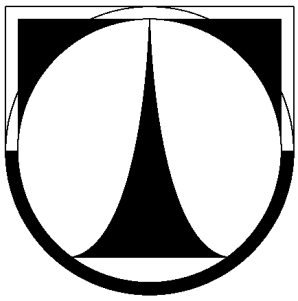
Technická univerzita v Liberci
Fakulta mechatroniky, informatiky a mezioborových studií
Ústav informačních technologií a elektroniky

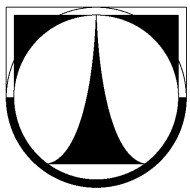


Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích

DISERTAČNÍ PRÁCE

Listopad 2011 Ing. Jan Silovský

Technická univerzita v Liberci
Fakulta mechatroniky, informatiky a mezioborových studií
Ústav informačních technologií a elektroniky



Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích

Generative and discriminative classifiers in the tasks of text-independent speaker recognition and diarization

Ing. Jan Silovský

Školitel: Prof. Ing. Jan Nouza, CSc.

Disertační práce předkládaná na Fakultě mechatroniky, informatiky a mezioborových studií
Technické univerzity v Liberci
v souvislosti s plněním požadavků
k získání akademického titulu „**Doktor**“
v oboru
2612V045 Technická kybernetika
studijního programu P2612 Elektrotechnika a informatika

Listopad 2011

Prohlášení

Prohlašuji, že jsem tuto disertační práci vypracoval samostatně s využitím uvedené literatury a na základě konzultací se svým školitelem a kolegy z Laboratoře počítačového zpracování řeči na Fakultě mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci.

Liberec, 30. listopadu 2011 *Jan Silovský*

Poděkování

Nejprve bych rád poděkoval panu Prof. Ing. Janu Nouzovi, CSc. za jeho pomoc, ochotu a čas věnovaný mi během celého doktorského studia. Děkuji kolegům z Laboratoře počítačového zpracování řeči na Fakultě mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci za jejich přátelství, pomoc a povzbuzení v mé práci. Dále děkuji kolegům ze Skupiny zpracování řeči působící na Fakultě informačních technologií Vysokého učení technického v Brně za podnětné diskuze a ochotu sdílet informace o výsledcích jejich práce. Na závěr vřele děkuji své rodině za trpělivost a podporu při psaní této práce.

Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích

Ing. Jan Silovský

Technická univerzita v Liberci, Liberec, Listopad 2011

Školitel: Prof. Ing. Jan Nouza, CSc.

Tato disertační práce se zabývá problematikou textově nezávislého rozpoznávání mluvčích. V úvodní části jsou ve stručnosti vysvětleny základní pojmy a úlohy rozpoznávání mluvčích, je stručně popsán současný stav problematiky, představena motivace pro využití informace o identitě mluvčích v systémech vyvíjených Laboratoří počítačového zpracování řeči na Technické univerzitě v Liberci (TUL) a na základě toho stanoveny cíle práce.

Samostatná kapitola je věnována metodám používaným pro vyhodnocování úspěšnosti rozpoznávání, včetně metod pro takzvané aplikačně nezávislé vyhodnocení, a metodám pro kalibraci a fúzi systémů.

V následující kapitole jsou postupně představeny metody založené na generativních modelech, od standardních metod využívajících modely reprezentované směsí Gaussovských rozložení, po moderní metody založené na různých formách faktorové analýzy. V kapitole věnované metodám založeným na diskriminativním principu je pozornost soustředěna na metody založené na podpůrných vektorech a speciální jádrové funkce navržené pro úlohu rozpoznávání mluvčích.

Na příkladu aplikace rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů jsou diskutovány některé rozdílné charakteristiky dat standardních evaluačních databází a reálných aplikací. Následně jsou předloženy výsledky experimentálního vyhodnocení několika systémů, založených na generativním i diskriminativním přístupu, na vytvořené evaluační databázi českých televizních a rozhlasových pořadů. Jazykové omezení umožňuje využití systémů vyvinutých na TUL pro získání automatického přepisu nahrávek a jeho použití při rozpoznávání mluvčích.

Následující kapitola shrnuje popis vývoje systémů pro účast TUL v evaluaci systémů pro rozpoznávání mluvčích pořádané americkým Úřadem pro standardy a technologii (NIST) v roce 2010.

Jedním z hlavních přínosů práce je pak návrh několika přístupů pro shlukování mluvčích v rámci úlohy diarizace audiozáznamů, včetně návrhu dvoufázového schématu shlukování s využitím těchto přístupů. Ty vycházejí z principů metod navržených pro rozpoznávání mluvčích a jsou založeny na faktorové analýze. Experimentální vyhodnocení prezentovaných přístupů je provedeno na základě databáze televizních a rozhlasových zpravodajských pořadů vytvořené s využitím dat korpusu COST278.

Generative and discriminative classifiers in the tasks of text-independent speaker recognition and diarization

Ing. Jan Silovský

Technical University of Liberec, Liberec, November 2011

Supervisor: Prof. Ing. Jan Nouza, CSc.

The thesis deals with problem of text-independent speaker recognition. The introduction part briefly explains the basic terms and tasks of speaker recognition and summarizes the state-of-the-art appoaches to the problem. Subsequently, motivation for utilization of knowledge about speaker's identity in systems that are being developed by the Laboratory of Computer Speech Processing at the Technical University of Liberec (TUL) is given. Finally, the main goals of this work are presented.

The next part of the thesis deals with survey of methods used for evaluation of speaker recognition systems, including methods for so called application independent evaluation, calibration and fusion of systems.

The next chapter provides an overview of generative classifiers used in text-independent speaker recognition, starting with standard methods based on Gaussian mixture models and with particular emphasis on methods based on factor analysis. The following chapter deals with discriminative classifiers and it is focused on support vector machines (SVMs) and kernel functions proposed for the purpose of speaker recognition.

Application of speaker recognition in broadcast domain is used to discuss some of aspects of different nature of standard evaluation databases and real applications. Creation of an evaluation database comprising recordings of czech television and radio programs is described and results of experimental evaluation of several systems, based on both generative and discriminative approaches, provided. The language restriction allows employment of systems for automatic speech recognition developed at TUL and utilization of automatic transcriptions within the task of speaker recognition.

Next chapter summarizes development of systems for participation of TUL in the speaker recognition evaluation conducted by National Institute of Standards and Technology (NIST) in 2010.

Inspired by the success of methods based on factor analysis in the speaker recognition task, several approaches to speaker clustering within the speaker diarization task are proposed. Further, two-stage clustering scenario is proposed with the aim to lower some of weaknesses of approaches based on factor analysis. Experimental validation of presented approaches was carried out using an evaluation database created based on data from the COST278 broadcast news corpus.

Obsah

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

APE	Applied Probability of Error.
BIC	Bayesovské informační kritérium (Bayesian Information Criterion).
CDS	skórování na základě kosinové vzdálenosti (Cosine Distance Scoring).
CMS	odečítání kepstrálního průměru (Cepstral Mean Subtraction).
DER	míra diarizační chyby (Diarization Error Rate).
DET	Detection Error Trade-off.
ECA	adaptace s využitím vlastních kanálů (Eigenchannel Adaptation).
ECE	Empirical Cross-Entropy.
EER	Equal Error Rate.
EM	Expectation-Maximization.
GD	závislý na pohlaví (Gender Dependent).
GLDS	Generalized Linear Discriminant Sequence.
GMM	směsi Gaussovských komponent (Gaussian Mixture Model).
HMM	skrytý Markovův model (Hidden Markov Model).
JFA	sdružená faktorová analýza (Joint Factor Analysis).
KKT	Karush-Kuhn-Tucker.
KL	Kullback-Leibler.
LDA	lineární diskriminativní analýza (Linear Discriminant Analysis.
LLR	logaritmus poměru věrohodností (log-likelihood-ratio).

MAP maximální aposteriorní pravděpodobnost (Maximum A Posteriori).

MD minimální divergence (Minimum Divergence).

MFCC Mel-frekvenční kepstrální koeficienty (Mel-Frequency Cepstral Coefficients).

ML maximální věrohodnost (Maximum Likelihood).

MLLR maximálně věrohodná lineární regrese (Maximum Likelihood Linear Regression).

NAP projekce rušivých atributů (Nuisance Attribute Projection).

NBE Normalized Bayes Error-rate.

NCE Normalized Cross-Entropy.

NIST Národní úřad pro standardy a technologii (National Institute of Standards and Technology).

PAV Pool Adjacent Violators.

PCA analýza hlavních komponent (Principal Component Analysis).

PLDA pravděpodobnostní lineární diskriminativní analýza (Probabilistic Linear Discriminant Analysis.

PPCA pravděpodobnostní analýza hlavních komponent (Probabilistic Principal Component Analysis).

QP kvadratické programování (Quadratic Programming).

RBF radiální bázová funkce (Radial Basis Function).

ROC Receiver Operating Characteristics.

RVM Relevance Vector Machines.

SMO Sequential Minimal Optimization.

SMS syntéza modelů mluvčích (Speaker Model Synthesis).

SNR odstup signálu od šumu (Signal-to-Noise Ratio).

SPKE míra chybně rozpoznaných mluvčích (Speaker Error rate).

SRE evaluace rozpoznávání mluvčích (Speaker Recognition Evaluation).

SVM metoda klasifikace založená na podpůrných vektorech (Support Vector Machines).

TUL Technická univerzita v Liberci.

TVR televizní a rozhlasový.

UBM univerzální hlasový model (Universal Background Model).

WCCN	normalizace kovariance uvnitř tříd (Within Class Covariance Normlization).
WER	míra chybně rozpoznaných slov (Word Error Rate).

Kapitola 1

Úvod

Mluvená řeč je nejpřirozenějším způsobem interakce mezi lidmi. Řečový signál však obsahuje velké množství informací různého druhu. Z pohledu vzniku řečového signálu se jedná o lingvistické informace, související s jazykem a obsahem promluvy, a dále informace o řečníkovi, dané vlastnostmi mluvidel (jedná se o souhrnné označení orgánů podílejících se na tvorbě řečového signálu) a zohledňující řečníkův emocionální a zdravotní stav a geografická specifika (rodný jazyk, dialekt). Z pohledu vnímaného signálu se dále přidává informace o prostředí, ve kterém řečový signál vznikl, a informace o způsobu jeho přenosu. Lidé jsou schopni přirozeně oddělit jednotlivé informace a snadno porozumět takto složitému signálu. Automatické zpracování však představuje značně komplikovanou úlohu. V případě automatického rozpoznávání mluvčích je důležitá pouze informace o řečníkovi a přítomnost ostatních informací v signálu (např. obsah promluvy nebo vliv přenosového kanálu) může snížit úspěšnost rozpoznávání. Dokonce některé složky informace o mluvčím ovlivňují negativním způsobem úspěšnost rozpoznávání. Jedná se o různé druhy variability lidského hlasu projevující se pouze po omezenou dobu nebo měnící se v čase. V krátkém časovém horizontu je hlas ovlivňován zdravotním stavem, suchostí úst nebo náladou mluvčího. V dlouhodobém časovém horizontu se hlas mění především s věkem. Snahou je provést rozpoznávání mluvčích pouze na základě takových charakteristik hlasu, které jsou víceméně neměnné. Jedná se tedy především o fyziologické vlastnosti mluvidel (např. tvar a velikost hlasového ústrojí, nosní dutiny, úst nebo rtů) a charakteristiky získané v průběhu osvojování si jazyka (ovlivněné například dosaženým vzděláním).

Metody zabývající se rozpoznáváním osob na základě jejich jedinečných fyziologických vlastností nebo charakteristik chování označujeme souhrnně jako biometrické. Nespornou výhodou v porovnání s jinými technikami pro identifikaci osob využívajícími různé bezpečnostní karty, klíče a hesla je, že u biometrických charakteristik nehrozí odcizení, ztráta, nebo zapomenutí. Síla biometriky není v utajení informací používaných pro autentizaci, ale v jedinečnosti těchto informací. Biometrické metody jsou využívány například k zabezpečení autorizovaného přístupu nebo v kriminalistice. Úvodem do problematiky biometrických metod se blíže zabývá například (Jain et al., 2004).

Mezi nejběžněji používané biometrické charakteristiky patří DNA, otisky prstů nebo obraz sítnice. Na rozdíl od hlasu tyto charakteristiky vykazují výrazně menší variabilitu. Výhodou hlaso-

vých biometrických systémů však je, že jsou dobře přijímány uživateli. Systémy jsou bezkontaktní a nehrozí zneužití citlivých osobních dat, například o zdravotním stavu osoby (Jain et al., 2004) (z obrazu sítnice lze získat informace o diabetu nebo vysokém krevním tlaku, z DNA informace o náchylnosti k různým chorobám). Systémy založené na rozpoznávání hlasu jsou vhodné pro vzdálenou autentizaci a často uváděným nasazením je tak přístup k informačním systémům pomocí telefonu. Podobně jako ostatní biometrické metody je rozpoznávání hlasu využíváno v již zmíněné kriminalistice (Alexander, 2005). Přitom vzorek hlasu je možné získat i bez spolupráce podezřelé osoby, např. formou telefonního odposlechu.

Využití se nabízí i v dalších oblastech. Rostoucí kapacita datových médií v oblasti výpočetní techniky umožňuje vytváření rozsáhlých digitálních multimediálních archivů. Užitná hodnota těchto archivů vzrůstá, pokud je umožněno efektivní vyhledávání informací a na významu proto nabývá indexace dat. S použitím technologie rozpoznávání osob můžeme nad automaticky přepsanými daty vyhledávat nejenom co bylo řečeno, ale také kým.

1.1 Úlohy rozpoznávání mluvčích

Rozpoznávání mluvčích představuje obecný pojem, pod který zahrnujeme různé úlohy související s rozlišováním osob na základě jejich hlasu. Pro odlišení těchto úloh se často používají termíny jako identifikace, verifikace (detekce), porovnávání nebo zachytávání mluvčích. Nejčastěji se však úlohy rozlišují na identifikační a verifikační.

Cílem *identifikace* mluvčích je rozhodnout, kterému mluvčímu ze skupiny mluvčích patří hlas v předložené nahrávce. Chápání rozpoznávání mluvčích jako problému identifikace se zdá být nejintuitivnější. Když slyšíme hlas někoho známého, bezprostředně se snažíme určit, resp. identifikovat mluvčího, kterému hlas přísluší. Úloha tedy odpovídá klasifikaci do N tříd, kde N je počet zvažovaných (zapsaných) mluvčích. Je zřejmé, že N má zásadní vliv na výsledek identifikace. Zatímco pro malý počet mluvčích může být úspěšnost značně vysoká (zejména pokud mají mluvčí výrazně rozdílné hlasy), s růstem N se pravděpodobnost správné identifikace snižuje. Tato závislost komplikuje vyhodnocení úspěšnosti systémů, protože základní a přitom těžko zodpověditelnou otázkou je, kolik mluvčích by mělo být uvažováno při vyhodnocení? Je také nutné uvažovat situaci, kdy hlas v předložené nahrávce nepřísluší žádnému ze zvažovaných mluvčích. V důsledku toho rozlišujeme identifikaci v *uzavřené množině*, pro kterou platí, že rozpoznávaný hlas vždy patří jednomu ze zvažovaných mluvčích, a identifikaci v *otevřené množině*, pokud není splněn uvedený předpoklad. Problém identifikace v uzavřené množině však představuje spíše pouze laboratorní úlohu, ke které je těžké nalézt praktické aplikace. Identifikace v otevřené množině odpovídá například automatické anotaci mluvčích v různých archivech multimediálních záznamů, může se jednat například o záznamy televizních a rozhlasových pořadů (Nouza et al., 2006) nebo záznamy z jednání (Luque – Hernando, 2007). Přestože lze nalézt i další praktické aplikace odpovídající úloze identifikace v otevřené množině, problémem zůstává metodika vyhodnocení systémů pro identifi-

kaci.

Cílem *verifikace* mluvčích je ověřit tvrzení o totožnosti osoby na základě záznamu jejího hlasu. Problém tedy odpovídá klasifikaci do dvou tříd, které označíme jako:

- *pravý mluvčí* – hlas v předloženém záznamu skutečně patří proklamované osobě,
- *nepravý mluvčí* – hlas patří jiné než proklamované osobě[1].

Úloha verifikace bývá proto také označována termínem *detekce* mluvčích. Této úloze odpovídá řada praktických aplikací. Typicky se jedná například o bezpečnostní systémy, kdy osoba usilující o povolení přístupu dobrovolně spolupracuje se systémem a předkládá tvrzení o své identitě například zadáním osobního identifikačního čísla (PIN) nebo pomocí bezpečnostní karty. Forenzní aplikace také častěji vyžadují posuzování jednotlivých mluvčích ze skupiny možných podezřelých oproti výběru jednoho z podezřelých. Výhodnou vlastností verifikace, vyplývající z interpretace úlohy jako klasifikace do dvou tříd, je, že výsledek není závislý na počtu mluvčích ověřovaných pro danou testovací nahrávku. Vyhodnocení verifikačních systémů je tak nezávislé na počtu nepravých mluvčích v testovacích datech. To ovšem neznamená, že úspěšnost není závislá na volbě těchto nepravých mluvčích. Budou-li pro vyhodnocení vybráni nepraví mluvčí s hlasem podobným pravému mluvčímu, bude přirozeně vyhodnocena nižší úspěšnost.

Systémy rozpoznávání mluvčích lze klasifikovat na základě různých kritérií, přičemž jedním ze základních kritérií je znalost textového přepisu promluv použitých při trénování modelů mluvčích a při rozpoznávání. Textově závislé systémy používají při zapsání (registraci) uživatele i jeho rozpoznávání stejnou promluvu. Případně systém používá omezený slovník a při zapsání uživatele jsou vytvořeny modely pro všechna slova ze slovníku. V průběhu rozpoznávání je pak uživatel vyzván, aby vyslovil vybraná slova (např. sekvenci čísel). Výhodou textově závislých systému je, že při rozpoznávání dochází k porovnávání shodných fonetických sekvencí. Textově závislé systémy jsou již dnes běžně nabízeny komerčními společnostmi a nacházejí uplatnění zejména v různých zabezpečovacích systémech. Nutným předpokladem je dobrovolná spolupráce uživatele se systémem. V řadě praktických aplikací však není možné řídit obsah promluv (multimediální archivy, telefonní odposlechy). Textově nezávislé systémy nekladou žádné nároky na obsah promluv použitých k registraci uživatele ani k rozpoznávání. Přesnost rozpoznávání těchto systémů je zpravidla nižší, protože je obtížné brát v modelech mluvčích v úvahu variabilitu akustických příznaků mezi fonémy. Tato práce se zabývá výhradně textově nezávislými systémy.

1.2 Vývoj a použití systémů pro rozpoznávání mluvčích

Vývoj a použití systémů pro rozpoznávání mluvčích lze obecně shrnout do následujících základních etap:

[1] V anglické literatuře se používá označení impostor, kterému odpovídá česky podvodník.

- *Učení systému* – trénování parametrů (příp. hyperparametrů) na datech odpovídajících zamýšlené aplikaci systému. Prakticky pro všechny systémy spočívá minimálně v natrénování univerzálního modelu (UBM), který reprezentuje rozložení zvolených příznaků bez ohledu na konkrétního mluvčího nebo nahrávku. Součástí této etapy je také provedení kalibrace systému.

- *Zapsání mluvčích* – mluvčí je zpravidla reprezentován v systému svým modelem. Zapsání mluvčího představuje proces stanovení parametrů jeho modelu na základě předložených trénovacích dat.

- *Rozpoznávání* – spočívající obvykle (bez ohledu na konkrétní úlohu) ve vyhodnocení skóre (ideálně ve formě logaritmu poměru věrohodností) pro dvojici tvořenou testovanou nahrávkou (segmentem) – modelem mluvčího.

Bylo snahou autora, aby při popisu systémů založených na různých principech bylo vždy zřejmé, jak jsou tyto etapy realizovány.

1.3 Současný stav problematiky

Problematice rozpoznávání mluvčích se v současné době ve světě věnuje řada laboratoří zabývajících se automatickým zpracováním řeči. Za jednu z hlavních příčin vzniku skutečně široké komunity, věnující se této problematice, lze považovat přirozenou jazykovou nezávislost vyvíjených systémů. Možnost opomenutí jazykových specifik umožňuje také snadné zapojení do pravidelných evaluací pořádaných americkým Národním úřadem pro standardy a technologii (NIST)[2]. Tyto evaluace představují unikátní platformu pro porovnání vyvíjených systémů a slouží jako významný stimul ve vývoji nových přístupů a metod.

Jak uvádí Furui (Furui, 2005, 2009), první snahy o automatické rozpoznávání mluvčích byly učiněny na počátku 60. let 20. století v Bellových laboratořích v USA. Byly založeny na použití bank filtrů a na měření korelace spektrogramů. V následujících 50 letech prošla problematika automatického rozpoznávání mluvčích intenzivním vývojem. Současné systémy nejčastěji využívají akustické příznaky odvozené od *kepstra* (Huang et al., 2001) signálu v krátkém časovém intervalu (v řádu desítek ms). Za standardní lze pak označit reprezentaci mluvčích statistickými modely, například v podobě směsi Gaussovských komponent (GMM). Problémem těchto systémů je především závislost příznaků, vyjadřujících pouze krátkodobou informaci o signálu, na fonetickém obsahu a také akustických podmínkách. V posledních několika letech byla představena řada přístupů a metod pro kompenzaci variability akustických podmínek, potlačení vlivu fonetického obsahu, využití příznaků vyšší úrovně (využívajících dlouhodobější informace o signálu) a aplikaci nových klasifikátorů. Následující podkapitoly stručně shrnují hlavní témata, kterým je v současnosti věnována pozornost.

[2] `http://www.itl.nist.gov/iad/mig/tests/sre/`

1.3.1 Různé formy kompenzace variability akustických podmínek

Značná pozornost byla věnována zvýšení robustnosti vůči variabilitě akustických podmínek. Z množství činitelů podílejících se na této variabilitě uveďme například vliv přenosové cesty na signál, rozdílnou charakteristiku snímacích mikrofonů nebo měnící se hluk prostředí. Byla navržena řada technik usilujících o kompenzaci variability akustických podmínek na několika úrovních rozpoznávacího procesu. Obecně rozlišujeme metody pracující na úrovni příznaků, modelů a skóre.

Tradičními metodami používanými na úrovni příznaků jsou metoda odečítání kepstrálního průměru (CMS) nebo tzv. RASTA filtrace (Hermansky – Morgan, 1994). Z později navržených metod zmiňme například metody označované jako *feature warping* (Pelecanos – Sridharan, 2001) nebo *feature mapping* (Reynolds, 2003).

