods for Automatic Speech Recognition*, pages 61–64, Sophia, 2001.

[Vit67] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[Wes99] Robert Westwood. Speaker adaptation using eigenvoices. Technical report, Cambridge University Engineering Department, 1999.

[WSMN01] Frank Wessel, Ralf Schlüter, Klaus Macherey, and Hermann Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions On Speech and Audio Processing*, 9(3):288–298, 2001.

[WSW98] Martin Westphal, Tanja Schultz, and Alex Waibel. Linear discriminant a new criterion for speaker normalization. In *IEEE International Conference on Spoken Language Processing*, page paper no. 755, Sydney, Australia, 1998.

[WW03] L. Wang and P.C. Woodland. Discriminative adaptive training using the mpe criterion. In *IEEE Automatic Speech Recognition and Understanding*, 2003.

[WW04] L. Wang and P.C. Woodland. Mpe-based discriminative linear transform for speaker adaptation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 321–324, 2004.

[YEG+06] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2001-2006.

[YMJFG06] Kai Yu and IEEE Mark J. F. Gales, Member. Discriminative cluster adaptive training. *IEEE International Conference on Spoken Language Processing*, 14(5):1694–1703, 2006.

[Yu06] Kai Yu. *Adaptive Training for Large Vocabulary Continuous Speech Recognition*. PhD thesis, Hughes Hall College and Cambridge University Engineering Department, 2006.

[ZS05] Jing Zheng and Andreas Stolcke. Improved discriminative training using phone lattices. In *Interspeech*, 2005.

[ZW97] Puming Zhan and Martin Westphal. Speaker normalization based on frequency warping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042, Munich, Germany, 1997.