Feature warping provádí transformaci distribuce hodnot kepstrálních příznaků přes určitý časový interval. Transformace je založena na přizpůsobení distribuční funkce tak, aby odpovídala normálnímu rozložení. Zobecnění tohoto principu představuje metoda nazvaná *short-time gaussianization* (Xiang et al., 2002). *Feature mapping* provádí zobrazení příznakových vektorů do tzv. neutrálního příznakového prostoru. Akustické podmínky jsou klasifikovány do tříd, například na základě typu snímacího mikrofonu, a jim příslušné modely jsou získány adaptací univerzálního modelu. Normalizace je pak provedena na základě parametrů těchto modelů. Nevýhodou tohoto přístupu je nutnost mít k trénovacím datům dostupnou informaci o příslušné třídě akustických podmínek. Pokud tyto informace nejsou dostupné, lze provést přiřazení do tříd vytvořených prostřednictvím shlukování akusticky blízkých nahrávek (Mason et al., 2005). Problémem však dále zůstává požadavek klasifikace akustických podmínek do diskrétních tříd. Při klasifikaci akustických podmínek na základě více kritérií (typ mikrofonu, prostředí, pohlaví mluvčího, atd.) narůstá velmi rychle počet modelů, které je nutné natrénovat. Navíc kategorizace na základě některých kritérií nemusí být vždy zcela jednoznačně proveditelná. Tímto problémem trpí i metoda provádějící normalizaci na úrovni modelů mluvčích označovaná jako *syntéza modelů mluvčích* (SMS) (Teunen et al., 2000).

Byla také představena řada metod provádějících normalizaci v prostoru skóre. Největší popularitu si získaly H-norm (Reynolds, 1997) (tato metoda je dnes již zastoupena jinými technikami), Z-norm nebo T-norm (Auckenthaler et al., 2000). *H-norm* (handset normalization) normalizace byla navržena pro telefonní data. Rozdílná charakteristika dvou použitých typů mikrofonů (elektretový a uhlíkový) se projevuje v prostoru skóre rozdílnými distribucemi. Pro každý model v databázi mluvčích je proto určena střední hodnota a rozptyl skóre vyhodnocených při rozpoznávání nahrávek z obou typů mikrofonů a na základě těchto hodnot je prováděna normalizace skóre. *Z-norm* (zero normalization) provádí normalizaci na základě shodných parametrů. Ty jsou však stanoveny na základě skóre vyhodnocených pro promluvy od nepravých mluvčích s cílem zjistit pro model mluvčího parametry rozložení skóre neoprávněných soudů. Princip metody *T-norm* je analogický. Odlišností je, že stanovení parametrů rozložení skóre neoprávněných soudů neprobíhá off-line v průběhu trénování modelů mluvčích, ale on-line v průběhu rozpoznávání, kdy je provedeno vyhodnocení skóre modelů nepravých mluvčích pro rozpoznávanou nahrávku. Ve srovnání s metodou Z-norm je

nevýhodou vyšší výpočetní náročnost rozpoznávání. Na druhé straně Z-norm může být negativně ovlivněn rozdílností akustických podmínek mezi testovanou promluvou a nahrávkami použitými ke stanovení normalizačních parametrů. To u metody T-norm nehrozí. Zobecněním procesu normalizace v prostoru skóre s využitím parametrů rozložení skóre nepravých mluvčích se zabývá (Mariethoz – Bengio, 2005).

Vlastností některých moderních systémů (především systémů využívajících tzv. i-vektory) je, že pro verifikační soudy poskytují *symetrická skóre*. Uvažujme, že model je natrénován s využitím jediné nahrávky, pro symetrické skóre pak platí $s(\text{train}, \text{test}) = s(\text{test}, \text{train})$. Jinými slovy je jedno, která z dvojice nahrávek je použita jako trénovací a která jako rozpoznávaná. Pro tento případ byla navržena *S-norm* normalizace (Brummer – Strasheim, 2009), která je založena na stanovení parametrů dvou rozložení. Jedná se o rozložení skóre vyhodnocených pro verifikační soudy tvořené trénovací nahrávkou a promluvami nepravých mluvčích a dále rozložení skóre vyhodnocených pro soudy tvořené testovanou nahrávkou a promluvami nepravých mluvčích. Sada promluv nepravých mluvčích je v obou případech shodná. Skóre určené systémem pro předložený soud je pak normalizováno zvlášť parametry obou rozložení (tedy rozložení stanovených pro trénovací i testovanou nahrávku) a výsledné normalizované skóre stanoveno jako součet těchto hodnot. Tímto způsobem je zachována symetričnost skóre i po provedení normalizace.

1.3.2 Metody založené na faktorové analýze

Metody založené na faktorové analýze (Kenny – Dumouchel, 2004a; Vogt et al., 2005) byly také navrženy s cílem omezení vlivu variability akustických podmínek na rozpoznávání. Z pohledu klasifikace použité v předchozí podkapitole se jedná o metody pracující na úrovni modelů mluvčích.

Zavádí se pojem tzv. *supervektoru*, který je tvořen zřetězením vektorů středních hodnot jednotlivých komponent GMM modelu. Jedním ze základních předpokladů metod faktorové analýzy je možnost vyjádření variability způsobené odlišností hlasů mluvčích v podprostoru prostoru supervektorů, tento podprostor je definován pomocí tzv. vlastních hlasů. Podobně je vyjádřena variabilita způsobená rozdílností akustických podmínek pomocí tzv. vlastních kanálů. Není přitom vyžadována žádná kategorizace akustických podmínek. Model sdružené faktorové analýzy (JFA) navíc zahrnuje člen charakterizující variabilitu hlasu mluvčích, která má individuální charakter a není tak popsatelná ve společném podprostoru. Vlastní hlasy, vlastní kanály a ostatní parametry modelu jsou standardně stanoveny pomocí EM (Expectation-Maximization) algoritmu (Kenny, 2005; Vogt – Sridharan, 2008).

Zjednodušenou variantou JFA je metoda adaptace s využitím vlastních kanálů (ECA) (Burget et al., 2007). V tomto případě se používá pouze prostor vlastních kanálů. Mluvčí nejsou modelováni v prostoru vlastních hlasů, ale tradičně pomocí modelů odvozených adaptací univerzálního hlasového modelu (UBM) metodou maximální aposteriorní pravděpodobnosti (MAP). Rozdíl spočívá také ve způsobu stanovení vlastních kanálů, které jsou v tomto případě stanoveny metodou analýzy hlavních komponent (PCA). V procesu rozpoznávání umožňuje metoda ECA přizpůso-

bení modelu mluvčího, trénovaného za libovolných akustických podmínek, odlišným akustickým podmínkám rozpoznávané promluvy.

1.3.3 Diskriminativní klasifikátory

Tradiční metody rozpoznávání mluvčích založené na GMM modelech i metody založené na faktorové analýze patří do třídy tzv. generativních klasifikátorů. V nedávné době byly v různých oblastech automatického rozpoznávání, včetně rozpoznávání mluvčích, úspěšně aplikovány systémy založené na diskriminativních SVM (Support Vector Machines) klasifikátorech. Diskriminativní povaha SVM vyhovuje především úloze verifikace mluvčích, nicméně aplikace v úloze identifikace je také možná.

Stěžejním krokem při aplikaci SVM klasifikátorů je volba vhodné jádrové funkce. První systémy pro rozpoznávání mluvčích založené na SVM klasifikátorech využívaly tradiční jádrové funkce (polynomiální nebo radiální bázové funkce) aplikované přímo na jednotlivé příznakové vektory (Schmidt – Gish, 1996; Wan – Campbell, 2000). Později byly navrženy jádrové funkce umožňující přímé porovnání sekvencí dat různé délky. Jedná se například o GLDS (Generalized Linear Discriminant Sequence) jádrovou funkci (Campbell, 2002) nebo Fisherovu jádrovou funkci (Fine et al., 2001), případně její modifikaci pracující s logaritmem poměru věrohodností (Wan – Renals, 2005). V současnosti velké množství systémů založených na SVM využívá jádrovou funkci odvozenou na základě aproximace Kullback-Leiblerovy (KL) divergence GMM modelů mluvčích (Campbell et al., 2006b).

Důležitým tématem v souvislosti s diskriminativními metodami je, stejně jako v případě generativních klasifikátorů, kompenzace variability akustických podmínek. Systémy založené na SVM klasifikátorech nejčastěji využívají techniku nazvanou projekce rušivých atributů (NAP) (Solomonoff et al., 2004, 2005).

1.3.4 Využití automatických přepisů

Jedním z popsaných zdrojů variability řečového signálu je fonetický obsah promluv. Přirozeným způsobem odstranění vlivu fonetického obsahu je použití akustických modelů specifických pro definovaný inventář akustických jednotek. Konkrétně se může jednat o fonémy (Park – Hazen, 2002) nebo slova (Sturim et al., 2002). Obě metody přináší dobré výsledky v případě dostatečného množství dat pro trénování a rozpoznávání. Nevýhodou je fragmentace dat, která může zkomplikovat trénování některých modelů kvůli nedostatku dat. V případě varianty využívající modely slov navíc hrozí, že rozpoznávané promluvy neobsahují stejná slova jako promluvy použité při trénování.

Moderní systémy rozpoznávání řeči využívají často adaptační metody pro přizpůsobení univerzálního akustického modelu (nezávislého na mluvčím) na mluvčího rozpoznávané nahrávky. Jednou z nejčastěji používaných metod je metoda maximálně věrohodné lineární regrese (MLLR) (Gales, 1998), která je založena na hledání afinní transformace vektorů středních hodnot akustického modelu (obecně umožňuje metoda MLLR provedení transformace všech parametrů modelu). Parametry této transformace je možné následně použít pro charakterizaci mluvčích a provést rozpoznávání

např. pomocí SVM (Stolcke et al., 2005, 2006). Výhodou této metody je, že nedochází k žádné fragmentaci dat.

1.3.5 Příznaky vyšší úrovně

Jak již bylo zmíněno, hlas ovlivňují jak anatomické vlastnosti orgánů podílejících se na tvorbě řeči, které jsou vrozené, tak i naučené vlastnosti získané v průběhu osvojování si jazyka. Ve většině moderních systémů používané akustické příznaky získávané zpravidla na základě spektrální, resp. spíše *kepstrální* analýzy signálu přes krátké úseky se vztahují především k fyziologickým charakteristikám mluvčího. Pravděpodobně nejčastěji používaným typem těchto příznaků jsou Mel-frekvenční kepstrální koeficienty (MFCC) (Huang et al., 2001). Mezi vlastnosti ovlivněné návyky patří například *prozódie* nebo *idiolekt*.

Termín prozódie je používán k souhrnné reprezentaci vlastností jakými jsou intonace, tempo řeči a přízvuk. Příznaky používané k vyjádření těchto vlastností jsou počítány přes delší časové úseky a jsou méně ovlivnitelné nestálostí akustických podmínek a hlukem. Metody založené na využití prozodických příznaků zpravidla pracují s příznaky odvozenými od základní hlasové frekvence (příp. tzv. pitch frekvence) a energie (Adami et al., 2003; Peskin et al., 2003; Shriberg et al., 2005). V textově nezávislých systémech je nejsnáze implementovatelné porovnání příznaků vyjadřujících globální (v rámci promluvy) statistiku zmíněných parametrů. V takovém případě ovšem nejsou zachyceny informace o dočasné dynamice sekvence prozodických příznaků. Tento nedostatek je možné v případě textově závislých systémů vyřešit přímým porovnáním sekvencí příznaků reprezentujících shodný text například pomocí metody dynamického borcení času (DTW). V textově nezávislých systémech je možné provádět linearizaci trajektorie základní hlasové frekvence a energie po úsecích a následné modelování pomocí n-gramů (Adami et al., 2003).

Idiolekt představuje charakteristický způsob užívání slov a frází specifických pro konkrétního mluvčího. Využitím idiolektu pro automatické rozpoznávání mluvčích se zabývá například Doddington (2001). Mluvčí jsou v tomto případě reprezentováni jazykovými n-gramy odvozenými od univerzálního jazykového modelu, stejného, jaký se používá při rozpoznávání řeči.

Žádná z uvedených metod však nedosahuje úspěšnosti rozpoznávání srovnatelné s metodami založenými na akustických příznacích, a proto se zpravidla nepoužívají samostatně, ale v kombinaci s jinými systémy.

1.3.6 Metody fúze a kalibrace systémů

Systémy využívající různé složky informace obsažené v řečovém signálu, např. akustické nebo lingvistické, jsou vzájemně komplementární. Vzájemná komplementarita ale byla zjištěna i u generativních a diskriminativních systémů využívajících shodnou složku řečové informace (např. akustickou). Pozornost tedy musí být věnována nalezení způsobu vhodné kombinace (fúze) výsledků různých systémů. Přirozeným způsobem je např. lineární kombinace výstupních skóre. Otázkou je však vhodná volba vah koeficientů použitých pro lineární kombinaci.

V současné době je velmi často využívaným způsobem odhadu těchto vah metoda založená na lineární logistické regresi (Brummer – du Preez, 2006), která umožňuje nejenom zlepšit úspěšnost rozpoznávání, ale současně také provádí kalibraci výsledných skóre.

1.4 Motivace a cíle disertační práce

Systémy rozpoznávání mluvčích nachází uplatnění v řadě aplikací. Vzhledem k zaměření Laboratoře počítačového zpracování řeči na Technické univerzitě v Liberci (TUL) je aktuální především rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů. Rozpoznávání mluvčích ale nelze provést pro celé záznamy. Nejprve je nutné určit segmenty, ve kterých nedochází ke změně mluvčích a až pro tyto segmenty lze následně provést identifikaci mluvčího. V případě, že daný mluvčí není systému známý, měl by ho systém vyhodnotit jako neznámého mluvčího. Pokud je těchto mluvčích v záznamu většina, stává se výstup systému příliš nepřehledný. Velmi užitečnou je tak v tomto případě informace o vzájemné příslušnosti segmentů v závislosti na mluvčích. Tato úloha je označována jako *diarizace* mluvčích.

Aktuálním tématem je také aplikace technologie rozpoznávání mluvčích pro potřeby bezpečnostních složek státu a pro získávání forenzních důkazů v justiční oblasti. Zejména ve druhém zmíněném případě je velmi důležitá forma, jakou jsou výsledky vyhodnocení rozpoznávání mluvčích předkládány k interpretaci soudu. Přestože není cílem práce zabývat se podrobně tématem forenzních aplikací, je snahou seznámit čtenáře s uceleným přehledem metod používaných pro vyhodnocování systémů pro rozpoznávání mluvčích.

S ohledem na výše uvedené byly stanoveny tyto hlavní cíle disertační práce:

- prozkoumat a jednotným způsobem popsat principy moderních metod pro rozpoznávání mluvčích, s důrazem na metody založené na faktorové analýze
- vypracovat ucelený přehled metod používaných pro vyhodnocování systémů pro rozpoznávání mluvčích
- vytvořit vlastní databázi záznamů televizních a rozhlasových pořadů umožňující vyhodnocení systémů pro rozpoznávání mluvčích
- provést efektivní implementaci a vyhodnocení systémů (založených jak na generativním, tak na diskriminativním principu) na této databázi
- provést experimentální ověření s využitím standardní mezinárodní evaluační databáze
- navrhnout způsob aplikace principů metod pro rozpoznávání mluvčích v úloze diarizace mluvčích a provést experimentální ověření.

Kapitola 2

Vyhodnocování systémů rozpoznávání mluvčích

Tato kapitola se zabývá metodami pro vyhodnocování kvality systémů pro rozpoznávání mluvčích. V posledních několika letech byla tomuto tématu věnována značná pozornost, podnícená mimo jiné již zmíněnými pravidelnými evaluacemi pořádanými americkým Národním úřadem pro standardy a technologii (NIST). Schopnost reprezentativního vyhodnocení systémů je však zásadní také pro forenzní aplikace.

Připomeňme, že hovoříme-li o rozpoznávání mluvčích, máme nejčastěji na mysli úlohu identifikace nebo verifikace. Nepříjemné aspekty související s identifikační úlohou byly uvedeny v podkapitole 1.1. Jedná se především o omezené množství praktických aplikací, které odpovídají této úloze, nutnost rozlišovat situaci identifikace v uzavřené a otevřené množině a závislost vyhodnocení na množství zvažovaných mluvčích. Naštěstí nic nebrání formulaci úlohy identifikace mluvčího v dané nahrávce jako opakované verifikace (detekce)[1] všech zvažovaných mluvčích. Tím pádem je možné provést vyhodnocení stejným způsobem jako v případě systémů pro verifikaci mluvčích.

Z tohoto důvodu je v této kapitole věnována pozornost především metodám pro vyhodnocení systémů pro verifikaci mluvčích. Nejprve jsou popsány druhy chyb, kterých se tyto systémy dopouští při nesprávné klasifikaci soudů. Následně jsou popsány parametry, které specifikují zamýšlenou aplikaci použití systémů. Verifikační systémy však nemusí nutně poskytovat pouze binární rozhodnutí o přijetí, nebo zamítnutí proklamované totožnosti mluvčího. Ideálním výstupem systémů jsou skóre, která odpovídají (logaritmu) poměru věrohodností hypotézy o klasifikaci soudu jako oprávněného a hypotézy o klasifikaci soudu jako neoprávněného. Tato kapitola postupně probírá různé metody vyhodnocení, od kritérií určených pro vyhodnocení chybové míry při konkrétně specifikovaných aplikačních parametrech, až po charakteristiky používané pro takzvané *aplikačně nezávislé* vyhodnocení (Brummer – du Preez, 2006), které reflektují úspěšnost systému v širokém rozsahu aplikací.

[1] Termíny verifikace a detekce jsou v následujícím textu používány jako ekvivalentní. Systém pro verifikaci mluvčích pak bývá označován také jako detektor mluvčích.

Cílem kapitoly je představit a jednotným způsobem popsat nejen metody používané pro vyhodnocení systémů v této práci, ale i další metody, se kterými se lze nejčastěji setkat v současné literatuře zabývající se problematikou rozpoznávání mluvčích.

Podrobné informace týkající se metod pro vyhodnocení a kalibraci systémů pro rozpoznávání mluvčích lze nalézt v (Brummer, 2010a). Velmi přehledný úvod do problematiky aplikačně nezávislého vyhodnocení zprostředkovává van Leeuwen – Brummer (2007). Vyhodnocením systémů pro rozpoznávání mluvčích z pohledu forenzní aplikace se zabývají práce (Alexander, 2005; Ramos, 2007). Způsob vyhodnocení navržený autorem (Ramos, 2007), jenž je založen na interpretaci výsledků ve smyslu množství informace, které je možné získat aplikací detektoru mluvčích, je popsán v závěru této kapitoly.

2.1 Vyhodnocení identifikace mluvčích

Pouze pro úplnost ve stručnosti uveďme způsob vyhodnocení systémů pro identifikaci mluvčích. V rámci vyhodnocení identifikace v uzavřené množině jsou systému předkládány nahrávky od zapsaných mluvčích (takových, pro které byl natrénován model) a systém pro každou z nich určí identitu domnělého mluvčího. V případě, že tato identita není shodná s referenční identitou, dopouští se systém chyby identifikace. Protože se jedná o jediný druh chyby, který může v této úloze nastat, je v tomto případě vyhodnocení velmi přímočaré a odpovídá jednoduše souhrnné míře identifikačních chyb P_{wrong_id}, kterou lze vyjádřit jako:

$$P_{wrong\_id} = \frac{N_{wrong\_id}}{N_{total}}, \tag{2.1}$$

kde N_{wrong_id} je počet chyb identifikace a N_{total} je celkový počet nahrávek předložených k identifikaci. Výhodou této metriky je především snadná interpretovatelnost.

Vyhodnocení úlohy identifikace v otevřené množině již však není tak triviální, protože se systém může dopustit několika druhů chyb. Stejně jako v případě identifikace v uzavřené množině může systém chybně identifikovat hlas jednoho ze zapsaných mluvčích jako hlas jiného mluvčího. Dále může být systému neznámý mluvčí identifikován jako jeden ze zapsaných mluvčích, nebo hlas zapsaného mluvčího označen jako neznámý. Uvedené chyby nemusí mít v zamýšlené aplikaci stejné důsledky, a proto jim mohou být při vyhodnocení přiřazeny odlišné váhy (pokuty). Přestože je možné definovat souhrnnou pokutovou funkci, která reflektuje míry uvedených chyb a příslušných pokut, zmíněná formulace identifikační úlohy jako úlohy verifikace (detekce) umožňuje výrazně snazší vyhodnocení a analýzu systémů s využitím standardních a v literatuře dobře zdokumentovaných metod.

2.2 Chyby v úloze detekce mluvčích

V rámci vyhodnocení verifikačního systému jsou mu předkládány k posouzení dva druhy *soudů*. V prvním případě řečová data v testované nahrávce skutečně pochází od ověřovaného mluvčího.

Tato situace se označuje jako *oprávněný soud* (target trial). V druhém případě nepochází řeč v testované nahrávce od ověřovaného mluvčího a tato situace se označuje jako *neoprávněný soud* (impostor trial). Při klasifikaci soudů se systém může dopustit dvou druhů chyb, jsou to:

- chybné zamítnutí (opomenutá detekce): oprávněný soud je klasifikován jako neoprávněný a
- chybné přijetí (falešná detekce): neoprávněný soud je klasifikován jako oprávněný soud.

V průběhu vyhodnocení systému jsou určeny míry výskytu obou typů chyb:

$$P_{miss} = \frac{N_{miss}}{N_{tar}} \quad \text{a} \quad P_{FA} = \frac{N_{FA}}{N_{non}}, \tag{2.2}$$

kde N_{miss} je počet soudů, ve kterých došlo k chybnému zamítnutí, a N_{tar} je počet předložených oprávněných soudů. Obdobně N_{FA} je počet soudů, ve kterých došlo k chybnému přijetí, a N_{non} je počet předložených neoprávněných soudů. Dvě rozdílné míry chybovosti však komplikují vzájemné porovnání úspěšnosti dvou různých systémů, případně možnost posoudit zlepšení úspěšnosti systému při změně jeho parametrů. Možnost porovnání systémů je však hlavním smyslem jejich vyhodnocování, a proto je nutné zabývat se způsoby vhodného vyjádření úspěšnosti systému na základě obou uvedených chybových měr, které by porovnání umožnily.

2.2.1 Statistická významnost vyhodnocení

V souvislosti s vyhodnocením je nutné brát zřetel na jeho statistickou významnost. Mají-li být výsledky vyhodnocení statisticky významné pro nějakou vymezenou populaci, je nutné, aby byl pro vyhodnocení detektoru použit dostatečně velký vzorek dané populace. Například, jsou-li pro vyhodnocení použity převážně nahrávky studentů, jejichž získání bývá nejsnazší, je sporné, do jaké míry jsou výsledky vyhodnocení statisticky významné pro populaci zahrnující osoby jiného věku a vzdělání. Faktorů, ovlivňujících výsledky rozpoznávání mluvčích, jako je věk nebo vzdělání, je však celá řada a navíc se vzájemně ovlivňují. Snaha o provedení statisticky významného vyhodnocení s ohledem na mnoho různých kritérií může vyústit v požadavek rozsáhlého a těžko realizovatelného experimentu.

Pro určení velikosti experimentální databáze, umožňující získání statisticky významného vyhodnocení, slouží *Doddingtonovo pravidlo třiceti* (zkráceně pouze Pravidlo 30) (Doddington et al., 2000). Označme neznámou odhadovanou chybovou míru jako p, jejíž odhad určíme jako $\widehat{p} = \frac{n}{N}$, kde n je počet chyb při rozhodování N nezávislých soudů. Doddingtonovo pravidlo 30 říká, že:

> Abychom zajistili 90% interval spolehlivosti $\pm$ 30% p pro odhad $\widehat{p}$, musí být $n \geq 30$.

Pravidlo lze odvodit z následujících předpokladů (Brummer, 2010a):

1. počet chyb n má binomické rozložení s parametry p (skutečná chybová míra) a N (celkový počet soudů)
2. p je malé

3. očekávaný počet chyb Np je dostatečně velký, aby bylo možné použít aproximaci binomického rozložení pomocí normálního rozložení.

Na základě aproximace normálním rozložením lze zapsat:

$$n \sim \mathcal{N}\left(Np, Np(1-p)\right), \tag{2.3}$$

$$\widehat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{N}\right) \tag{2.4}$$

kde $\sim \mathcal{N}(\mu, \sigma^2)$ značí, že proměnná má normální rozložení se střední hodnotou μ a rozptylem σ^2. Nyní označme normalizovanou chybu odhadu $\widehat{p}$ jako $e = \frac{\widehat{p}-p}{\widehat{p}}$, takže platí

$$e \sim \mathcal{N}(0, \sigma_e^2), \tag{2.5}$$

kde

$$\sigma_e^2 = \frac{(1-p)}{n} \approx \frac{1}{n}. \tag{2.6}$$

Na základě aproximace $\sigma_e^2 \approx \frac{1}{n}$ je rozptyl normalizované chyby odhadu závislý pouze na počtu chyb n. Interval $\langle 0{,}7p; 1{,}3p\rangle$ požadovaný pro odhad $\widehat{p}$ odpovídá v případě e intervalu $\langle -0{,}3; 0{,}3\rangle$. Protože požadujeme, aby se e nacházelo v tomto intervalu s pravděpodobností ≥ 90 %, stanovíme s využitím inverzní chybové funkce (Press et al., 1992) podmínku

$$\sigma_e^2 \leq \left(\frac{0{,}3}{\mathrm{erf}^{-1}(0{,}9)\sqrt{2}}\right)^2 = 0{,}033. \tag{2.7}$$

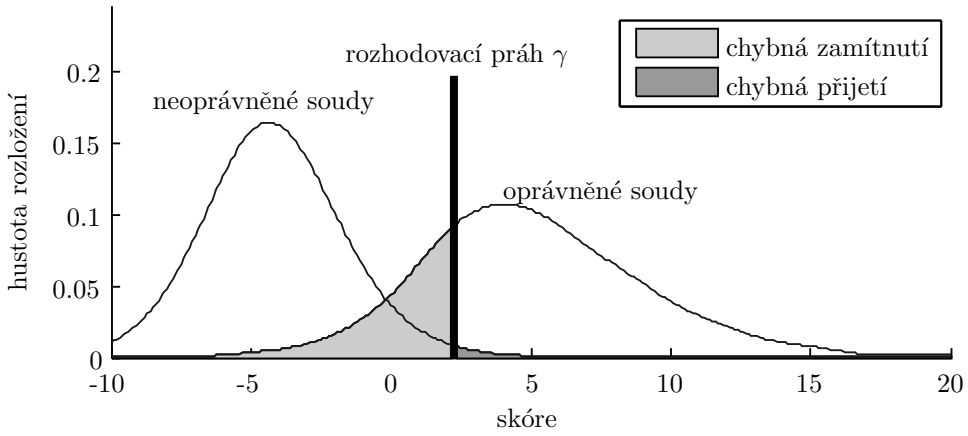
Z aproximace (2.6) pak již vyplývá požadavek $n \geq 30$.

Pro příklad předpokládejme, že požadavkům zamýšlené aplikace detektoru bude odpovídat pracovní bod, pro který bude míra chybného zamítnutí 1 % a míra chybného přijetí 0,1 %. Je tedy nutné, aby v rámci vyhodnocení bylo detektoru předloženo k vyhodnocení alespoň 3 000 oprávněných soudů a 30 000 neprávněných soudů.

Pro úplnost je nutné uvést, že předpoklad nezávislosti soudů není ve skutečnosti obvykle splněn, protože jedna testovací nahrávka je použita v několika soudech s různými mluvčími a jeden mluvčí je použit v soudech s různými nahrávkami. Důsledkem ustoupení z požadavku nezávislosti soudů je pak oslabení platnosti tvrzení o statistické významnosti výsledků.

2.2.2 ROC charakteristika

Pro každý předložený soud poskytuje detektor mluvčích jako svůj výstup *skóre*, které vyjadřuje míru podpory hypotézy, že se jedná o oprávněný soud, oproti hypotéze, že se jedná o neoprávněný soud. Jediným požadavkem, který na toto skóre prozatím budeme klást je, že vyšší hodnota skóre představuje vyšší míru podpory hypotézy o oprávněném soudu. Zatímco čím nižší je hodnota skóre, tím více je podporována hypotéza o neoprávněném soudu. Současně požadujeme, aby hodnoty skóre byly konečné. Hypotéza o oprávněném soudu je přijata, pokud skóre převyšuje zvolený práh.



Obrázek 2.1: *Rozložení skóre pro neoprávněné (vlevo) a oprávněné (vpravo) soudy. Světle šedá plocha vlevo od rozhodovacího prahu reprezentuje míru P_{miss} a tmavá plocha vpravo míru P_{FA}*

Obr. 2.1 ilustruje typické rozložení skóre určených detektorem mluvčích pro oprávněné soudy $P(s|tar)$ a rozložení skóre pro neoprávněné soudy $P(s|non)$. Rozložení skóre pro oprávněné soudy má ve srovnání s rozložením skóre pro neoprávněné soudy vyšší střední hodnotu ($\mu_{tar} > \mu_{non}$), rozptyly obou rozložení (σ^2_{tar} a σ^2_{non}) se obecně liší a rozložení se vzájemně překrývají. Rozhodnutí o přijetí jedné z hypotéz se řídí porovnáním skóre se zvoleným prahem γ a chybové míry P_{FA} a P_{miss} tak můžeme vyjádřit v závislosti na hodnotě γ jako:

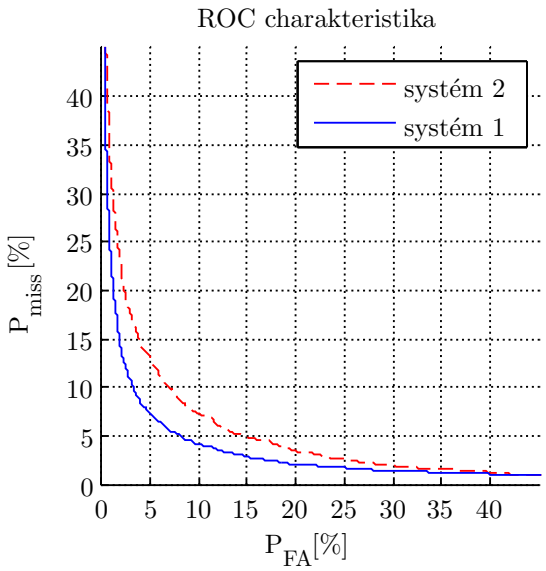
$$P_{miss}(\gamma) = P(s < \gamma|tar) = \int_{-\infty}^{\gamma} P(s|tar)ds \tag{2.8}$$

$$P_{FA}(\gamma) = P(s \geq \gamma|non) = \int_{\gamma}^{\infty} P(s|non)ds. \tag{2.9}$$

Tedy P_{FA} je možné odečíst z obr. 2.1 jako podíl skóre určených pro neoprávněné soudy, které převyšují daný práh. Obdobně P_{miss} jako podíl skóre nižších než daný práh, určených pro oprávněné soudy. Je patrné, že se změnou rozhodovacího prahu dochází ke změně hodnot P_{FA} a P_{miss}. Tyto hodnoty se přitom mění v opačném směru a existuje tak mezi nimi závislost (zprostředkovaná přes hodnotu rozhodovacího prahu).

Budeme-li měnit hodnotu rozhodovacího prahu γ v celém možném rozsahu, tj. v intervalu od $-\infty$ do ∞ (hodnota P_{miss} bude postupně růst a P_{FA} klesat), a provedeme zobrazení příslušných hodnot P_{miss} a P_{FA}, dostaneme charakteristiku označovanou jako ROC (Receiver Operating Characteristics) diagram (Fawcett, 2006). Hodnota P_{FA} se vynáší tradičně na vodorovnou osu a hodnota P_{miss} na svislou osu. Příklad ROC diagramu je uveden na obr. 2.2. V případě ROC diagramu je tradičně vynášena na svislou osu hodnota $1 - P_{miss}$, která odpovídá míře úspěšnosti detekce. Zde použitá míra neúspěšnosti detekce však odpovídá významu os v DET diagramu, který bude popsán v následující podkapitole. Koncovými body ROC křivky jsou body $(P_{FA}; P_{miss}) = (0; 1)$ a $(1; 0)$.

Jelikož je počet soudů provedených v rámci experimentálního vyhodnocení systému omezený, je omezený i počet hodnot skóre určených pro tyto soudy. Při vyhodnocení ROC charakteristiky



Obrázek 2.2: *Příklad ROC charakteristik vyhodnocených pro dva ukázkové systémy*

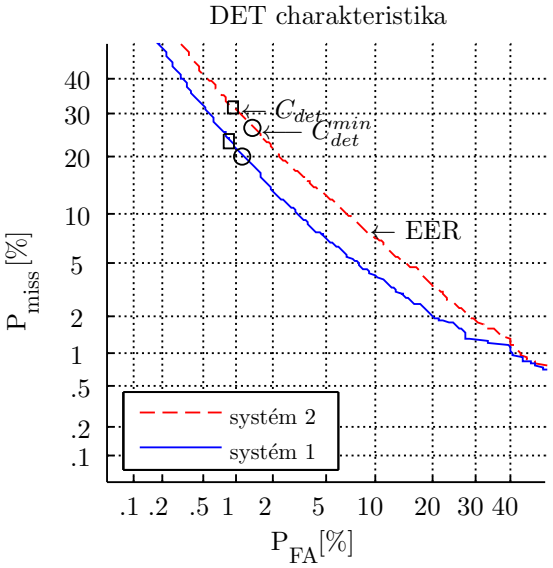
proto není nutné uvažovat všechny možné hodnoty prahu γ v intervalu od $-\infty$ do ∞. Stačí provést vyhodnocení pro hodnoty $-\infty$, ∞ a dále pak již pouze pro jednu hodnotu z intervalů mezi sousedními hodnotami v uspořádané posloupnosti skóre určených detektorem. Z omezeného počtu soudů použitých pro vyhodnocení také vyplývá, že $P_{miss}(\gamma)$ roste schodovitě v diskrétních krocích $\frac{1}{N_{tar}}$ a $P_{FA}(\gamma)$ klesá v diskrétních krocích $\frac{1}{N_{non}}$, kde N_{tar} představuje opět počet oprávněných soudů a N_{non} počet neoprávněných soudů použitých pro evaluaci detektoru. ROC křivka tak není spojitá, ale představuje posloupnost diskrétních bodů $(P_{FA}; P_{miss})$.

2.2.3 DET charakteristika

V rozpoznávání mluvčích je ROC diagram používán zřídka a obvykle je převeden na DET (Detection Error Trade-off) diagram (Martin et al., 1997), viz obr. 2.3. Ten se od ROC diagramu liší především měřítkem os, které odpovídá kvantilové funkci (inverzní distribuční funkci) standardního normálního rozložení. Tato funkce se označuje také jako probit funkce a je definována jako

$$\text{probit}(p) = \sqrt{2}\,\text{erf}^{-1}(2p-1), \tag{2.10}$$

kde p je hodnota P_{FA} nebo P_{miss} a erf^{-1} je inverzní funkce k chybové funkci erf. DET křivka tak vyjadřuje vzájemnou závislost $\text{probit}(P_{FA})$ a $\text{probit}(P_{miss})$. Původní rozsah os pokrývající interval $\langle 0; 1\rangle$ je tak tranformován do intervalu $(-\infty; \infty)$. Je tedy nutné omezit rozsah obou os na konečný interval, který nás zajímá. Dalším důvodem pro omezení rozsahu os je, že v oblastech odpovídajících nízkým chybovým mírám již bude absolutní počet chyb detektoru velmi malý a příslušný odhad chybové míry tak bude nepřesný (viz Doddingtonovo pravidlo 30). Transformace měřítka os diagramu má několik důsledků. Za předpokladu, že skóre určená pro oprávněné a neoprávněné



Obrázek 2.3: *Příklad DET charakteristik vyhodnocených pro stejné systémy, které byly použity v případě obr. 2.2*

soudy mají normální rozložení, DET charakteristika bude reprezentována přibližně lineární křivkou se směrnicí $-\sigma_{non}/\sigma_{tar}$. Dalším pozitivním důsledkem transformace měřítka os je vyšší rozlišení v oblastech nízkých chybových měr. To je výhodné také s ohledem na možnost snazšího porovnání rozdílů charakteristik vynesených pro několik systémů do jednoho diagramu. V případě ROC diagramu dochází zejména v oblastech nízkých chybových měr ke zhuštění křivek a situace se stává nepřehlednou. DET křivka představuje stejně jako ROC křivka spojnici diskrétních bodů a má tedy schodovitý průběh.

Na základě hodnot P_{FA} a P_{miss} vyhodnocených pro zvolený práh γ můžeme v DET diagramu vyznačit pracovní bod systému. Kolem tohoto bodu je někdy vykreslován obdélník udávající 95% interval spolehlivosti odhadu P_{FA} a P_{miss}. Určení tohoto intervalu vychází opět z předpokladu nezávislosti soudů a binomického rozložení chyb detektoru (van Leeuwen et al., 2006). Dalším významným bodem je průsečík DET křivky (popř. ROC křivky) s diagonálou. Tento bod udává hodnotu míry EER (Equal Error Rate) a lze jej definovat jako bod, pro který platí $(P_{FA}; P_{miss}) = (\mathrm{EER}; \mathrm{EER})$. Míra EER tak představuje souhrnnou reprezentaci DET charakteristiky jedinou hodnotou.

DET diagramy se v úloze rozpoznávání mluvčích staly velmi populárním nástrojem pro vyhodnocení systémů a jejich porovnávání. Výhodou je zejména jejich snadná interpretovatelnost. Čím blíže je DET křivka blíže dolnímu levému rohu diagramu, tím lepší je schopnost rozlišovat mezi pravými a nepravými soudy a tím pádem lepší detektor. V případě, že DET křivka vykazuje výrazné odchylky od linearity, může to být nejenom indikátorem špatné práce detektoru, ale také naznačovat chybné reference evaluačních dat. Výhodou DET diagramu je také fakt, že pro vykreslení DET křivky není potřeba volba rozhodovacího prahu. DET diagramy i hodnota EER ale indikují pouze *rozlišovací schopnosti*[2] detektoru.

[2]Ve smyslu schopnosti rozlišovat oprávněné a neoprávněné soudy.

2.2.4 Vyhodnocení binární klasifikace soudů

Volba rozhodovacího prahu musí reflektovat aplikaci, ve které je detektor použit. Určité aplikace mohou vyžadovat velmi nízkou míru chybného přijetí, které lze dosáhnout dostatečným zvýšením rozhodovacího prahu. Musí pak však samozřejmě akceptovat vyšší míru chybného zamítnutí. Zpravidla specifikujeme aplikaci stanovením hodnot:

- P_{tar}: apriorní pravděpodobnosti výskytu oprávněného soudu,
- C_{FA}: pokuty za chybné přijetí neoprávněného soudu a
- C_{miss}: pokuty chybného zamítnutí oprávněného soudu.

Vhodná volba rozhodovacího prahu s ohledem na stanovené parametry zamýšlené aplikace je stejně významnou součástí návrhu detektoru jako vývoj klasifikačních metod.

Na základě uvedených parametrů je definována očekávaná pokuta detekčních chyb C_{det} jako

$$C_{det}(P_{miss}, P_{FA}) = C_{miss}P_{miss}P_{tar} + C_{FA}P_{FA}(1 - P_{tar}), \tag{2.11}$$

kde P_{miss} a P_{FA} představují chybové míry stanovené na základě počtu příslušných chybných rozhodnutí detektoru při zvoleném rozhodovacím prahu. Upozorněme, že hodnota P_{tar} je stanovena (stejně jako hodnoty C_{miss} a C_{FA}) s ohledem na zamýšlenou aplikaci a neodpovídá poměru pravých soudů v evaluačních datech, na základě kterých byly stanoveny hodnoty P_{miss} a P_{FA}. Optimálně zvolený práh je takový, pro který je hodnota detekční pokutové funkce C_{det} minimální. Pro stanovení vhodné hodnoty prahu se používají vývojová data. Hodnota prahu, pro kterou je C_{det} minimální na vývojových datech, však nemusí být shodná s optimální hodnotou pro neviděná evaluační data. Cílem návrhu detektoru mluvčích tedy je, aby skóre poskytovaná systémem byla normalizovaná tak, aby byl práh stanovený na vývojových datech blízký optimální hodnotě prahu pro nová neviděná data.

Kritérium C_{det} bylo dosud použito jako primární metrika pro vyhodnocení systémů pro rozpoznávání mluvčích ve všech ročnících NIST evaluací, tedy již od prvního ročníku pořádaného v roce 1996. Pro zvýšení interpretovatelnosti C_{det} je možné provést normalizaci vydělením hodnotou C_{det0}, která odpovídá pokutě detekčních chyb *referenčního detektoru*. Referenční detektor nebere vůbec v úvahu řečová data a všechny soudy klasifikuje shodně jako oprávněné nebo neoprávněné v závislosti na tom, které rozhodnutí přinese nižší pokutu vzhledem ke stanoveným aplikačním parametrům. Platí tedy:

$$C_{det0} = \min\{C_{miss}P_{tar}, C_{FA}(1 - P_{tar})\}. \tag{2.12}$$

2.2.5 Význam kalibrace detektoru

Pokutová funkce C_{det} charakterizuje současně rozlišovací schopnosti i *kalibraci* detektoru. K dosažení nízké hodnoty je nutná nejenom nízká hodnota EER (příp. DET křivka blízká dolnímu levému

rohu diagramu), ale také správný detekční práh zvolený s ohledem na stanovené parametry zamýšlené aplikace. Na základě znalosti referencí evaluačních dat je možné stanovit optimální hodnotu rozhodovacího prahu, pro kterou je C_{det} minimální, na základě kritéria

$$C_{det}^{min} = \min_{-\infty \leq \gamma \leq \infty} C_{miss} P_{miss}(\gamma) P_{tar} + C_{FA} P_{FA}(\gamma)(1 - P_{tar}). \tag{2.13}$$

Při vyhodnocení tak dochází k přizpůsobení se hodnotám skóre poskytovaných systémem a na rozdíl od C_{det} nejsou hodnocena pevná rozhodnutí detektoru. Hodnota C_{det}^{min} tak reflektuje pouze rozlišovací schopnosti detektoru podobně jako hodnota EER. Na rozdíl od EER je však hodnota C_{det}^{min} závislá na stanovených parametrech zamýšlené aplikace. Bod odpovídající hodnotě C_{det}^{min} bývá vyznačen v DET diagramu pro porovnání ideálního pracovního bodu s pracovním bodem definovaným zvoleným rozhodovacím prahem γ (viz obr. 2.3). Je zřejmé, že platí $C_{det}^{min} \leq C_{det}$.

Význam C_{det}^{min} spočívá především v možnosti posouzení podílu vlivu rozlišovacích schopností a kalibrace detektoru na hodnotě C_{det}. Ztráta způsobená chybnou kalibrací je dána rozdílem $C_{det} - C_{det}^{min}$.

Přestože definice reálných aplikací pomocí trojice parametrů $(P_{tar}, C_{miss}, C_{FA})$ je přirozená, z pohledu vyhodnocení nese tato trojice redundantní informace. Funkci C_{det} je namísto uvedené trojice možné parametrizovat trojicí $(\widetilde{P}_{tar}, \widetilde{C}_{miss} = 1, \widetilde{C}_{FA} = 1)$, kde $\widetilde{P}_{tar}$ označíme jako efektivní apriorní pravděpodobnost, která zahrnuje stanovené pokuty C_{miss} a C_{FA}. V důsledku jednotkových pokut $\widetilde{C}_{miss}$ a $\widetilde{C}_{FA}$ můžeme pokutovou funkci (2.11) interpretovat jako očekávanou souhrnnou chybovou míru P_{err}, tedy

$$P_{err}(P_{miss}, P_{FA}) = \widetilde{P}_{tar} P_{miss} + (1 - \widetilde{P}_{tar}) P_{FA}, \tag{2.14}$$

kde hodnotu efektivní apriorní pravděpodobnosti určíme dle vztahu

$$\widetilde{P}_{tar} = \frac{P_{tar} C_{miss}}{P_{tar} C_{miss} + (1 - P_{tar}) C_{FA}}. \tag{2.15}$$

Protože platí $C_{det} = k P_{err}$, kde k je pozitivní koeficient, poskytují obě kritéria ekvivalentní vyhodnocení systémů. Ekvivalence vyhodnocení znamená, že všechny porovnávané systémy budou při vyhodnocení prováděném na základě obou kritérií seřazeny ve shodném pořadí. Konkrétně platí

$$k \ = P_{tar} C_{miss} + (1 - P_{tar}) C_{FA}. \tag{2.16}$$

Ve dvanácti ročnících NIST evaluací uspořádaných v průběhu let 1996 až 2008 byly vždy specifikovány aplikační parametry $C_{miss} = 10$, $C_{FA} = 1$ a $P_{tar} = 0{,}01$ (připomeňme, že tato hodnota je zvoleným parametrem zamýšlené aplikace a nesouvisí nijak s podílem oprávněných soudů v evaluačních datech). Těmto parametrům odpovídá hodnota efektivní apriorní pravděpodobnosti $\widetilde{P}_{tar} \approx 0{,}092$. V prozatím posledním ročníku pořádaném v roce 2010 (NIST, 2010) byly aplikační parametry změněny na $C_{miss} = 1$, $C_{FA} = 1$ a $P_{tar} = 0{,}001$, kterým odpovídá hodnota $\widetilde{P}_{tar} = 0{,}001$.

2.2.6 Skóre vs. logaritmus poměru věrohodností

Připomeňme, že jediným dosud kladeným požadavkem na skóre poskytovaná detektorem pro předložené soudy byla vyšší hodnota skóre pro soudy klasifikované jako oprávněné a nižší hodnota pro soudy klasifikované jako neoprávněné. Pokud bude ke všem skóre přičtena libovolná hodnota nebo budou skóre vynásobena libovolným koeficientem, nebudou hodnoty EER a C_{det}^{min} ani DET křivka touto transformací nijak ovlivněny. Bude-li nastavený rozhodovací práh transformován stejným způsobem jako hodnoty skóre, nebude ovlivněna ani hodnota C_{det}. Uvedené evaluační metriky nebudou ovlivněny v případě transformace skóre libovolnou ryze rostoucí funkcí. Jediné co musí být zachováno je pořadí skóre a konkrétní hodnoty tak nemají žádný význam.

Změní-li se parametry aplikace nebo chceme-li detektor použít v nové aplikaci, je nutné vyzkoušet chování detektoru v nových podmínkách pro zjištění příslušného rozhodovacího prahu. Jeho hodnota je samozřejmě závislá na stanovených parametrech aplikace (P_{tar}, C_{FA} a C_{miss}), jejich vztah však není možné nijak vyjádřit. V případě, že pak rozhodnutí detektoru působí jako příliš jednostranná, není možné posoudit, zda to není způsobeno nepatřičnou volbou parametrů aplikace.

Podívejme se nyní, jakým způsobem je možné použít skóre s určené detektorem pro předložený soud, abychom provedli optimální klasifikaci tohoto soudu. Očekávaná pokuta klasifikace soudu jako oprávněného je $(1-P(tar|s))C_{FA}$, zatímco očekávaná pokuta klasifikace soudu jako neoprávněného je $P(tar|s)C_{miss}$, kde $P(tar|s)$ představuje rozložení posteriorní pravděpodobnosti, že předložený soud je oprávněný, bylo-li vyhodnoceno skóre s. Rozhodnutí o klasifikaci soudu, vedoucí k minimální očekávané pokutě, je známo jako Bayesovské rozhodnutí. Pokud bude Bayesovské rozhodnutí učiněno pro všechny předložené soudy, bude minimalizována i celková očekávaná pokuta detekčních chyb přes všechny soudy, tedy hodnota C_{det}. Předložený soud klasifikujeme jako oprávněný, pokud platí

$$(1 - P(tar|s))C_{FA} \leq P(tar|s)C_{miss}. \tag{2.17}$$

Pro $P(tar|s)$ na základě Bayesovy věty platí

$$\begin{aligned} P(tar|s) &= \frac{P(s|tar)P_{tar}}{P(s)} \\ &= \frac{P(s|tar)P_{tar}}{P(s|tar)P_{tar} + P(s|non)(1 - P_{tar})}. \end{aligned} \tag{2.18}$$

Toto vyjádření je však poněkud nepraktické a proto vyjádříme $P(tar|s)$ pomocí funkce logit[3] a dostáváme

$$\operatorname{logit} P(tar|s) = \mathcal{L}(s) + \operatorname{logit} P_{tar}, \tag{2.19}$$

kde

$$\mathcal{L}(s) = \log \frac{P(s|tar)}{P(s|non)} \tag{2.20}$$

[3] Je-li p pravděpodobnost, pak $\frac{p}{1-p}$ je příslušná hodnota *šance*. Funkce logit je definována jako $\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right)$ a odpovídá tedy logaritmu šance.

představuje *logaritmus poměru věrohodností*[4] LLR (log-likelihood-ratio) pro skóre s. Rozhodovací pravidlo pro klasifikaci soudů (2.17) je možné s využitím vztahu (2.19) vyjádřit v závislosti na hodnotě $\mathcal{L}(s)$ stanovené detektorem. Předložený soud je klasifikován v závislosti na podmínce:

$$\mathcal{L}(s) \begin{cases} \geq -\widetilde{\gamma} & \text{jako } \textit{oprávněný} \text{ soud a} \\ < -\widetilde{\gamma} & \text{jako } \textit{neoprávněný} \text{ soud,} \end{cases} \tag{2.21}$$

kde $\widetilde{\gamma}$ představuje rozhodovací práh stanovený na základě parametrů aplikace jako

$$\begin{aligned} \widetilde{\gamma} &= \log\left(\frac{P_{tar}}{(1-P_{tar})}\frac{C_{miss}}{C_{FA}}\right) \\ &= \operatorname{logit}\left(\widetilde{P}_{tar}\right). \end{aligned} \tag{2.22}$$

Významným důsledkem (2.21) je oddělení hodnoty poskytované detektorem pro předložený soud, reprezentované $\mathcal{L}(s)$, a rozhodovacího prahu $\widetilde{\gamma}$, který je nyní funkcí parametrů zamýšlené aplikace. Účelem skóre s je pouze reprezentovat informace související s úlohou detekce mluvčích, které se podařilo získat z předložených řečových dat. $\mathcal{L}(s)$ je pak hodnota skóre transformovaná do podoby, která vyhovuje standardizovanému procesu rozhodování o klasifikaci soudů. Hovoříme o *kalibraci* skóre. Informace reprezentovaná hodnotou $\mathcal{L}(s)$ je tzv. *aplikačně nezávislá*, protože veškeré aplikačně závislé parametry jsou zahrnuty v rozhodovacím prahu $\widetilde{\gamma}$.

2.2.7 Aplikačně nezávislé vyhodnocení

Pokud výstup detektoru skutečně odpovídá logaritmu poměru věrohodností, není nutné požadovat po detektoru, aby poskytoval pevná rozhodnutí o klasifikaci soudů. Rozhodovací práh lze totiž určit nezávisle na použitém detektoru na základě parametrů aplikace podle vztahu (2.22). Pro zvolený práh $\widetilde{\gamma}$ je pak možné na základě hodnot $\mathcal{L}(s)$ určených pro jednotlivé soudy provést pevná rozhodnutí o jejich klasifikaci nezávisle na detektoru. Na základě těchto rozhodnutí lze pak stanovit míry P_{miss} a P_{FA} a vyhodnotit příslušnou hodnotu C_{det} nebo P_{err}. Samozřejmě můžeme vyhodnotit také hodnotu C_{det}^{min}, která by v ideálním případě (pro perfektně kalibrovaný systém) měla být pro libovolně zvolené aplikační parametry shodná s hodnotou C_{det}. Tímto způsobem však provedeme vyhodnocení kvality poskytovaných LLR pouze z pohledu zvoleného pracovního bodu definovaného hodnotou $\widetilde{\gamma}$. Po systému však požadujeme, aby poskytoval správné hodnoty v celém možném rozsahu pracovních bodů. van Leeuwen – Brummer (2007) tedy definují nové kritérium pro souhrnné vyhodnocení detektoru mluvčích na základě poskytovaných hodnot LLR[5]:

$$C_{llr} = C_0 \int_{-\infty}^{\infty} P_{err}\left(P_{miss}(\widetilde{\gamma}), P_{FA}(\widetilde{\gamma}), \widetilde{\gamma}\right) d\widetilde{\gamma}, \tag{2.23}$$

kde $C_0 > 0$ je normalizační konstanta. Protože P_{err} hodnotí rozhodnutí detektoru o klasifikaci soudů, reflektuje funkce C_{llr} jak rozlišovací schopnosti, tak kvalitu kalibrace detektoru. Kritérium

[4] Poměr věrohodností není symetrický a malá skóre jsou navíc koncentrována kolem 0, proto je použit logaritmus.

[5] Toto kritérium bývá označováno jako log-likelihood-ratio cost function

C_{llr} je možné interpretovat jako celkovou chybovou míru vyhodnocenou přes celý rozsah možných pracovních bodů. Pro dosažení nízké hodnoty C_{llr} je nutné, aby byla nízká hodnota EER a systém poskytoval dobře kalibrované hodnoty $\mathcal{L}(s)$ v celém rozsahu pracovních bodů. C_{llr} tak poskytuje *aplikačně nezávislé* vyhodnocení.

Definice C_{llr} podle (2.23) je dobře interpretovatelná z hlediska významu, integrál ale představuje komplikaci z hlediska praktického výpočtu. Naštěstí je však možné ukázat (Brummer, 2010a), že C_{llr} lze vyjádřit analyticky v uzavřené formě jako

$$C_{llr}(\{\mathcal{L}'_t\}) = \frac{1}{2\log 2}\left(\frac{1}{N_{tar}}\sum_{t\in\text{tar}}\log\left(1+e^{-\mathcal{L}'_t}\right)+\frac{1}{N_{non}}\sum_{t\in\text{non}}\log\left(1+e^{\mathcal{L}'_t}\right)\right), \tag{2.24}$$

kde hodnota $\mathcal{L}'_t$ je výsledkem snahy detektoru o určení hodnoty LLR pro soud t, „tar“ reprezentuje soubor všech N_{tar} oprávněných soudů a „non“ soubor všech N_{non} neoprávněných soudů předložených ve vyhodnocení. Uveďme nyní tři případy hodnot, kterých nabývá člen $\log\left(1+e^{-\mathcal{L}'_t}\right)$ v závislosti na stanovené hodnotě $\mathcal{L}'_t$ pro oprávněný soud t:

- dobrý detektor přiřadí oprávněnému soudu hodnotu $\mathcal{L}'_t \gg 1$ a tedy člen $\log\left(1+e^{-\mathcal{L}'_t}\right) \approx 0$, jeho příspěvek k celkové hodnotě C_{llr} vymizí;
- špatný detektor přiřadí oprávněnému soudu hodnotu $\mathcal{L}'_t \ll -1$ a tedy člen $\log\left(1+e^{-\mathcal{L}'_t}\right) \approx -\mathcal{L}'_t$, jeho příspěvek tedy lineárně (neomezeně) roste s hodnotou $|\mathcal{L}'_t|$;
- jako *referenční* označíme detektor, který přiřadí soudu neutrální hodnotu $\mathcal{L}'_t = 0$ a tedy člen $\log\left(1+e^{-\mathcal{L}'_t}\right) = \log 2$.

Obdobně lze ukázat chování druhého součtového členu $\log\left(1+e^{\mathcal{L}'_t}\right)$ v závislosti na hodnotách $\mathcal{L}'_t$ určených detektorem pro neoprávněné soudy. V případě referenčního detektoru, odpovídá příspěvek obou členů hodnotě $\log 2$. S ohledem na zvolenou normalizační konstantu v (2.24) tak tomuto detektoru odpovídá referenční hodnota $C_{llr} = 1$.

Stejně jako nás v případě funkce C_{det} zajímalo, jakým způsobem se na dané hodnotě podílí rozlišovací schopnosti a kalibrace detektoru, zajímá nás analogicky vliv těchto složek i v případě funkce C_{llr}. Připomeňme, že C_{det} provádí vyhodnocení rozhodnutí o klasifikaci soudů a hodnota C_{det}^{min} odpovídá klasifikaci podle ideálního rozhodovacího prahu nalezeného na základě znalosti referenční klasifikace testovacích dat. Funkce C_{llr} provádí vyhodnocení LLR hodnot $\mathcal{L}'_t$ stanovených systémem. Kritérium C_{llr}^{min} odpovídá vyhodnocení transformovaných hodnot $\mathcal{L}''_t = w(\mathcal{L}'_t)$, kde w představuje ryze rostoucí transformační funkci, která je nalezena na základě znalosti referenční klasifikace testovacích dat tak, aby byla minimalizována funkce C_{llr}. Platí tedy $C_{llr}^{min} = \min_w C_{llr}(\{w(\mathcal{L}'_t)\})$. Transformační funkce je omezena na ryze rostoucí z následujících důvodů:

- Aplikace transformace nemění setřídění hodnot odpovídajících jednotlivým soudům.
- Jednomu rozhodovacímu prahu pro $\mathcal{L}'_t$ odpovídá jeden rozhodovací práh pro $\mathcal{L}''_t$.

- DET charakteristika je invariantní vůči transformaci skóre ryze monotónní funkcí.
- Ryze monotónní funkce je invertovatelná a její aplikace tak mění pouze formální podobu (kalibraci) skóre, nikoliv jeho významovou složku (rozlišovací schopnosti).

Požadavek monotónně rostoucí funkce nijak nebrání tomu, aby byla každá hodnota $\mathcal{L}'_t$ optimalizována nezávisle a transformační funkce byla definována neparametrickým způsobem. Efektivní způsob nalezení transformační funkce umožňuje PAV algoritmus[6] (Ayer et al., 1955). Tento algoritmus hledá optimální zobrazení vstupních hodnot na hodnoty posteriorních pravděpodobností. Výsledkem je maximálně věrohodné řešení problému, které je navíc unikátní. Posteriorní pravděpodobnosti lze jednoduchým způsobem převést na hodnoty ve formátu LLR.

Na základě znalosti C_{llr}^{min} je možné provést rozklad C_{llr} na ztrátu způsobenou rozlišovacími schopnostmi, která odpovídá právě hodnotě C_{llr}^{min}, a ztrátu způsobenou špatnou kalibrací, která odpovídá rozdílu $C_{llr} - C_{llr}^{min}$. Tento rozdíl je vždy nezáporný. Pro dobře kalibrovaný systém se blíží nule a neomezeně roste se zvyšující se mírou špatné kalibrace.

2.2.8 PAV algoritmus

Vstupem PAV (Pool Adjacent Violators) algoritmu (Wilbur et al., 2005; Fawcett – Niculescu-Mizil, 2007; Brummer, 2010a) jsou dvojice hodnot $(\mathcal{L}'_t, y_t)$, tedy hodnoty $\mathcal{L}'_t$ určené systémem pro předložené soudy t a jejich referenční klasifikace $y_t \in \{0, 1\}$, kde $t = 1, \ldots T$. Výstupem jsou hodnoty posteriorní pravděpodobnosti p_t, které odpovídají hypotéze, že soud t je oprávněný.

Na začátku činnosti algoritmu jsou dvojice hodnot $(\mathcal{L}'_t, y_t)$ uspořádány ve vzestupném pořadí podle hodnoty $\mathcal{L}'_t$, takže platí $\mathcal{L}'_1 < \mathcal{L}'_2 < \ldots < \mathcal{L}'_T$. Pro následující činnost algoritmu již hodnoty $\mathcal{L}'_i$, kde $i = 1, \ldots T$, nemají význam a příslušná sekvence hodnot y_i se použije pro inicializaci hodnot p_i. Je zřejmé, že sekvence p_i není za reálných podmínek[7] monotónní. V průběhu výkonu algoritmu jsou iterativně vyhledávány sousední hodnoty narušující monotónní průběh a jejich hodnota je upravována. Nyní označíme podsekvenci začínající indexem i a končící indexem j jako $G_{i:j}$ a dále označíme aktuální hodnotu posteriorní pravděpodobnosti společnou pro všechny členy podsekvence jako $p_{i:j}$, platí tedy $p_{i:j} = p_i = p_{i+1} = \ldots = p_j$. Pokud existují podsekvence $G_{i:k}$ a $G_{k+1:j}$ takové, že $p_{i:k} \geq p_{k+1:j}$, provede se sloučení obou těchto podsekvencí a příslušná hodnota $p_{i:j}$ se určí jako

$$p_{i:j} = \frac{\sum_{k=i}^{j} y_k}{j - i + 1}. \tag{2.25}$$

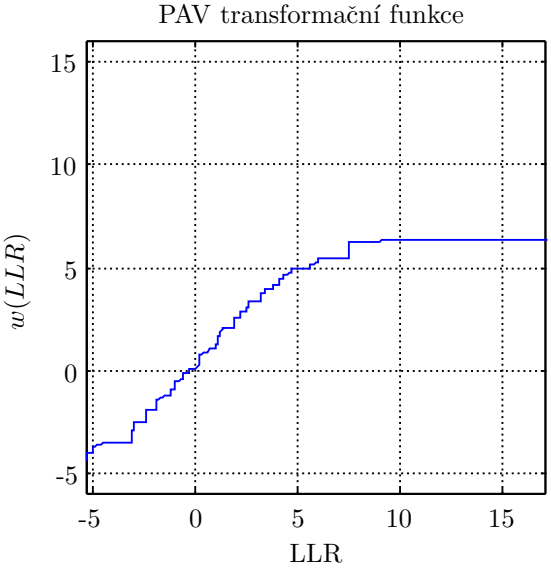
Činnost algoritmu končí v okamžiku, kdy je dosaženo monotónnosti celé sekvence. Přímá implementace popsaného řešení má nelineární časovou náročnost (bez zahrnutí času potřebného na setřídění sekvence před zahájením činnosti algoritmu). Existuje však i efektivní verze PAV algoritmu, která má lineární časovou náročnost v závislosti na počtu soudů T, viz algoritmus 1.

[6] PAV optimalizační algoritmus řeší problém izotonické regrese.

[7] Tedy vyjma nereálného detektoru, který by byl schopen bezchybně rozlišovat oprávněné a neoprávněné soudy podle vhodně zvoleného rozhodovacího prahu. Jinými slovy vyjma situace, kdy nedochází k žádnému překryvu distribucí skóre určených systémem pro oprávněné a neoprávněné soudy.

Algoritmus 1 PAV algoritmus

Setřídění dvojic $(\mathcal{L}'_t, y_t)$ ve vzestupném pořadí podle $\mathcal{L}'_t$, takže platí $\mathcal{L}'_1 < \mathcal{L}'_2 < \ldots < \mathcal{L}'_T$
$g = 1$, $start_g = 1$, $length_g = 1$, $p^*_g = y_1$
for $k = 2 \to T$ **do**
  $g = g + 1$, $start_g = k$, $length_g = 1$, $p^*_g = y_k$
  while $g \geq 2$ **and** $p^*_{\max(g-1,1)} \geq p^*_g$ **do**
    $nw = length_{g-1} + length_g$
    $p^*_{g-1} = p^*_{g-1} + \frac{length_g}{nw}(p^*_g - p^*_{g-1})$
    $length_{g-1} = nw$
    $g = g - 1$
  end while
end for
Přiřazení výstupních hodnot v závislosti na příslušnosti k jednotlivým subsekvencím
$n = T$
while $n \geq 1$ **do**
  for $k = start_g \to n$ **do**
    $p_k = p^*_g$
  end for
  $n = start_g - 1$, $g = g - 1$
end while



Obrázek 2.4: *Příklad transformační funkce nalezené pomocí PAV algoritmu*

Výsledek PAV algoritmu ovšem není zcela ideální. Výsledkem je funkce provádějící zobrazení skóre poskytovaných systémem na hodnoty posteriorní pravděpodobnosti, zatímco naším cílem bylo získat zobrazení na hodnoty odpovídající LLR hodnotám. Navíc sekvence výstupních posteriorních

pravděpodobností nesplňuje požadavek ryzí monotónnosti, protože platí $p_1 \leq p_2 \leq \ldots \leq p_T$.

První problém lze snadno vyřešit aplikací Bayesovy věty, takže platí $\mathcal{L}''_t = \text{logit}(p_t) - \text{logit}(\frac{N_{tar}}{T})$. V případě druhého problému lze ukázat, že přestože nalezené řešení nepatří do prostoru přípustných transformačních funkcí, který je definován podmínkou ryzí monotónnosti. Leží řešení nalezené pomocí PAV algoritmu na hranici tohoto prostoru (Brummer, 2010a). Protože bylo naším cílem nalézt transformační funkci $w(\mathcal{L}')$ minimalizující $C_{llr}(\{w(\mathcal{L}'_t)\})$, lze nalezené řešení interpretovat jako infimum namísto minima optimalizačního problému. Skutečně pak platí $C_{llr}^{min} = C_{llr}(\{PAV(\mathcal{L}'_t, y_t)\})$. Výpočet hodnoty C_{llr}^{min} podle (2.24) pro ideálně kalibrované hodnoty $\mathcal{L}''$ není nijak ovlivněn rozporem s požadavkem ryzí monotónnosti. V případě, že bychom však prováděli vyhodnocení rozlišovacích schopností detektoru (například pomocí DET charakteristiky) na základě hodnot $\mathcal{L}''$, nebude výsledek shodný s původním vyhodnocením pro $\mathcal{L}'$. V praxi lze tento problém vyřešit například přičtením malé hodnoty v závislosti na pořadí. Obr. 2.4 znázorňuje příklad transformační funkce $w(\mathcal{L}')$ nalezené pomocí PAV algoritmu.

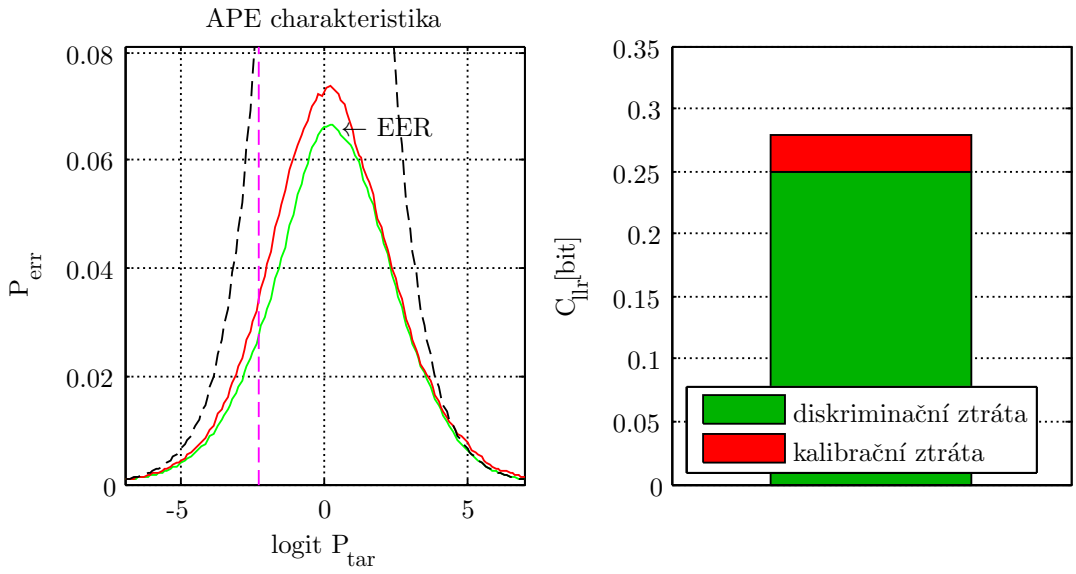
2.2.9 APE charakteristika

Souhrnné skalární vyjádření rozlišovacích schopností a kvality kalibrace přes všechny pracovní body detektoru pomocí funkce C_{llr} umožňuje okamžité porovnání systémů. Navíc je ohodnocení systémů pomocí skalárního kritéria nezbytné například pro definici optimalizačního kritéria při hledání parametrů kalibrační transformace. Na druhé straně skalární vyjádření přirozeně neposkytuje informace o chování detektoru v jednotlivých pracovních bodech. Za tímto účelem Brummer – du Preez (2006) navrhli APE (Applied Probability of Error) charakteristiku. Základem této charakteristiky je znázornění závislosti P_{err} na hodnotě rozhodovacího prahu $\widetilde{\gamma}$, který lze podle (2.22) určit jako logaritmus šance odpovídající efektivní apriorní pravděpodobnosti $\widetilde{P}_{tar}$. Protože rozsah hodnot $\widetilde{\gamma}$ není nijak omezen, představuje vodorovnou osu charakteristiky celá reálná osa. V praxi je vykreslován pouze omezený interval v blízkosti nulové hodnoty $\widetilde{\gamma}$. Důvodem pro omezení zobrazovaného intervalu je především nízký počet chyb v oblastech $|\widetilde{\gamma}| \gg 1$, který neumožňuje spolehlivý odhad chybové míry. Jedná se tedy o analogický důvod, který vede k omezení zobrazovaného intervalu v případě DET charakteristiky. Příklad této charakteristiky ilustruje obr. 2.5. APE diagram zahrnuje charakteristiku následujících systémů:

- skutečný detektor – znázorněn červenou čarou. Křivka představuje chybovou míru P_{err} odpovídající provedení rozhodnutí o klasifikaci soudů na základě hodnot $\mathcal{L}'$. Plocha pod křivkou (v celém intervalu pracovních bodů, nejenom pod znázorněných intervalem) je úměrná hodnotě C_{llr}[8].

- ideálně kalibrovaný detektor – znázorněn zelenou čarou. Křivka představuje chybovou míru P_{err}^{min}, která odpovídá provedení rozhodnutí o klasifikaci soudů na základě ideálně kalibrovaných hodnot $\mathcal{L}''$ (nalezených s použitím PAV algoritmu). Plocha pod křivkou je úměrná

[8] Je nutné brát v úvahu normalizační konstantu C_0 z (2.23).



Obrázek 2.5: *Příklad APE diagramu*

hodnotě C_{llr}^{min}. Plocha mezi plnou a čerchovanou čarou odpovídá celkové kalibrační ztrátě. Křivka má globální maximum, které odpovídá hodnotě EER.

- referenční detektor – znázorněn čerchovanou čarou. Připomeňme, že se jedná o detektor, který pro své rozhodnutí nebere vůbec v úvahu řečová data a každému předloženému soudu přidělí hodnotu $\mathcal{L}'_t = 0$. Detektor se rozhoduje o binární klasifikaci soudů pouze na základě apriorní pravděpodobnosti $\widetilde{P}_{tar}$. Očekávanou chybovou míru referenčního detektoru označíme jako P_{err0} a lze ji jednoduše určit jako $P_{err0}(\widetilde{\gamma}) = \min\{\widetilde{P}_{tar}(\widetilde{\gamma}), 1 - \widetilde{P}_{tar}(\widetilde{\gamma})\}$. Křivka dosahuje maximální hodnoty 0,5 pro $\widetilde{\gamma} = 0$. Plocha pod touto křivkou je úměrná jedné (se stejným koeficientem úměrnosti jaký platí pro ostatní křivky).

Svislá čára odpovídá pracovnímu bodu příslušnému $\widetilde{P}_{tar} = 0{,}092$. Hodnota odečtená v průsečíku s červenou čarou je úměrná hodnotě C_{det} a hodnota odečtená v průsečíku se zelenou čarou hodnotě C_{det}^{min} s koeficientem úměrnosti určeným dle (2.16).

Pro úplnost uveďme, že kritérium P_{err}^{min} lze také definovat analogicky k definici C_{det}^{min} podle (2.13) jako:

$$P_{err}^{min}(\widetilde{P}_{tar}) = \min_{-\infty \leq \widetilde{\gamma} \leq \infty} \widetilde{P}_{tar} P_{miss}(\widetilde{\gamma}) + (1 - \widetilde{P}_{tar}) P_{FA}(\widetilde{\gamma}). \tag{2.26}$$

Z této definice je patrné, že platí $P_{err}^{min}(\widetilde{P}_{tar}) \leq P_{err}(\widetilde{P}_{tar})$. Pokud jsou si obě hodnoty blízké, je systém dobře kalibrovaný. Od dobrého detektoru požadujeme, aby byl dobře kalibrovaný v celém rozsahu pracovních bodů. Z předpokladu, že všechna skóre poskytovaná systémem jsou konečná, vyplývá, že je vždy možné zvolit rozhodovací práh tak, aby jedna z chybových měr P_{miss} nebo P_{FA} byla nulová, a proto platí $P_{err}^{min}(\widetilde{P}_{tar}) \leq P_{err0}(\widetilde{P}_{tar})$. Z tohoto pozorování lze vyvodit závěr, že pokud je úspěšnost detektoru ve srovnání s referenčním detektorem nižší, je to způsobeno pouze špatnou kalibrací.

Z APE charakteristiky je zřejmé, že P_{err} v oblasti okolo $\widetilde{\gamma} = 0$ (tj. $\widetilde{P}_{tar} = 0{,}5$) přispívá k hodnotě C_{llr} největší měrou. To by naznačovalo, že úloha detekce mluvčích je v této oblasti nejtěžší a to i s ohledem na kalibraci skóre. Na druhé straně je však největší přínos použití detektoru ve srovnání s referenčním detektorem dosažen také právě v této oblasti. Zatímco pro $|\widetilde{\gamma}| \gg 1$ je již velké množství informace obsaženo v apriorní pravděpodobnosti a je obtížné získat na základě analýzy řečového signálu další užitečnou informaci (přestože je pravděpodobnost chyby v těchto oblastech nižší). V případě APE charakteristiky znázorněné na obr. 2.5 je možné pozorovat, že pro $\widetilde{\gamma} > 5$ poskytuje ukázkový systém horší výsledky než referenční detektor.

APE charakteristika zprostředkovává podobně jako DET charakteristika informaci o chování detektoru v závislosti na pracovním bodě. Nepředstavuje však náhradu DET charakteristiky, ale nabízí komplementární informace. DET charakteristika je vhodným nástrojem především pro analýzu a porovnání rozlišovacích schopností detektorů, je z ní také dobře zřejmý vzájemný vztah mezi P_{miss} a P_{FA}. APE charakteristika je vhodná především pro analýzu kvality kalibrace detektoru. Obě charakteristiky nabízejí některé společné informace, z obou lze například odečíst hodnotu EER.

Nepříjemnou vlastností APE charakteristiky je nízká rozlišovací schopnost v oblastech $|\widetilde{\gamma}| \gg 1$. To představuje problém například s ohledem na nově (v roce 2010) definovaný pracovní bod v NIST evaluacích, s ním se totiž přesunula pozornost na úspěšnost systémů při nízké apriorní pravděpodobnosti výskytu oprávněných soudů.

2.2.10 NBE charakteristika

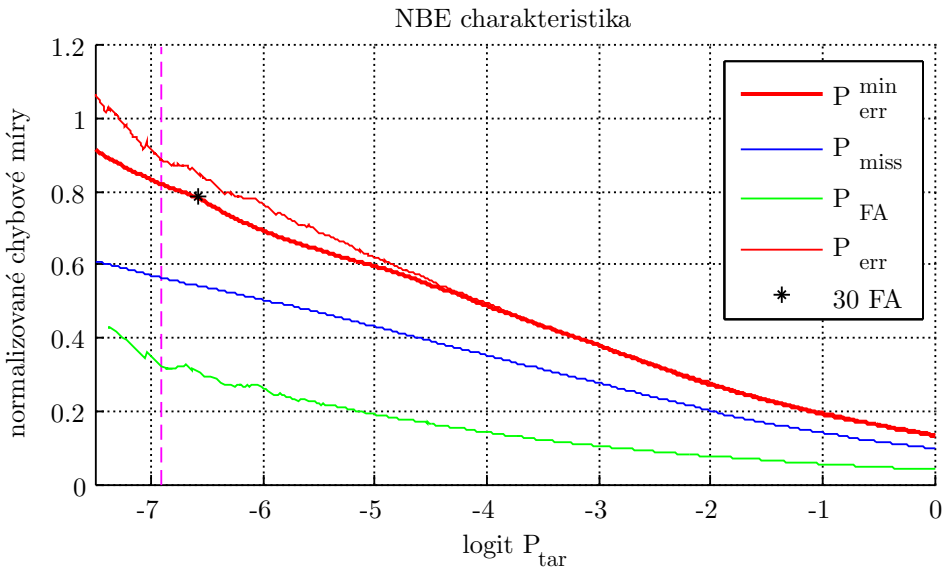
Vhodným nástrojem pro analýzu systémů i v oblastech $\widetilde{\gamma} \ll -1$ je NBE (Normalized Bayes Error-rate) charakteristika (Brummer, 2010a). Tato charakteristika opět sleduje kritérium P_{err} jako funkci efektivní apriorní pravděpodobnosti $\widetilde{P}_{tar}$. Pro možnost nezávislého posouzení vlivu rozlišovacích schopností a kvality kalibrace na hodnotě P_{err} je sledována i hodnota P_{err}^{min}. Rozdíl ve srovnání s APE charakteristikou spočívá především v normalizaci obou kritérií hodnotou P_{err0}. Je možné jednoduše ukázat, že pro normalizované hodnoty platí

$$\frac{P_{err}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})} \geq \frac{P_{err}^{min}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})} \qquad \text{a také} \qquad 1 \geq \frac{P_{err}^{min}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})}. \tag{2.27}$$

Příklad NBE charakteristiky je znázorněn na obr. 2.6. Protože nás zajímá chování systému v oblasti nízkých hodnot $\widetilde{P}_{tar}$ je oproti APE diagramu rozsah vodorovné osy omezen na oblast $\widetilde{P}_{tar} < 0{,}5$, resp. $\widetilde{\gamma} < 0$. Pro tuto oblast lze vyjádřit normalizovanou hodnotu $P_{err}(\widetilde{P}_{tar})$ následujícím způsobem:

$$\begin{aligned} \frac{P_{err}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})} &= \frac{\widetilde{P}_{tar} P_{miss}(\widetilde{P}_{tar}) + (1 - \widetilde{P}_{tar}) P_{FA}(\widetilde{P}_{tar})}{\widetilde{P}_{tar}} \\ &= P_{miss}(\widetilde{P}_{tar}) + \exp(-\operatorname{logit} \widetilde{P}_{tar}) P_{FA}(\widetilde{P}_{tar}) \\ &= P_{miss}(\operatorname{logit}^{-1} \widetilde{\gamma}) + \exp(-\widetilde{\gamma}) P_{FA}(\operatorname{logit}^{-1} \widetilde{\gamma}). \end{aligned} \tag{2.28}$$

Efektem normalizace je tedy exponenciální zesílení chybové míry P_{FA}. Ve srovnání s APE charakteristikou je díky tomuto zesílení efekt kalibrace dobře pozorovatelný i v oblastech nízkých hodnot



Obrázek 2.6: *Příklad NBE diagramu*

P_{FA} a lze tak zobrazit širší rozsah pracovních bodů. Svislá čára v obr. 2.6 odpovídá pracovnímu bodu příslušnému $\widetilde{P}_{tar} = 0{,}001$.

Stále je však nutné dbát na spolehlivost odhadu chybových měr, která je závislá na absolutním počtu zaznamenaných chyb pro daný pracovní bod. V NBE diagramu je proto vyznačen bod, který odpovídá 30 chybným přijetím. Bod je vyznačen na křivce odpovídající normalizovanému kritériu P_{err}^{min}, protože je brán v úvahu počet chybných přijetí při optimální volbě rozhodovacího prahu (resp. za předpokladu ideální kalibrace detektoru).

2.2.11 Vyhodnocení množství informace poskytované detektorem

Ještě před zpracováním řečových dat detektorem je nejistota o klasifikaci soudu podmíněná pouze apriorní znalosti, která je představována hodnotou apriorní pravděpodobnosti výskytu oprávněného soudu $\widetilde{P}_{tar}$. V krajních případech, kdy $\widetilde{P}_{tar} = 0$ nebo $\widetilde{P}_{tar} = 1$, není žádná nejistota o klasifikaci soudů a detektor tuto klasifikaci nemůže ovlivnit. Posteriorní pravděpodobnost bude v těchto případech rovna 0 nebo 1. Informační (Shannonova) entropie vyjadřuje nejistotu o klasifikaci předloženého soudu v jednotkách bit. Entropie je vzhledem k apriorní pravděpodobnosti konkávní funkce a maxima 1 bitu dosahuje pro $\widetilde{P}_{tar} = 0{,}5$. Hodnota entropie je tedy maximální, když je maximální nejistota o klasifikaci soudů. Entropie referenčního detektoru[9] je stanovena v závislosti na $\widetilde{P}_{tar}$ jako (Bishop, 2006):

$$H(\widetilde{P}_{tar}) = -\widetilde{P}_{tar} \log_2 \widetilde{P}_{tar} - (1 - \widetilde{P}_{tar}) \log_2 (1 - \widetilde{P}_{tar}). \tag{2.29}$$

Cílem detektoru je získat z řečových dat užitečnou informaci, která by pomohla snížit nejistotu o klasifikaci soudů. Průměrnou entropii přes posteriorní pravděpodobnosti určené (na

[9] Připomeňme, že se jedná o detektor, který nebere v úvahu řečová data a hodnota posteriorní pravděpodobnosti je tak shodná s apriorní pravděpodobností.

základě poměrů věrohodností stanovených detektorem a dané apriorní pravděpodobnosti) skutečným detektorem pro všechny soudy předložené ve vyhodnocení je možné vyjádřit jako (Ramos – Gonzalez-Rodriguez, 2008)[10,11]:

$$\begin{aligned} ECE(\mathcal{L}'_t, \widetilde{P}_{tar}) =& \frac{\widetilde{P}_{tar}}{N_{tar}} \sum_{t \in tar} \log_2 \left(1 + e^{(-\mathcal{L}'_t - \operatorname{logit} \widetilde{P}_{tar})}\right) \\ &+ \frac{(1-\widetilde{P}_{tar})}{N_{non}} \sum_{t \in non} \log_2 \left(1 + e^{(\mathcal{L}'_t + \operatorname{logit} \widetilde{P}_{tar})}\right). \end{aligned} \tag{2.30}$$

Čím vyšší je hodnota ECE (Empirical Cross-Entropy), tím více informace je potřeba ke zjištění správné klasifikace soudů. Rozdíl $H(\widetilde{P}_{tar}) - ECE(\mathcal{L}'_t, \widetilde{P}_{tar})$ je možné interpretovat jako průměrnou (přes všechny předložené soudy) informaci získanou aplikací detektoru.

Je možné ukázat (Brummer – du Preez, 2006; Ramos – Gonzalez-Rodriguez, 2008), že ECE entropie je tvořena dvěma složkami:

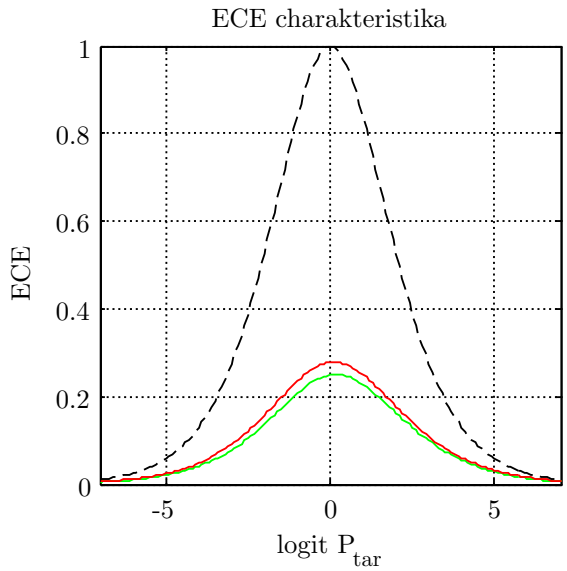
- průměrnou entropií posteriorních pravděpodobností stanovených ideálně kalibrovaným systémem. K provedení této kalibrace je třeba znalost referenční klasifikace soudů. K nalezení funkce, která převádí hodnoty skóre stanovené skutečným detektorem na optimální hodnoty LLR $\mathcal{L}''_t$, lze použít PAV algoritmus. Rozdíl $H(\widetilde{P}_{tar}) - ECE(\mathcal{L}''_t, \widetilde{P}_{tar})$ tak představuje informativní potenciál skóre. Platí (Brummer, 2010a), že $0 \leq H(\widetilde{P}_{tar}) - ECE(\mathcal{L}''_t, \widetilde{P}_{tar})$.

- rozdílem rozložení hodnot $\mathcal{L}'_t$ poskytovaných systémem od rozložení ideálních hodnot $\mathcal{L}''_t$ nalezených PAV algoritmem. Přírůstek entropie způsobený tímto rozdílem lze vyjádřit na základě průměrné (přes všechny předložené soudy) Kullback-Leiblerovy divergence mezi skutečně stanovenou a ideální posteriorní pravděpodobností (Brummer – du Preez, 2006). Pokud není detektor ideálně kalibrován, dochází ke ztrátě informace a proto platí $ECE(\mathcal{L}'_t, \widetilde{P}_{tar}) \geq ECE(\mathcal{L}''_t, \widetilde{P}_{tar})$. Rozdíl $H(\widetilde{P}_{tar}) - ECE(\mathcal{L}'_t, \widetilde{P}_{tar})$ pak představuje průměrnou informaci, kterou lze skutečně získat aplikací detektoru. V tomto případě nemá uvedený rozdíl dolní mez a může být obecně záporný. V takovém případě je použití detektoru na škodu, protože referenční detektor poskytuje lepší výsledky, a měla by být provedena kalibrace systému.

Tento rozklad tedy přesně odpovídá konceptu rozkladu C_{llr} na složku odpovídající vyhodnocení rozlišovacích schopností a kalibrace detektoru. To však není jediná souvislost. Z porovnání vztahů (2.30) a (2.24) je zřejmé, že platí $ECE(\mathcal{L}'_t, \widetilde{P}_{tar} = 0{,}5) = C_{llr}(\mathcal{L}'_t)$. Z toho vyplývá další možná interpretace C_{llr}. Rozdíl $1 - C_{llr}$ představuje průměrnou míru informace, kterou lze získat

[10] Vyjádření v (Ramos – Gonzalez-Rodriguez, 2008) je provedeno ve tvaru pracujícím s poměry věrohodností, v souladu s definicí ostatních kritérii je zde však použit logaritmus poměru věrohodností.

[11] Ekvivalentního vyjádření, s jediným rozdílem spočívajícím v normalizaci $\log 2$, lze dosáhnout odvozením kritéria pro vyhodnocení kvality poměru věrohodností poskytovaných detektorem na základě logaritmického skórovacího pravidla (Brummer, 2010a). ECE také úzce souvisí s již dříve navrženou (Campbell et al., 2005) normalizovanou křížovou entropií NCE (Normalized Cross-Entropy), která vyjadřuje relativní snížení nejistoty o klasifikaci soudů po použití detektoru.



Obrázek 2.7: *Příklad ECE diagramu vyhodnoceného pro stejný systém, který byl použit v případě obr. 2.5*

použitím detektoru ve srovnání s referenčním detektorem, při maximální úrovni apriorní nejistoty o klasifikaci soudů (Brummer – du Preez, 2006), tj. když $\widetilde{P}_{tar} = 0{,}5$.

2.2.12 ECE charakteristika

ECE charakteristika, navržená autory (Ramos – Gonzalez-Rodriguez, 2008), vyjadřuje průběh hodnoty ECE v závislosti na hodnotě $\widetilde{P}_{tar}$. Příklad ECE charakteristiky je uveden na obr. 2.7.

Z porovnání obr. 2.7 a 2.5 je patrná podoba ECE charakteristiky s APE charakteristikou. V ECE diagramu je navíc stejně jako v případě APE diagramu znázorněn průběh sledovaného kritéria nejenom pro skutečný systém (červenou čarou), ale také pro ideálně kalibrovaný (zelenou čarou) a referenční systém (čerchovanou čarou). Oba diagramy tak poskytují podobnou představu o kalibraci systému v širokém rozsahu pracovních bodů.

ECE charakteristika je však vhodnějším nástrojem pro forenzní interpretaci vyhodnocení spolehlivosti systémů. S ohledem na justiční oblast je vhodnější hovořit o průměrné informaci, kterou je systém schopen určit na základě analýzy řečových dat (důkazu), než o chybové míře, která je sledována v případě APE charakteristiky. Navíc v případě APE charakteristiky dochází k vyhodnocení binárních rozhodnutí o klasifikaci soudů při dané hodnotě apriorní pravděpodobnosti, zatímco v případě ECE charakteristiky dochází k vyhodnocení kvality hodnot LLR stanovených systémem.

Ramos – Gonzalez-Rodriguez (2008) také přímo navrhuje formulaci, kterou může forenzní specialista použít, aby soudu srozumitelně vysvětlil vyhodnocení systému pomocí ECE diagramu.

2.3 Shrnutí

Metody pro vyhodnocování systémů by měly být schopné v první řadě odpovědět na otázku, který ze dvou systémů je lepší. To ovšem není zcela triviální, protože srovnání systémů může být odlišné v závislosti na zamýšlené aplikaci systému. V případě forenzních aplikací navíc není dostatečné vyhodnocení, které umožňuje provést relativní posouzení úspěšnosti dvou systémů. Aby mohl být výsledek vyhodnocení shody mluvčích mezi nahrávkami podezřelého a pachatele použit jako důkazní materiál v soudních případech, je nutné mít k dispozici míru umožňující vyhodnotit váhu a přesnost takového důkazu.

V literatuře zabývající se problematikou rozpoznávání mluvčích jsou dosud stále velmi často pro porovnání systémů využívány hodnota EER a DET charakteristika. DET charakteristika znázorňuje vyhodnocení míry chybného přijetí P_{FA} a zamítnutí P_{miss} v širokém rozsahu pracovních bodů. Hodnota EER pak představuje horní mez souhrnné chybové míry P_{err}^{min} přes všechny pracovní body. V obou případech je tak prováděno aplikačně nezávislé vyhodnocení. Jsou ale předpokládána verifikační rozhodnutí provedená při optimální volbě rozhodovacího prahu a tím pádem reflektovány pouze rozlišovací schopnosti detektoru. Není hodnocena kvalita jeho kalibrace. V publikacích prezentujících výsledky systémů s využitím NIST evaluačních dat je zpravidla uváděna také hodnota kritéria C_{det}^{min} (příp. její normalizovaná varianta) pro aplikační parametry specifikované v zadání příslušné evaluace, méně často je pak uváděna hodnota C_{det}.

V této kapitole bylo představeno několik kritérií a charakteristik, které poskytují mnohem větší množství informací o chování systému, včetně aplikačně nezávislého vyhodnocení kvality kalibrace. Preference kritérií posuzujících pouze rozlišovací schopnosti systémů však není způsobena nezájmem autorů o problematiku kalibrace, přestože byla donedávna poněkud opomíjeným tématem. Je spíše důsledkem možnosti provést nezávislý návrh (a vyhodnocení) systému nejprve s ohledem na co nejlepší rozlišovací schopnosti a následně s ohledem na kvalitu kalibrace. Kalibrovaná skóre ve formátu poměru věrohodností (resp. preferovaného logaritmu tohoto poměru) přináší velkou výhodu v možnosti snadného stanovení hodnoty rozhodovacího prahu v závislosti na parametrech zamýšlené aplikace. Kalibrace systémů a vhodný způsob vyhodnocení jsou nezbytné pro forenzní aplikaci systémů pro automatické rozpoznávání mluvčích.

Tabulka 2.1 shrnuje skalární kritéria používaná pro vyhodnocení systémů pro detekci mluvčích a uvádí také rozsah hodnot, kterých mohou tato kritéria nabývat. V případě kritérií C_{det} a P_{err} jsou uvažovány nenormalizované varianty.

S výjimkou hodnot EER (horní mez souhrnné chybové míry) a C_{llr} (integrál souhrnné chybové míry) jsou všechna kritéria uvedená v tabulce 2.1 závislá na volbě pracovního bodu. Zatímco skalární forma reprezentace aplikačně nezávislého vyhodnocení je důležitá pro diskriminativní trénování systémů, neposkytuje informace o chování systému v jednotlivých pracovních bodech. Míra C_{llr} například neumožňuje získat dobrou představu o chování systému s ohledem na pracovní bod definovaný v evaluaci pořádané NIST v roce 2010, protože hodnota P_{err} v tomto bodě představuje

Tabulka 2.1: *Porovnání skalárních kritérií používaných pro vyhodnocení verifikačních systémů*

kritérium	rozsah	hodnotí			zahrnuje vyhodnocení kalibrace
		bin. rozh.	skóre	LLR	
EER	$0 \ldots 0{,}5$		•		ne
C_{det}	$C_{det}^{min} \ldots \max(C_{miss}P_{tar}, C_{FA}(1-P_{tar}))$	•			ano
C_{det}^{min}	$0 \ldots \min(C_{miss}P_{tar}, C_{FA}(1-P_{tar}))$		•		ne
P_{err}	$P_{err}^{min} \ldots \max(\widetilde{P}_{tar}, 1-\widetilde{P}_{tar}) \leq 1$			•	ano
P_{err}^{min}	$0 \ldots \min(\widetilde{P}_{tar}, 1-\widetilde{P}_{tar}) \leq EER$		•		ne
C_{llr}	$C_{llr}^{min} \ldots \infty$			•	ano
C_{llr}^{min}	$0 \ldots 1$		•		ne
ECE	$0 \ldots \infty$			•	ano

Tabulka 2.2: *Porovnání grafických charakteristik používaných pro vyhodnocení verifikačních systémů přes rozsah pracovních bodů (aplikačně nezávislé vyhodnocení)*

charakteristika	sledovaná kritéria	zahrnuje vyhodnocení kalibrace
ROC / DET	$P_{miss}(\gamma), P_{FA}(\gamma)$	ne
APE	$P_{err}(\widetilde{P}_{tar}), P_{err}^{min}(\widetilde{P}_{tar}), P_{err0}(\widetilde{P}_{tar})$	ano
NBE	$\frac{P_{err}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})}, \frac{P_{err}^{min}(\widetilde{P}_{tar})}{P_{err0}(\widetilde{P}_{tar})}$	ano
ECE	$ECE(\mathcal{L}_t', \widetilde{P}_{tar}), ECE(\mathcal{L}_t'', \widetilde{P}_{tar}), H(\widetilde{P}_{tar})$	ano

pouze malý přírůstek k celkové hodnotě C_{llr}. Za účelem získání přehledu o úspěšnosti systémů v závislosti na volbě pracovního bodu bylo prezentováno několik grafických charakteristik, jejichž přehled je uveden v tabulce 2.2.

Kapitola 3

Kalibrace a fúze systémů pro detekci mluvčích

Termínem kalibrace rozumíme jak vlastnost hodnot LLR určených systémem vystupovat jako správné hodnoty LLR, kterou je možné vyhodnotit některým z kritérií uvedených v předchozí kapitole, tak i proces hledání transformační funkce provádějící převod z hodnot skóre na hodnoty LLR.

Parametry transformační funkce jsou trénovány na základě *vývojových* dat, pro které je známa referenční klasifikace. Statistické rozložení vývojových dat může být samozřejmě odlišné od rozložení dat použitých pro vyhodnocení. V takovém případě bude vyhodnocena špatná hodnota kalibračního kritéria. Neznamená to ovšem selhání kalibračního procesu pro stanovení parametrů transformační funkce. Kalibrace na rozdíl od různých forem normalizace skóre (např. S-norm nebo T-norm) nebere v úvahu aktuální data. Jediným vstupem kalibrační transformace je skóre určené detektorem a rozdíl v rozložení dat tak vede ke špatné kalibraci.

Jednu z možných transformačních funkcí představuje neparametrická funkce nalezená pomocí PAV algoritmu. V praxi je ovšem PAV algoritmus využíván především jako součást vyhodnocení systémů s cílem stanovení kvality kalibrace na základě příslušných kritérií.

3.1 Logistická regrese

Pravděpodobně nejčastěji používaný způsob kalibrace detektorů mluvčích je založen na lineární logistické regresi (Pigeon et al., 2000; Brummer et al., 2007). Snahou je nalézt parametry afinní transformace[1], které by zajistily optimální hodnotu účelové funkce. Transformační funkce má podobu

$$\mathcal{L}'_t = w_{lr}(s_t, \boldsymbol{\lambda}) = \alpha_0 + \alpha_1 s_t, \tag{3.1}$$

[1] Afinní transformace je složena z lineárního zobrazení a posunutí.

kde s_t je skóre určené detektorem pro verifikační soud t, $\boldsymbol{\lambda} = \{\alpha_0, \alpha_1\}$ představuje parametry funkce w_{lr} a $\mathcal{L}'_t$ je kalibrované skóre ve formátu LLR. Protože funkce w_{lr} je ryze monotónní, nedochází k ovlivnění rozlišovacích schopností detektoru. Účelová funkce je definována jako (Brummer et al., 2007)[2]:

$$\begin{aligned} Q(\boldsymbol{\lambda}, \widetilde{P}_{tar}) =& \frac{\widetilde{P}_{tar}}{N_{tar}} \sum_{t \in tar} \log_2 \left(1 + e^{(-w_{lr}(s_t, \boldsymbol{\lambda}) - \operatorname{logit} \widetilde{P}_{tar})}\right) \\ &+ \frac{(1 - \widetilde{P}_{tar})}{N_{non}} \sum_{t \in non} \log_2 \left(1 + e^{(w_{lr}(s_t, \boldsymbol{\lambda}) + \operatorname{logit} \widetilde{P}_{tar})}\right). \end{aligned} \tag{3.2}$$

Z porovnání s (2.30) je zřejmé, že účelová funkce logistické regrese přesně odpovídá ECE entropii. Hodnota $\widetilde{P}_{tar}$ ve vztahu (3.2) představuje parametr účelové funkce, který umožňuje provedení optimalizace s ohledem na zamýšlenou aplikaci nezávisle na poměru oprávněných a neoprávněných soudů ve vývojových datech. Pro fixní hodnotu $\widetilde{P}_{tar}$ je funkce $Q(\boldsymbol{\lambda}, \widetilde{P}_{tar})$ v závislosti na $\boldsymbol{\lambda}$ konvexní a má proto globální minimum. Pokud není zamýšlená aplikace detektoru předem známá, je možné zvolit libovolnou hodnotu $\widetilde{P}_{tar}$. Typickou volbou je v tomto případě hodnota $\widetilde{P}_{tar} = 0{,}5$, účelová funkce pak odpovídá kritériu C_{llr}. Výsledek logistické regrese není výrazně závislý na zvolené hodnotě $\widetilde{P}_{tar}$ a kalibrace provedená pro danou hodnotu zpravidla poskytuje dobrou kalibraci pro široký rozsah apriorních pravděpodobností (Brummer, 2010a).

Optimalizaci logistické regrese nelze vyřešit analyticky a je proto založena na iterativním numerickém řešení. Porovnáním řady optimalizačních metod se zabývá (Minka, 2003). Na základě této práce vytvořil Niko Brümmer nástroj nazvaný Focal toolkit[3] (Brummer et al., 2007), který používá pro optimalizaci algoritmus založený na metodě konjugovaných gradientů.

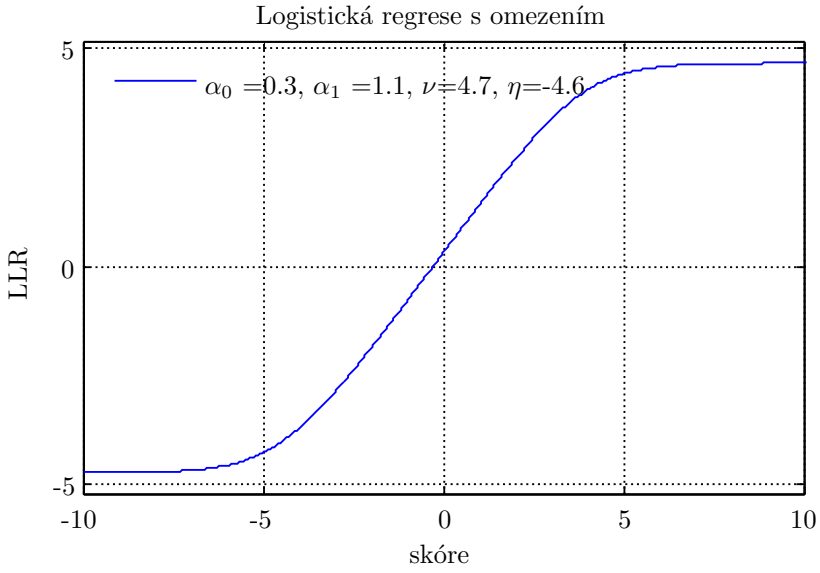
Budeme-li uvažovat výstup detektoru již ve formátu LLR, budou znamenat kladné hodnoty podporu hypotézy o klasifikaci soudu jako oprávněného. Naopak záporné hodnoty podporu opačné hypotézy. Pokud detektor mylně vyjádří vysokou míru podpory chybné hypotéze (určí vysokou hodnotu LLR s chybným znaménkem), bude důsledkem vysoká pokuta podílející se na hodnotě účelové funkce. V případě podpory správné hypotézy bude pokuta přirozeně malá. Vysoká míra podpory chybné hypotézy však nutně nemusí znamenat chybu detektoru, ale může být způsobena chybnou referenční klasifikací. Za účelem potlačení vlivu těchto chyb a jim příslušných vysokých pokut byla navržena (Brummer et al., 2007) nelineární varianta logistické regrese. Příslušná transformační funkce je definována jako:

$$\mathcal{L}'_t = w_{scal}(s_t, \boldsymbol{\lambda}) = \log \frac{\left(\operatorname{logit}^{-1} \nu\right)\left(\exp(\alpha_0 + \alpha_1 s_t) - 1\right) + 1}{\left(\operatorname{logit}^{-1} \eta\right)\left(\exp(\alpha_0 + \alpha_1 s_t) - 1\right) + 1}, \tag{3.3}$$

kde η a ν jsou parametry funkce. Pokud je splněna podmínka $\eta < 0 < \nu$, má funkce tvar sigmoidy a omezuje rozsah hodnot $\mathcal{L}'_t$ na interval $(-\nu; -\eta)$. Příklad této funkce je uveden na obr. 3.1.

[2] Zde uvedené kritérium se liší normalizací $\log 2$, která však z pohledu optimalizace nemá na výsledek žádný vliv.

[3] Jedná se o sadu nástrojů pro kalibraci a fúzi systémů v jazyce Matlab. Focal toolkit si získal v komunitě zabývající se rozpoznáváním mluvčích (a jazyků) velkou popularitu a byl použit pro kalibraci a fúzi systémů i v této práci.



Obrázek 3.1: *Ukázka průběhu transformační funkce nalezené nelinearní variantou logistické regrese*

Všechny parametry transformační funkce $\boldsymbol{\lambda} = \{\alpha_0, \alpha_1\, \nu, \eta\}$ jsou nalezeny optimalizačním algoritmem. Transformační funkce w_{scal} je za předpokladu splnění uvedené podmínky monotónně rostoucí a tím pádem invertovatelná.

Brummer et al. (2010) nabízí interpretaci parametrů ν a η ve vztahu k možným chybám v referenční anotaci vývojových dat. Parametr ν je interpretován jako logaritmus šance, že pravý soud bude v datech chybně označen jako nepravý, a η jako logaritmus šance, že nepravý soud bude označen jako pravý.

3.2 Příklady dalších kalibračních transformací

V literatuře lze nalézt řadu dalších metod pro kalibraci hodnot skóre na hodnoty LLR. Často užívaným způsobem ve forenzních aplikacích je například reprezentace rozložení skóre příslušných oprávněným a neoprávněným soudům pomocí pravděpodobnostních generativních modelů (Ramos, 2007). Pravděpodobně nejjednodušší přístup je založen na použití unimodálních Gaussovských rozložení (Botti et al., 2004). Lze však uvažovat i vícemodální modely (Ramos-Castro et al., 2005) nebo neparametrické modely založené na jádrových hustotních funkcích (kernel density function) (Meuwly – Drygajlo, 2001; Alexander, 2005; Gonzalez-Rodriguez et al., 2005). Výsledkem uvedených přístupů jsou obecně neinvertovatelné transformace.

(Campbell et al., 2005) se zabývá přístupem založeným na neuronových sítích. Vícevrstvý perceptron je natrénován pro transformaci hodnot skóre na hodnoty LLR s ohledem na optimalizaci účelové funkce představované NCE entropií (jedná se o již zmíněnou normalizovanou verzi ECE). Podle (Ramos, 2007) je tak tento přístup možné chápat jako nelineární variantu logistické regrese, resp. logistickou regresi lze chápat jako variantu tohoto přístupu, kdy je neuronová síť představována

jednovrstvým perceptronem. I v tomto případě je výsledkem obecně neinvertovatelná transformace a je tak vhodné provést kontrolu vlivu transformace na rozlišovací schopnosti detektoru.

3.3 Fúze systémů

Cílem fúze (kombinace) systémů je vytěžit na základě komplementárnosti různých systémů větší množství informace související s úlohou detekce mluvčích, než jsou schopné samostatné systémy. V současné době je pro fúzi systémů pravděpodobně nejčastěji aplikována lineární logistická regrese (Brummer et al., 2007), jak dokládá například četné využití v posledním ročníku NIST evaluace (Brummer et al., 2010; Scheffer et al., 2011; Castaldo et al., 2011; Sturim et al., 2011).

Výsledné skóre stanovené na základě kombinace skóre určených K dílčími systémy pomocí lineární logistické regrese představuje vážený součet

$$\begin{aligned} \mathcal{L}'_t = w_f(\boldsymbol{s}_t, \boldsymbol{\lambda}) &= \alpha_0 + \sum_{k=1}^{K} \alpha_k s_{t,k} \\ &= \boldsymbol{\alpha}^T \boldsymbol{s}_t, \end{aligned} \tag{3.4}$$

kde $s_{t,k}$ je skóre určené systémem k pro soud t, vektor $\boldsymbol{s}_t = [1, s_{t,1}, \ldots, s_{t,K}]^T$ a vektor $\boldsymbol{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_K]^T$. Parametry $\boldsymbol{\alpha}$ jsou stanoveny na základě optimalizace účelové funkce. Tradičně je používána stejná účelová funkce jako v případě hledání parametrů kalibrační transformace monolitických systémů ($K = 1$), tedy funkce (3.2). Je zřejmé, že funkce w_f není nutně invertovatelná a může tak docházet k ovlivnění rozlišovacích schopností detektoru. Zlepšení rozlišovacích schopností detektoru znamená, že je systém schopen získat více informace na základě analýzy předložených řečových dat. Jak již bylo uvedeno, účelová funkce (3.2) odpovídá kritériu ECE, které hodnotí nejenom rozlišovací schopnosti ale také kalibraci detektoru. Aby bylo dosaženo minima účelové funkce, měly by být výstupem funkce w_f hodnoty odpovídající logaritmu poměru věrohodností. Výsledkem fúze jsou tak hodnoty ve formátu LLR, jak je explicitně zdůrazněno ve vztahu (3.4).

Dle (Brummer et al., 2010) lze dosáhnout lepší kvality kalibrace systémů aplikací nelineární transformace omezující rozsah hodnot LLR podle (3.3). I v tomto případě je však nejprve provedena lineární kombinace dílčích systémů jako vážený součet příslušných skóre a až následně je aplikována saturující sigmoidní transformace.

Zajímavé je, že na hodnoty vah příslušných dílčím systémům α_k ($k = 1, \ldots, K$) nejsou v případě kritéria (3.2) kladena žádná omezení, a může se tak stát, že některé váhy budou záporné. Jak však uvádí (Brummer et al., 2007), je možné nalézt řešení zahrnující dílčí systémy se zápornou vahou přinášející dobré výsledky.

3.4 Podmíněná kalibrace a fúze systémů

V praxi často není aplikována pro všechny soudy shodná kalibrace nebo fúze systémů. V NIST evaluacích jsou například používána data zaznamenaná na mikrofon i prostřednictvím telefonního

kanálu. Rozložení skóre určených detektorem pro oprávněné a neoprávněné soudy jsou odlišná v závislosti na typu kanálu dat použitých pro trénování modelu ověřovaného mluvčího a testované nahrávky. Rozložení jsou však rozdílná i v závislosti na pohlaví mluvčích. Z tohoto důvodu je vhodné aplikovat specifické kalibrace/fúze v závislosti na podmínkách předloženého soudu (Brummer et al., 2010). Ve forenzních aplikacích je pak nutné věnovat každému případu zvláštní pozornost, a proto je vhodné stanovit kalibraci specifickou pro aktuálně posuzovaný případ (Ramos, 2007).

Jak již bylo uvedeno, fúze systémů představuje obecně neinvertovatelnou transformaci a podmíněná aplikace na tom nic nemění. Jak je to ovšem s kalibrační transformací? Pokud detektor používá různé (podmíněné) kalibrační transformace, pak kalibrace detektoru představuje z pohledu zobrazení všech skóre na hodnoty LLR neinvertovatelnou funkci, a to i v případě, že jednotlivé kalibrační transformace jsou invertovatelné.

Podmíněná fúze je jednoznačně užitečná v případech, kdy systém pracuje s různými typy dat. K podmíněné fúzi je však třeba přistupovat opatrně, protože specifikováním velkého počtu úzce vymezených kategorií dochází k velké fragmentaci vývojových dat. Omezené množství dat dostupných pro trénování parametrů jednotlivých transformačních funkcí pak může vést k problému *přetrénování*.

Podmíněná fúze vyžaduje natrénování parametrů $\boldsymbol{\alpha}_\psi = [\alpha_{0,\psi}, \alpha_{1,\psi}, \ldots, \alpha_{K,\psi}]^T$ specifických pro každou z definovaných kategorií $\psi = 1, \ldots, \Psi$. Za účelem zabránění problému přetrénování navrhuje Ferrer et al. (2008) vyjádření parametrů $\boldsymbol{\alpha}_\psi$ na základě rozdílu od globálního vektoru vah $\boldsymbol{\alpha}$ jako

$$\boldsymbol{\alpha}_\psi = \boldsymbol{\alpha} + \Delta\boldsymbol{\alpha}_\psi. \tag{3.5}$$

Účelová funkce pro trénování parametrů fúze je přitom rozšířena o regularizační člen penalizující rostoucí $\Delta\boldsymbol{\alpha}_\psi$ a současně je s cílem zamezení možného přetrénování požadováno, aby všechny váhy byly kladné.

Dalším možným řešením je postupná aplikace podmíněné kalibrace v několika krocích (Burget et al., 2008), kdy je v každém kroku specifikován pouze malý počet kategorií. V prvním kroku může být například provedena kalibrace specifická v závislosti na pohlaví a v dalším kroku v závislosti na typu přenosového kanálu.

3.5 Využití přídavných informací

Aplikací podmíněné fúze dochází k efektivnímu využití přídavných informací, které byly systému poskytnuty společně s předloženým soudem, případně automaticky zjištěny systémem. Bez ohledu na způsob získání těchto přídavných informací je pro aplikaci podmíněné fúze nutné, aby byly tyto informace kategorizovatelné. Existuje však řada údajů, o kterých lze předpokládat, že mají potenciál informačního přínosu pro rozhodnutí detektoru, a měly by být tedy brány v potaz, ale mají spojitou povahu (ať již v podobě celé nebo reálné hodnoty). Jako příklad lze uvést délku nahrávky, počet rozpoznaných hlásek (naznačuje množství užitečné informace v nahrávce), odhad spolehlivosti automatického přepisu nebo hodnotu SNR (odstup signálu od šumu).

Jedno z možných řešení pro využití těchto údajů spočívá ve vytvoření kategorií na základě kvantizace spojitých hodnot podle předem definovaných intervalů (případně lze použít metody pro shlukování) a následné aplikaci podmíněné fúze (Ferrer et al., 2008). S kategorizací spojitých hodnot ovšem opět vyvstává problém s fragmentací trénovacích dat a možným přetrénováním.

V (Campbell et al., 2005) jsou hodnoty přídavných informací společně s hodnotami skóre určenými dílčími systémy rovnocennými vstupy vícevrstvé neuronové sítě.

Způsob navržený v (Brummer et al., 2010) zahrnuje hodnoty přídavných informací přímo do výpočtu vážené kombinace skóre:

$$w_f(\boldsymbol{s}_t, \boldsymbol{\lambda}) = \boldsymbol{\alpha}^T \boldsymbol{s}_t + \boldsymbol{\beta}^T \boldsymbol{r}_t + \boldsymbol{q}_{t\_model}^T \boldsymbol{\Gamma} \boldsymbol{q}_{t\_test}, \tag{3.6}$$

kde dosud nespecifikované proměnné mají následující význam: $\boldsymbol{r}_t$ představuje vektor hodnot vyjadřujících přídavné informace příslušné soudu[4] t, $\boldsymbol{q}_{t\_test}$ vyjadřuje informace související s předloženou testovanou nahrávkou a konečně $\boldsymbol{q}_{t\_model}$ reprezentuje vektor hodnot vyjadřujících informace související s nahrávkami použitými pro natrénování modelu mluvčího. V (Brummer et al., 2010) je v souvislosti se základním zadáním NIST evaluací uvažováno použití pouze jedné trénovací nahrávky, což činí výklad snáze interpretovatelný. Vektor $\boldsymbol{\beta}$ a matice[5] $\boldsymbol{\Gamma}$ představují stejně jako vektor $\boldsymbol{\alpha}$ parametry fúze, které lze souhrnně reprezentovat pomocí $\boldsymbol{\lambda} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Gamma}\}$. Všechny parametry jsou stanoveny optimalizací účelové funkce na trénovacích datech.

Popsaný způsob využití přídavných informací nepředstavuje náhradu přístupu založeného na podmíněné fúzi. Pokud existuje nějaké kritérium, které umožňuje přirozenou definici malého počtu kategorií (např. na základě typu přenosového kanálu) tak, aby nedocházelo k velké fragmentaci dat, je vhodné použít přístup založený na podmíněné fúzi. Oba přístupy jsou tak komplementární a je možné je kombinovat, jak dokládá například (Brummer et al., 2010).

[4] Zdůrazněme, že se jedná o informace specifické pro konkrétní soud, tvořený kombinací modelu a testované nahrávky. Informace specifické pro model bez ohledu na testovanou nahrávku a informace specifické pro testovanou nahrávku bez ohledu na předložený model jsou reprezentovány vektory $\boldsymbol{q}_{t\_model}$ a $\boldsymbol{q}_{t\_test}$.

[5] Matice $\boldsymbol{\Gamma}$ je symetrická.

Kapitola 4

Generativní klasifikátory

Cílem klasifikátorů je určit pro vstupní data $\boldsymbol{o} \in \mathcal{O}$ označení třídy $y \in \mathcal{Y}$, ke které daná data přísluší. Přístupy, které explicitně nebo implicitně modelují rozložení vstupních dat stejně jako výstupních, jsou označovány jako generativní, protože je možné vygenerovat na jejich základě syntetická data ve vstupním prostoru (Bishop, 2006, s. 43). V případě *generativních klasifikátorů* je sdružené rozložení $P(\boldsymbol{o}, y)$ modelováno na základě funkcí hustot podmíněné pravděpodobnosti $P(\boldsymbol{o}|y)$, naučených zvlášť pro každou třídu $y \in \mathcal{Y}$, a apriorních pravděpodobností tříd $P(y)$. Posteriorní pravděpodobnost pro třídu y_k je jednoduše určena aplikací Bayesovy věty jako

$$P(y_k|\boldsymbol{o}) = \frac{P(\boldsymbol{o}|y_k)P(y_k)}{P(\boldsymbol{o})}. \tag{4.1}$$

Na základě znalosti posteriorních pravděpodobností je pak provedeno rozhodnutí o příslušnosti vstupních dat k jedné ze tříd.

Naproti tomu *diskriminativní klasifikátory*, o kterých pojednává následující kapitola, modelují přímo posteriorní pravděpodobnosti $P(y|\boldsymbol{o})$, nebo se dokonce učí přímo zobrazení z vstupních dat $\boldsymbol{o}$ na označení tříd.

4.1 Metody odhadu parametrů modelů

Podmíněná pravděpodobnost $P(\boldsymbol{o}|y)$ je zpravidla reprezentována formou nějakého parametrického rozložení, o kterém se předpokládá, že je známé. Parametry rozložení jsou obvykle uspořádány do vektoru parametrů $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_N)$. Skutečné hodnoty těchto parametrů $\boldsymbol{\theta}_0$ pro daný objekt nejsou přímo pozorovatelné a tak musí být odhadovány (estimovány) na základě pozorovaných dat $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T\}$. Cílem je nalézt takový odhad $\tilde{\boldsymbol{\theta}}$, který by byl co nejblíže skutečným parametrům. Nejčastěji používanými metodami odhadu parametrů jsou metoda *maximální věrohodnosti* a metoda *maximální aposteriorní pravděpodobnosti*.

4.1.1 Metoda maximální věrohodnosti

Odhad parametrů metodou maximální věrohodnosti (ML) hledá takové hodnoty parametrů rozložení $P(\boldsymbol{O}|\boldsymbol{\theta})$, které zajišťují nejvyšší pravděpodobnost vygenerování pozorovaných (trénovacích) dat, tedy

$$\tilde{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\boldsymbol{\theta}) \tag{4.2}$$

Za předpokladu, že pozorovaná data $\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T$ jsou považována za nezávislé a identicky rozložené náhodné veličiny, lze vyjádřit funkci hustoty sdružené pravděpodobnosti pro data jako

$$P(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T|\boldsymbol{\theta}) = P(\boldsymbol{o}_1|\boldsymbol{\theta})P(\boldsymbol{o}_2|\boldsymbol{\theta}) \ldots P(\boldsymbol{o}_T|\boldsymbol{\theta}). \tag{4.3}$$

Pokud budeme uvažovat (4.3) jako funkci parametrů rozložení $\boldsymbol{\theta}$ a nikoliv pozorovaných dat[1], označíme ji jako věrohodnostní funkci

$$L(\boldsymbol{\theta}|\boldsymbol{O}) = P(\boldsymbol{O}|\boldsymbol{\theta}) = \prod_{i=1}^{T} P(\boldsymbol{o}_i|\boldsymbol{\theta}). \tag{4.4}$$

V praxi je často výhodné pracovat s logaritmem věrohodnostní funkce, pro který platí

$$\ell(\boldsymbol{\theta}|\boldsymbol{O}) = \log L(\boldsymbol{\theta}|\boldsymbol{O}) = \sum_{i=1}^{T} \log P(\boldsymbol{o}_i|\boldsymbol{\theta}). \tag{4.5}$$

Kritérium (4.2) lze pak přepsat pomocí věrohodnosti jako

$$\tilde{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}|\boldsymbol{O}) \tag{4.6}$$

a proto hovoříme o *maximálně věrohodném* odhadu.

4.1.2 Metoda maximální aposteriorní pravděpodobnosti

Zatímco v případě ML odhadu se parametry předpokládají jako pevné, ale neznámé, v případě metody maximální aposteriorní pravděpodobnosti (MAP) jsou odhadované parametry považovány za náhodné veličiny, pro které je předpokládána znalost jejich apriorního rozložení $P(\boldsymbol{\theta})$ (Gauvain – Lee, 1994; Heijden, 2004). V případě ML odhadu není k dispozici žádná znalost o $\boldsymbol{\theta}$. Ekvivalentní situace nastává v případě neinformativního apriorního rozložení parametrů[2].

Odhad parametrů metodou MAP je definován jako modus funkce hustoty posteriorního rozložení $\boldsymbol{\theta}$ (Gauvain – Lee, 1994), tedy platí

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_{MAP} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\boldsymbol{\theta}|\boldsymbol{O}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \frac{P(\boldsymbol{O}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{O})} \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\boldsymbol{O}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \end{aligned} \tag{4.7}$$

[1] Pozorovaná data $\boldsymbol{O}$ lze v tomto případě chápat jako pevné „parametry“ funkce.

[2] Příkladem je rovnoměrné rozložení, kdy $P(\boldsymbol{\theta}) = \text{konst.}$

Metoda MAP představuje speciální případ bayesovského odhadu (bayesian inference), jehož výsledkem je bodový odhad parametrů. Pro plně bayesovský přístup je typická reprezentace pomocí rozložení, a pokud je cílem stanovení bodového odhadu, je upřednostňována spíše střední hodnota[3] nebo medián spíše než modus.

MAP odhad nachází uplatnění zejména v situacích, kdy nelze předpokládat, že dostupné množství dat je dostatečné pro získání spolehlivého odhadu parametrů metodou ML. Nedostatek dat je pak kompenzován informací o apriorním rozložení. V důsledku znalosti apriorního rozložení pak často bývá proces odhadu parametrů metodou MAP označován jako *adaptace* modelu namísto trénování modelu. MAP odhad parametrů je využíván například jako adaptační metoda v úloze rozpoznávání *řeči* pro přizpůsobení (adaptaci) univerzálního akustického modelu, typicky reprezentovaného skrytými Markovovými modely (HMM), na hlas mluvčího rozpoznávané promluvy (Červa, 2007). Široké uplatnění však MAP adaptace modelů nachází i v úloze rozpoznávání mluvčích.

4.2 Směsi Gaussových rozložení

V úloze automatického rozpoznávání mluvčích (i řeči) jsou akustické příznakové vektory obvykle modelovány Gaussovými (normálními) rozloženími. Parametry těchto rozložení jsou vektor středních hodnot a kovarianční matice. S ohledem na velkou variabilitu řečového signálu, a v souvislosti s ní i akustických příznakových vektorů, je nutné uvažovat rozložení jako vícemodální. Směs Gaussových rozložení (GMM) představuje model tvořený váženou kombinací C Gaussových rozložení[4]. Ta jsou charakterizována vahou w_c, vektorem středních hodnot $\boldsymbol{\mu}_c$ a kovarianční maticí $\boldsymbol{\Sigma}_c$, kde $c = 1, \ldots, C$. Váhu w_c lze interpretovat jako apriorní pravděpodobnost c-té komponenty modelu. Kovarianční matice $\boldsymbol{\Sigma}_c$ jsou obecně plné, ale nejčastěji pouze diagonální[5]. Věrohodnostní funkce GMM pro příznakový vektor $\boldsymbol{o}$ je vyjádřena vztahem

$$P(\boldsymbol{o}|\boldsymbol{\theta}) = \sum_{c=1}^{C} w_c \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \tag{4.8}$$

kde $\boldsymbol{\theta} = \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$ představuje vektor všech parametrů modelu. Váhy komponent musí splňovat podmínky $w_c \geq 0$ a $\sum_{c=1}^{C} w_c = 1$. Pro F-rozměrné příznakové vektory je funkce hustoty pravděpodobnosti c-té Gaussovy komponenty rovna

$$\mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \frac{1}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} \exp\left(-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{o} - \boldsymbol{\mu}_c)\right), \tag{4.9}$$

kde $|\cdot|$ symbolizuje determinant matice. V praxi je zpravidla z výpočetních a numerických důvodů využíván logaritmus věrohodnosti, pak lze (4.9) přepsat jako

$$\log \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = -\frac{1}{2} F \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_c| - \frac{1}{2} \boldsymbol{\mu}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu} - \frac{1}{2} \boldsymbol{o}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{o} + \boldsymbol{o}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_c \tag{4.10}$$

[3] $\mathbb{E}[\boldsymbol{\theta}] = \int_{\boldsymbol{\theta}} \boldsymbol{\theta} P(\boldsymbol{\theta}|\boldsymbol{O}) d\boldsymbol{\theta}$.

[4] V následujícím textu budou jednotlivá rozložení označována též jako komponenty modelu.

[5] Diagonální kovarianční matici lze reprezentovat vektorem rozptylů.

kde jsou na aktuálním příznakovém vektoru $\boldsymbol{o}$ závislé pouze poslední dva členy a ostatní mohou být předpočítány a uloženy společně s logaritmem váhy w_c v podobě konstanty specifické pro danou komponentu modelu. Platí tedy

$$\log w_c \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = \kappa_c - \frac{1}{2}\boldsymbol{o}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{o} + \boldsymbol{o}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c, \tag{4.11}$$

kde

$$\kappa_c = \log w_c - \frac{1}{2}F\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_c| - \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}. \tag{4.12}$$

Nyní předpokládejme sekvenci příznakových vektorů $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ reprezentující zparametrizovaný signál. Za zjednodušujícího předpokladu nezávislosti příznakových vektorů můžeme vyjádřit logaritmus věrohodnostní funkce pro tuto sekvenci jako

$$\begin{aligned} P(\boldsymbol{O}|\boldsymbol{\theta}) &= \sum_{t=1}^{T} \log P(\boldsymbol{o}_t|\boldsymbol{\theta}) \\ &= \sum_{t=1}^{T} \log \sum_{c=1}^{C} \exp\left(\kappa_c - \frac{1}{2}\boldsymbol{o}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{o} + \boldsymbol{o}^T\boldsymbol{\Sigma}_c^{-1}\boldsymbol{\mu}_c\right). \end{aligned} \tag{4.13}$$

4.2.1 EM algoritmus

Zatímco v případě unimodálního Gaussova rozložení je možné vyjádřit ML odhad parametrů v uzavřené formě, v případě vícemodálního modelu takové vyjádření není možné, protože není známá příslušnost trénovacích dat k jednotlivým komponentám modelu. Odhad parametrů GMM tak spadá do obecné třídy problémů chybějících dat (missing data problem). Jedním z možných řešení je použití gradientních optimalizačních metod (Bishop, 2006). Pravděpodobně nejčastěji využívaným způsobem získávání maximálně věrohodného odhadu parametrů GMM je však EM algoritmus (Dempster et al., 1977).

Cílem EM (Expectation-Maximization) algoritmu je provést odhad parametrů, který závisí na skrytých (latentních) proměnných $\mathcal{Z}$ (v našem případě představují skryté proměnné informace o příslušnosti trénovacích dat ke komponentám modelu).

EM algoritmus představuje iterativní postup, kdy jsou v rámci každé iterace prováděny dva kroky:

- *E (expectation) krok*: provádí vyhodnocení očekávané logaritmické věrohodnosti kompletních dat Q na základě podmíněného posteriorního rozložení skrytých proměnných $\mathcal{Z}$ vzhledem k pozorovaným datům $\boldsymbol{O}$ a stávajícímu odhadu parametrů modelu $\hat{\boldsymbol{\theta}}$:

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathcal{Z}|\boldsymbol{O},\hat{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}|\boldsymbol{O}, \mathcal{Z})]. \tag{4.14}$$

- *M (maximization) krok*: hledá parametry, které maximalizují očekávanou věrohodnost Q:

$$\boldsymbol{\theta}^{new} = \mathrm{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}). \tag{4.15}$$

Přestože mezi jednotlivými iteracemi nedochází nikdy k poklesu věrohodnostní funkce (Bishop, 2006), není zaručeno, že EM algoritmus najde globální maximum. Věrohodnostní funkce může mít obecně několik lokálních extrémů (maxim), ve kterých může EM algoritmus v závislosti na inicializaci uvíznout. Činnost algoritmu je ukončena v okamžiku, kdy nedochází k žádné nebo pouze malé změně věrohodnostní funkce[6] pod stanoveným prahem.

4.2.2 Odhad parametrů metodou maximální věrohodnosti

Předpokládejme, že jsou k dispozici trénovací data $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$. Dále budeme uvažovat C-rozměrnou binární náhodnou veličinu $\boldsymbol{z}$, pro jejíž složky platí $z_i \in \{0, 1\}$ a $\sum_{c=1}^{C} z_c = 1$ (jinými slovy, pouze jedna složka je vždy rovna 1 a ostatní jsou rovny 0). Proměnná $\boldsymbol{z}_t$ vyjadřuje příslušnost příznakového vektoru $\boldsymbol{o}_t$ k jedné z komponent modelu. Logaritmus věrohodnosti modelu pro trénovací data je pak roven

$$\ell(\boldsymbol{\theta}|\boldsymbol{O}) = \log P(\boldsymbol{O}|\boldsymbol{\theta}) = \log \sum_{\mathcal{Z}} P(\boldsymbol{O}, \mathcal{Z}|\boldsymbol{\theta}). \tag{4.16}$$

Přímá optimalizace parametrů modelu vzhledem k funkci $\log P(\boldsymbol{O}|\boldsymbol{\theta})$ je však zpravidla velmi obtížná. Úloha by přitom byla značně jednodušší, pokud by byla známa skutečná příslušnost všech trénovacích dat ke komponentám GMM modelu, tj. $\mathcal{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T\}$. Pár $\{\boldsymbol{O}, \mathcal{Z}\}$ pak představuje *kompletní* data a logaritmus věrohodnostní funkce modelu nabývá podoby $\log P(\boldsymbol{O}, \mathcal{Z}|\boldsymbol{\theta})$. Za předpokladu nezávislosti trénovacích dat lze věrohodnost pro kompletní data vyjádřit jako:

$$\log P(\boldsymbol{O}, \mathcal{Z}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \sum_{c=1}^{C} z_{t,c} \left(\log w_c + \log \mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \right), \tag{4.17}$$

kde $z_{t,c}$ představuje c-tou složku $\boldsymbol{z}_t$. Protože všechny složky $\boldsymbol{z}_t$ kromě jedné jednotkové jsou nulové, je možné interpretovat logaritmus věrohodnostní funkce pro kompletní data jednoduše jako součet přes C nezávislých rozložení. Odhad parametrů modelu s ohledem na maximalizaci věrohodnostní funkce je tak možné provést nezávisle pro jednotlivé komponenty (unimodální rozložení) na základě dat příslušných k dané komponentě. V případě znalosti kompletních dat tedy existuje řešení v uzavřené formě.

Ve skutečnosti jsou však známa pouze *nekompletní* data v podobě $\boldsymbol{O}$. Jediná znalost o hodnotách skrytých proměnných je představována posteriorním rozložením $P(\mathcal{Z}|\boldsymbol{O}, \boldsymbol{\theta})$. Namísto neznámých hodnot je tak možné pracovat s očekávanými hodnotami skrytých proměnných stanovenými na základě tohoto posteriorního rozložení. Lze ukázat (Bishop, 2006), že očekávaná hodnota c-té složky $\boldsymbol{z}_t$ je rovna[7]

$$\mathbb{E}[z_{t,c}] = \gamma_{t,c} = \frac{w_c \mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{k=1}^{C} w_k \mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{4.18}$$

[6] Zpravidla bývá sledována relativní změna průměrné hodnoty věrohodnostní funkce, tj. $\bar{\ell} = \frac{1}{T}\ell$.

[7] V praxi se obvykle provádí vyhodnocení logaritmu věrohodnosti podle (4.11), označíme-li $q_c = \log w_c \mathcal{N}(\boldsymbol{o}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, vztah (4.18) je možné vyjádřit v podobě $\gamma_c = \frac{\exp q_c}{\sum_{k=1}^{C} \exp q_k}$ a tato funkce bývá označována softmax.

Hodnota $\gamma_{t,c}$ vyjadřuje míru příslušnosti příznakového vektoru $\boldsymbol{o}_t$ k c-té komponentě (stanovenou na základě stávajících parametrů modelu $\hat{\boldsymbol{\theta}}$) a bývá označována jako okupační pravděpodobnost komponenty. Očekávanou hodnotu logaritmu věrohodnostní funkce pro kompletní data lze na základě (4.18) vyjádřit jako[8]

$$\begin{aligned}\mathbb{E}_{\mathcal{Z}|\boldsymbol{O},\hat{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}|\boldsymbol{O},\mathcal{Z})] &= \sum_{\mathcal{Z}} P(\mathcal{Z}|\boldsymbol{O},\hat{\boldsymbol{\theta}}) \log P(\boldsymbol{O},\mathcal{Z}|\boldsymbol{\theta}) \\ &= \sum_{t=1}^{T}\sum_{c=1}^{C} \gamma_{t,c} \left(\log w_c + \log \mathcal{N}(\boldsymbol{o}_t|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\right).\end{aligned} \tag{4.19}$$

Budeme-li uvažovat hodnoty $\gamma_{t,c}$ jako neměnné, je možné provést maximalizaci (4.19) vzhledem k parametrům GMM $\{w_c,\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c\}_{c=1}^{C}$ položením derivací podle příslušných proměnných rovných nule. V případě vah komponent je nutné respektovat podmínku $\sum_{c=1}^{C} w_c = 1$ čehož lze dosáhnout zavedením příslušného Lagrangeova multiplikátoru (Bishop, 2006).

EM algoritmus pro získání odhadu parametrů GMM metodou maximální věrohodnosti lze shrnout v následujících krocích:

1. Inicializace hodnot vah w_c, vektorů středních hodnot $\boldsymbol{\mu}_c$ a kovariančních matic $\boldsymbol{\Sigma}_c$ jednotlivých komponent. Pro inicializaci vektorů středních hodnot je možné aplikovat například dobře známý K-means algoritmus (Bishop, 2006) používaný pro shlukovou analýzu dat. Inicializační kovarianční matice je pak možné stanovit na základě dat přiřazených do jednotlivých shluků.

2. E krok: vyhodnocení okupačních pravděpodobností $\gamma_{t,c}$ podle (4.18) a hodnoty očekávané věrohodnosti $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ podle (4.19).

3. M krok: výpočet nových odhadů parametrů modelu:

$$\boldsymbol{\mu}_c^{ML} = \frac{\sum_{t=1}^{T}\gamma_{t,c}\boldsymbol{o}_t}{\sum_{t=1}^{T}\gamma_{t,c}} \tag{4.20}$$

$$\boldsymbol{\Sigma}_c^{ML} = \frac{\sum_{t=1}^{T}\gamma_{t,c}(\boldsymbol{o}_t-\boldsymbol{\mu}_c^{ML})(\boldsymbol{o}_t-\boldsymbol{\mu}_c^{ML})^T}{\sum_{t=1}^{T}\gamma_{t,c}} \tag{4.21}$$

$$w_c^{ML} = \frac{1}{T}\sum_{t=1}^{T}\gamma_{t,c}. \tag{4.22}$$

4. Ukončení v případě konvergence, v opačném případě návrat do kroku 2.

V praxi je výhodné provést v rámci E kroku pro postupně předkládaná data $\boldsymbol{o}_t$ akumulaci

[8] Obecně platí $\mathbb{E}[g(x)] = \sum_{i=1}^{\infty} g(x_i)p_i$, kde p_i je pravděpodobnost výskytu x_i.

následujících proměnných:

$$N_c = \sum_{t=1}^{T} \gamma_{t,c} \tag{4.23a}$$

$$\boldsymbol{F}_c = \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t \tag{4.23b}$$

$$\boldsymbol{S}_c = \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t \boldsymbol{o}_t^T \tag{4.23c}$$

Tyto proměnné označíme v uvedeném pořadí jako *postačující statistiky*[9] nultého, prvního a druhého řádu. Je zřejmé, že N_c je skalár, $\boldsymbol{F}_c$ je F-dimenzionální vektor a $\boldsymbol{S}_c$ je matice rozměru $F \times F$. V případě v praxi běžněji používaných GMM s diagonálními kovariančními maticemi však stačí provést výpočet v podobě $\boldsymbol{S}_c = \sum_{t=1}^{T} \gamma_{t,c} \operatorname{diag}(\boldsymbol{o}_t \boldsymbol{o}_t^T)$. Rozměr $\boldsymbol{S}_c$ pak odpovídá F-dimenzionálnímu vektoru[10]. Na základě uvedených postačujících statistik je možné v rámci M kroku provést výpočet nových parametrů modelu podle vztahů:

$$\boldsymbol{\mu}_c^{ML} = \frac{1}{N_c} \boldsymbol{F}_c \tag{4.24}$$

$$\boldsymbol{\Sigma}_c^{ML} = \frac{1}{N_c} \boldsymbol{S}_c - \boldsymbol{\mu}_c^{ML} (\boldsymbol{\mu}_c^{ML})^T \tag{4.25}$$

$$w_c^{ML} = \frac{N_c}{T}. \tag{4.26}$$

Dále definujme *centralizované postačující statistiky* prvního a druhého řádu

$$\begin{aligned} \tilde{\boldsymbol{F}}_c(\boldsymbol{\mu}_c) &= \sum_{t=1}^{T} \gamma_{t,c} (\boldsymbol{o}_t - \boldsymbol{\mu}_c) \\ &= \boldsymbol{F}_c - N_c \boldsymbol{\mu}_c \end{aligned} \tag{4.27a}$$

$$\begin{aligned} \tilde{\boldsymbol{S}}_c(\boldsymbol{\mu}_c) &= \sum_{t=1}^{T} \gamma_{t,c} (\boldsymbol{o}_t - \boldsymbol{\mu}_c)(\boldsymbol{o}_t - \boldsymbol{\mu}_c)^T \\ &= \boldsymbol{S}_c - \boldsymbol{F}_c \boldsymbol{\mu}_c^T - \boldsymbol{\mu}_c \boldsymbol{F}_c^T + N_c \boldsymbol{\mu}_c \boldsymbol{\mu}_c^T. \end{aligned} \tag{4.27b}$$

Na jejich základě je možné vyjádřit očekávanou hodnotu logaritmu věrohodnostní funkce pro kompletní data (4.19) jako

$$\mathbb{E}_{\mathcal{Z}|\boldsymbol{O},\hat{\boldsymbol{\theta}}}[\ell(\boldsymbol{\theta}|\boldsymbol{O},\mathcal{Z})] = \sum_{c=1}^{C} N_c \log \frac{w_c}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} - \frac{1}{2} \sum_{c=1}^{C} \operatorname{tr}(\boldsymbol{\Sigma}_c^{-1} \tilde{\boldsymbol{S}}_c(\boldsymbol{\mu}_c)), \tag{4.28}$$

kde tr(.) značí stopu matice.

Nepříjemnou vlastností ML odhadu parametrů GMM je jeho náchylnost k přetrénování (overfitting). Projevem tohoto jevu je například vznik singulárních kovariančních matic. Pro ilustraci

[9] Jedná se o základní charakteristiky, které jsou nutné pro určení odhadu parametrů modelu.

[10] Pro přehlednost zápisu je i v tomto případě zacházeno v dalším výkladu s $\boldsymbol{S}_c$ jako s maticí, praktický výpočet se však zjednodušuje.

uvažujme například následující situaci. V případě nějaké komponenty ležící na okraji prostoru příznaků je možné, že v rámci E kroku jedné z iterací EM algoritmu bude této komponentě s nenulovou okupační pravděpodobností přiřazen pouze jeden z trénovacích vektorů. Dochází tak ke „zborcení“ komponenty do jednoho bodu a výsledkem M kroku pak bude nulová (singulární) kovarianční matice této komponenty. K odhadu singulárních (nebo singulárním blízkých) kovariančních matic však ve skutečnosti běžně dochází i za méně extrémních situací. V praxi je tak nutné aplikovat vhodné heuristické přístupy kontrolující například číslo podmíněnosti kovariančních matic komponent[11]. V případě GMM s diagonálními kovariančními maticemi je možné rozpoznat bortící se komponentu například na základě kontroly minimálního rozptylu jednotlivých složek. Bortící se komponentu je možné přesunout na jiné náhodně zvolené místo v příznakovém prostoru s nově inicializovanou kovarianční maticí nebo zcela vypustit z modelu. Lze očekávat, že efektem obou operací bude snížení celkové věrohodnosti pro trénovací data, které porušuje tvrzení o neklesající věrohodnosti mezi jednotlivými iteracemi EM algoritmu. Cílem je však v tomto případě zabránění problému přetrénování modelu na trénovací data.

Nepříliš vítanou vlastností ML odhadu také je, že ke každému maximálně věrohodnému odhadu parametrů pro model s C komponentami existuje $C!$ ekvivalentních řešení, které se liší pouze přiřazením parametrů k jednotlivým komponentám (to lze provést $C!$ způsoby). V *prostoru parametrů* tak pro každý bod (představující řešení maximálně věrohodného odhadu parametrů) existuje dalších $C! - 1$ bodů, které vedou na shodné rozložení reprezentované směsí.

4.2.3 Odhad parametrů metodou maximální aposteriorní pravděpodobnosti

Odhad parametrů GMM metodou maximální aposteriorní pravděpodobnosti představuje také problém nekompletních dat. Opět totiž není známa příslušnost trénovacích dat k jednotlivým komponentám směsi. I v tomto případě představuje možné řešení aplikace EM algoritmu (Dempster et al., 1977). E krok je shodný pro ML i MAP odhad parametrů a jeho cílem je tak stanovit hodnotu pomocné funkce $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ podle (4.14). V rámci E kroku dochází také k akumulaci postačujících statistik N_c, $\boldsymbol{F}_c$ a $\boldsymbol{S}_c$ podle vztahů (4.23). Cílem M kroku je v případě MAP odhadu parametrů GMM maximalizace funkce $Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) + P(\hat{\boldsymbol{\theta}})$. Od ML odhadu se tak liší zavedením regularizačního členu $P(\hat{\boldsymbol{\theta}})$.

Klíčové otázky představují volba vhodného apriorního rozložení a určení jeho parametrů. Přirozenou volbou je konjugované apriorní rozložení[12] (Gauvain – Lee, 1994), pro které je možné vyjádřit posteriorní rozložení v uzavřené formě. Konjugované apriorní rozložení pro váhy komponent GMM představuje Dirichletovo rozložení. Pro parametry vektoru středních hodnot a inverzní kovarianční

[11] Čím je číslo podmíněnosti (condition number) větší, tím více se matice blíží singulární matici.

[12] Pokud platí, že posteriorní rozložení $P(\boldsymbol{\theta}|\boldsymbol{O})$ je reprezentováno, až na rozdílné parametry, stejnou funkcí jako apriorní rozložení $P(\boldsymbol{\theta})$, říkáme, že posteriorní a apriorní rozložení představují konjugovaná rozložení. Apriorní rozložení, které splňuje tuto podmínku s ohledem k dané věrohodnostní funkci $\boldsymbol{\theta} \mapsto P(\boldsymbol{O}|\boldsymbol{\theta})$ se označuje jako konjugované apriorní rozložení. Například pro Gaussovskou věrohodnostní funkci a Gaussovské apriorní rozložení je posteriorní rozložení opět Gaussovské.

matice normálního vícerozměrného rozložení je to pak Gaussovo-Wishartovo rozložení.

Parametry GMM jsou na základě metody maximální aposteriorní pravděpodobnosti v M kroku aktualizovány podle následujících vztahů (Gauvain – Lee, 1994):

$$\boldsymbol{\mu}_c^{MAP} = \frac{\tau_c \boldsymbol{\nu}_c + \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t}{\tau_c + \sum_{t=1}^{T} \gamma_{t,c}} \tag{4.29}$$

$$(\boldsymbol{\Sigma}_c^{MAP})^{-1} = \frac{\boldsymbol{\Psi}_c + \sum_{t=1}^{T} \gamma_{t,c} (\boldsymbol{o}_t - \boldsymbol{\mu}_c^{MAP})(\boldsymbol{o}_t - \boldsymbol{\mu}_c^{MAP})^T + \tau_c (\boldsymbol{\nu}_c - \boldsymbol{\mu}_c^{MAP})(\boldsymbol{\nu}_c - \boldsymbol{\mu}_c^{MAP})^T}{(\alpha_c - F) + \sum_{t=1}^{T} \gamma_{t,c}} \tag{4.30}$$

$$w_c^{MAP} = \frac{(\zeta_c - 1) + \sum_{t=1}^{T} \gamma_{t,c}}{\sum_{k=1}^{C} \left((\zeta_k - 1) + \sum_{t=1}^{T} \gamma_{t,k} \right)} \tag{4.31}$$

kde $\{\zeta_c\}_{c=1}^{C}$ představují parametry Dirichletova apriorního rozložení a $\{\tau_c, \boldsymbol{\nu}_c, \alpha_c, \boldsymbol{\Psi}_c\}_{c=1}^{C}$ parametry Gaussova-Wishartova apriorního rozložení, kde $\boldsymbol{\nu}_c$ je F-rozměrný vektor a $\boldsymbol{\Psi}_c$ pozitivně definitní matice rozměru $F \times F$. Parametry musí splňovat podmínky $\zeta_c > 0$, $\alpha_c > F - 1$ a $\tau_c > 0$. Parametry apriorních rozložení se pro odlišení od parametrů modelu označují jako *hyperparametry*. Nelze-li považovat hyperparametry za známé, například v důsledku teoretické úvahy o zkoumaném procesu, je možné provést jejich maximálně věrohodný odhad přímo na základě pozorovaných dat. Jedná se však o netriviální úlohu. Jednou z příčin možných komplikací je také fakt, že počet hyperparametrů převyšuje počet parametrů samotného modelu. V praxi jsou tak často aplikována různá omezení a svazování hodnot hyperparametrů. Část apriorní znalosti může být také začleněna přímo do modelu v podobě předpokladu, že některé parametry jsou známé a neměnné. Tím dochází ke zvýšení robustnosti odhadu ostatních parametrů.

V případě MAP odhadu parametrů GMM jsou často vybrané hyperparametry položeny rovny parametrům nějakého známého modelu $\hat{\boldsymbol{\theta}} = \{\hat{w}_c, \hat{\boldsymbol{\mu}}_c, \hat{\boldsymbol{\Sigma}}_c\}_{c=1}^{C}$. V takové situaci hovoříme zpravidla o adaptaci modelu. Metody rozpoznávání mluvčích zpravidla využívají pouze adaptaci vektorů středních hodnot. Jednou z motivací je malé množství dat, které je typicky k dispozici pro provedení MAP odhadu. To komplikuje získání robustního odhadu všech parametrů, zejména parametrů kovariančních matic. Ovšem fakt, že modely odvozené adaptací od jednoho modelu se v ostatních parametrech neliší, má i další výhody a aplikace, které budou patrné z pozdějšího výkladu. Položením $\boldsymbol{\nu}_c = \hat{\boldsymbol{\mu}}_c$ je možné přepsat vztah (4.29) ve tvaru

$$\boldsymbol{\mu}_c^{MAP} = \frac{\tau_c \hat{\boldsymbol{\mu}}_c + \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t}{\tau_c + \sum_{t=1}^{T} \gamma_{t,c}}. \tag{4.32}$$

Je možné pozorovat, že MAP odhad $\boldsymbol{\mu}_c^{MAP}$ odpovídá váženému součtu původního odhadu $\hat{\boldsymbol{\mu}}_c$ a nového ML odhadu stanoveného na základě trénovacích dat $\boldsymbol{O}$ (viz (4.20)). Vliv odhadu stanoveného na základě předložených dat se přitom zvyšuje s množstvím těchto dat, pro $T \to \infty$ je $\sum_{t=1}^{T} \gamma_{t,c} \to \infty$ a MAP odhad parametrů se tak asymptoticky blíží ML odhadu. Naopak, běžnou situací v rámci MAP odhadu je, že v rámci E kroku nejsou komponentě c přiřazena žádná data. Platí tedy $\sum_{t=1}^{T} \gamma_{t,c} = 0$. Pak odpovídá MAP odhad parametrům původního modelu $\boldsymbol{\mu}_c^{MAP} = \hat{\boldsymbol{\mu}}_c$. Toto chování je velmi žádoucí, protože v případě komponent, pro které je k dispozici dostatek dat

a lze tak předpokládat získání spolehlivého odhadu parametrů, bude zdůrazňován nový odhad založený na adaptačních datech, zatímco v případě komponent, kterým bylo přiděleno malé množství dat, bude více spoléháno na původní parametry.

Zaměřme se nyní na význam hodnoty τ_c. Za předpokladu, že bychom pro jednotlivé komponenty modelu $\hat{\boldsymbol{\theta}}$ znali hodnoty $\hat{N}_c = \sum_{t=1}^{T} \gamma_{t,c}^{bkg}$, určené podle (4.23a) při trénování tohoto modelu na základě dat $\boldsymbol{O}^{bkg} = \{\boldsymbol{o}_1^{bkg}, \ldots, \boldsymbol{o}_{T^{bkg}}^{bkg}\}$, a položili $\tau_c = \hat{N}_c$, získáme vztah

$$\boldsymbol{\mu}_c^{MAP} = \frac{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg} \hat{\boldsymbol{\mu}}_c + \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t}{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg} + \sum_{t=1}^{T} \gamma_{t,c}}, \tag{4.33}$$

kde

$$\hat{\boldsymbol{\mu}}_c = \frac{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg} \boldsymbol{o}_t^{bkg}}{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg}} \tag{4.34}$$

Dosazením (4.34) do (4.33) dostáváme

$$\boldsymbol{\mu}_c^{MAP} = \frac{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg} \boldsymbol{o}_t^{bkg} + \sum_{t=1}^{T} \gamma_{t,c} \boldsymbol{o}_t}{\sum_{t=1}^{T^{bkg}} \gamma_{t,c}^{bkg} + \sum_{t=1}^{T} \gamma_{t,c}}. \tag{4.35}$$

Vztah (4.35) tak odpovídá ML odhadu parametrů vektoru středních hodnot při sloučení dat $\boldsymbol{O}^{bkg}$ (použitých k natrénování původního modelu $\hat{\boldsymbol{\theta}}$) a adaptačních dat $\boldsymbol{O}$. Pro adaptaci se však ve skutečnosti používají nižší hodnoty τ_c než by odpovídaly $\hat{N}_c$, typicky se volí hodnoty v rozmezí 1 až 20. S ohledem na množství dat $\boldsymbol{O}^{bkg}$, které zpravidla značně převyšuje množství adaptačních dat $\boldsymbol{O}$, je tím pádem vliv adaptačních dat posílen. Bude-li hodnota $\tau_c = 1$, znamená to, že po přiřazení jednoho příznakového vektoru ke komponentě c v rámci E kroku se bude MAP odhad nacházet právě uprostřed mezi tímto vektorem a původním odhadem $\hat{\boldsymbol{\mu}}_c$. Pro zjednodušení je často uvažována shodná hodnota τ_c pro všechny komponenty, tj. $\tau_c = \tau$. Hodnota τ bývá označována jako *relevanční faktor* (Reynolds et al., 2000) a v souvislosti s tímto označením hovoříme o relevančním MAP odhadu parametrů GMM.

Pro snazší interpretovatelnost relevančního faktoru τ jako váhy řídící vliv mezi ML odhadem a původním odhadem parametrů definujme proměnou β_c jako (Van Vuuren, 1999)

$$\beta_c = \frac{\sum_{t=1}^{T} \gamma_{t,c}}{\sum_{t=1}^{T} \gamma_{t,c} + \tau}, \tag{4.36}$$

vztah (4.32) je pak možné přepsat ve tvaru

$$\boldsymbol{\mu}_c^{MAP} = (1 - \beta_c)\hat{\boldsymbol{\mu}}_c + \beta_c \boldsymbol{\mu}_c^{ML}, \tag{4.37}$$

kde $\boldsymbol{\mu}_c^{ML}$ je určeno podle (4.20), příp. (4.24). Z uvedeného vyjádření je patrné, že s rostoucí hodnotou relevančního faktoru τ klesá β_c a tím pádem se snižuje vliv odhadu založeného na adaptačních datech, a naopak.

4.2.4 Aplikace ML a MAP odhadu parametrů GMM

Metoda maximálně věrohodného odhadu parametrů GMM se v systémech pro rozpoznávání mluvčích využívá především pro natrénování univerzálního hlasového modelu, označovaného zpravidla jako UBM (Universal Background Model). Cílem je získat model reprezentující rozložení příznakových vektorů v prostoru nezávislé na mluvčím. UBM je tak trénován na řečových datech mnoha mluvčích (typicky jsou použity desítky až stovky hodin nahrávek) a v důsledku velkého objemu dat lze předpokládat získání robustního odhadu parametrů i pro modely s vysokým počtem komponent. Limitujícím faktorem při volbě počtu komponent pak již nemusí být množství dostupných dat ale také výpočetní nároky práce s takovým modelem. V době vzniku této práce lze považovat za typický univerzální model s 1024 nebo 2048 komponentami[13].

Výběr trénovacích dat pro UBM by měl odpovídat doméně, ve které bude systém nasazen. Pokud však není tato doména (ve smyslu pohlaví mluvčích, typu používaných mikrofonů nebo přenosových kanálů) předem známa, je vhodné zahrnout do trénovacích dat nahrávky různého charakteru. Přitom je nutné sledovat, aby nahrávky odpovídající různým akustickým podmínkám, stejně jako nahrávky obou pohlaví, byly v trénovacích datech zastoupeny pokud možno rovnoměrně. Protože v opačném případě by mohlo dojít k výraznému vychýlení modelu směrem k převládající skupině dat. V praxi lze tento problém řešit trénováním modelů pro různé skupiny dat (např. v závislosti na pohlaví mluvčích) odděleně a následným sloučením těchto modelů (Reynolds et al., 2000).

Přestože je možné použít ML odhad parametrů GMM i pro trénování modelů jednotlivých mluvčích, s ohledem na typické množství dostupných trénovacích dat nelze předpokládat získání spolehlivého odhadu parametrů pro všechny zapisované mluvčí. Při volbě modelů s vyšším počtem komponent lze očekávat častý výskyt problému přetrénování. Na druhé straně modely s malým počtem komponent nemusí poskytovat dostatečnou variabilitu pro popis všech specifických charakteristik řeči mluvčích.

Pro odvození modelů mluvčích je zpravidla využívána MAP adaptace UBM modelu. Nevýhodou MAP odhadu je, že adaptovány jsou pouze komponenty odpovídající řečovým jevům přítomným v adaptačních datech (tato nevýhoda je ovšem společná i pro ML odhad). Komponenty odpovídající neviděným jevům zůstávají neadaptovány. V případě výskytu těchto jevů v rozpoznávané promluvě je jejich příspěvek k celkové hodnotě skóre téměř nulový. Tento problém nastává zejména, je-li k dispozici velmi malé množství dat pro stanovení modelu mluvčího.

UBM-GMM systém

V souladu s (Reynolds et al., 2000) bude systém založený na výše popsaném schématu, tedy využívající modely mluvčích odvozené adaptací UBM modelu metodou MAP, v rámci této práce označen jako UBM-GMM systém. Jedna z výhod provázanosti UBM modelu s modely mluvčích

[13] Volba počtu komponent odpovídající mocnině 2 však není v žádném případě nutností a je dána spíše zvyklostí.

spočívá v možnosti snížení paměťových nároků na uložení modelů mluvčích. Lze totiž očekávat, že vektory středních hodnot budou adaptovány pouze pro část komponent a ostatní budou převzaty beze změny. Stačí tak uložit pouze rozdíl adaptovaného modelu oproti UBM.

V procesu rozpoznávání je míra podpory hypotézy, že řeč v předložené nahrávce (příp. řečovém segmentu) reprezentované sekvencí příznakových vektorů $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ pochází od ověřovaného řečníka s reprezentovaného modelem $\boldsymbol{\theta}_s$, stanovena jako

$$s_{UBM-GMM} = \frac{1}{T}\sum_{t=1}^{T}\left(\log P(\boldsymbol{o}_t|\boldsymbol{\theta}_s) - \log P(\boldsymbol{o}_t|\boldsymbol{\theta}_{UBM})\right). \tag{4.38}$$

Skóre tedy odpovídá průměrné hodnotě logaritmu poměru věrohodností[14] GMM určených podle (4.8). UBM model zde reprezentuje akustický prostor možných neoprávněných mluvčích.

Vzájemné provázanosti modelů mluvčích a UBM lze využít pro zrychlené vyhodnocení (4.38) označované jako top-C skórování (Reynolds et al., 2000; Brummer et al., 2007). Na celkové věrohodnosti příznakového vektoru stanovené dle (4.8) se totiž převážnou měrou podílí vždy pouze malý počet komponent a příspěvek ostatních je zanedbatelný. V důsledku odvození modelů mluvčích adaptací UBM lze předpokládat, že komponenty UBM s nejvyšším příspěvkem budou mít (resp. jejich adaptované protějšky) nejvyšší příspěvek i v případě modelů mluvčích, protože jsou si vzájemně blízké. Pro každý příznakový vektor rozpoznávané sekvence je tak na základě UBM identifikováno k (typicky se volí k v rozmezí 5 až 10) komponent s nejvyšší okupační pravděpodobností a při výpočtu věrohodnosti jsou uvažovány pouze tyto komponenty.

V případě UBM-GMM systému lze ve stručnosti shrnout náplň etap uvedených v podkapitole 1.2 následovně:

- *Učení systému* – spočívá v natrénování univerzální modelu (UBM) metodou ML na základě nahrávek od velkého množství mluvčích a volbě relevančního faktoru τ.

- *Zapsání mluvčích* – spočívá v odvození modelů mluvčích na základě adaptace vektorů středních hodnot UBM modelu metodou MAP dle (4.37).

- *Rozpoznávání* – protože je uvažována úloha detekce mluvčích, je cílem systému vyjádřit míru podpory hypotézy, že předložená nahrávka byla skutečně vyslovena domnělým mluvčím. Skóre určené UBM-GMM systémem je stanoveno dle (4.38).

[14] Přestože je skóre v tomto případě definováno přímo jako (průměrný) logaritmus poměru věrohodností, je u UBM-GMM systémů zpravidla nutné provést jeho následnou kalibraci pro získání hodnoty LLR ve smyslu diskutovaném v podkapitole 2.2.6. Nekalibrovaný výstup systémů je v rámci této práce pro větší jednoznačnost označován jako skóre. Upozorněme, že před provedením kalibrace bývá v praxi aplikována některá z forem normalizace skóre, např. T-norm nebo Z-norm, případně jejich kombinace.

4.3 Faktorová analýza

Faktorová analýza je oblíbeným statistickým nástrojem v různých vědních disciplínách včetně psychologie (resp. psychometrie)[15], sociologie nebo marketingu. V nedávné době se jí však, především zásluhou Patricka Kennyho (Kenny, 2005), dostalo velké popularity i v problematice automatického rozpoznávání mluvčích. Systémy založené na faktorové analýze dosáhly dobrých výsledků v posledních ročnících NIST evaluací pořádaných v letech 2008 a 2010.

Cílem faktorové analýzy je provést vyjádření variability D pozorovatelných proměnných $x_1, \ldots, x_D$ na základě menšího počtu M skrytých (latentních) proměnných $z_1, \ldots, z_M$, tzv. *faktorů*. Pozorovatelné proměnné jsou v rámci faktorové analýzy modelovány na základě generativního procesu (Bishop, 2006)

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\epsilon}, \tag{4.39}$$

kde $\boldsymbol{x} = [x_1, \ldots, x_D]^T$, $\boldsymbol{z} = [z_1, \ldots, z_M]^T$, $\boldsymbol{W}$ je tzv. *zátěžová matice* rozměru $D \times M$, $\boldsymbol{\mu}$ je konstantní D-rozměrný vektor a $\boldsymbol{\epsilon}$ je D-rozměrná proměnná s normálním rozložením s nulovou střední hodnotou a diagonální kovarianční maticí $\boldsymbol{\Sigma}$. Pozorovatelné proměnné jsou tedy tvořeny lineární transformací $\boldsymbol{z}$ a přídavného šumu reprezentovaného $\boldsymbol{\epsilon}$.

Podmíněné rozložení pozorovatelných proměnných $\boldsymbol{x}$ za předpokladu znalosti skrytých proměnných $\boldsymbol{z}$ odpovídá

$$P(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{W}\boldsymbol{z} + \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{4.40}$$

Apriorní rozložení skrytých proměnných je definováno jako normální s nulovou střední hodnotou a jednotkovou kovarianční maticí, tedy

$$P(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I}). \tag{4.41}$$

Pro stanovení ML odhadu parametrů $\boldsymbol{\mu}$, $\boldsymbol{W}$ a $\boldsymbol{\Sigma}$ je nutné vyjádřit marginální rozložení pozorovatelných dat $P(\boldsymbol{x})$. To vyžaduje provedení marginalizace přes skryté proměnné, tedy

$$P(\boldsymbol{x}) = \int P(\boldsymbol{x}|\boldsymbol{z})P(\boldsymbol{z})d\boldsymbol{z}. \tag{4.42}$$

Výsledkem je opět normální rozložení (Bishop, 2006)

$$P(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{\Sigma}). \tag{4.43}$$

Model faktorové analýzy tedy popisuje kovarianci pozorovatelných proměnných jednak pomocí nezávislých rozptylů (reprezentovaných pro každou pozorovatelnou proměnnou příslušným diagonálním členem matice $\boldsymbol{\Sigma}$) a dále kovariancí mezi proměnnými reprezentovanou maticí $\boldsymbol{W}$. Sloupce matice $\boldsymbol{W}$ bývají označovány jako *faktorové zátěže* a definují podprostor v prostoru pozorovatelných proměnných.

[15] Průkopnictví v aplikaci faktorové analýzy je přisuzováno Charlesi Spearmanovi, který ji poprvé použil pro testy inteligence (Spearman, 1904).

Odhad parametrů nelze v důsledku skrytých proměnných vyjádřit v uzavřené formě a možné řešení tak opět spočívá v aplikaci EM algoritmu. Konkrétní výpočet parametrů modelu bude uveden později.

Faktorová analýza úzce souvisí s pravděpodobnostní analýzou hlavních komponent (PPCA) (Bishop, 2006), kdy rozdíl spočívá v definici podmíněného rozložení $P(\boldsymbol{x}|\boldsymbol{z})$. Zatímco v případě faktorové analýzy je uvažována pro toto rozložení diagonální kovarianční matice $\boldsymbol{\Sigma}$, v případě PPCA je uvažována izotropní kovarianční matice $\sigma^2\boldsymbol{I}$, kde $\boldsymbol{I}$ je jednotková matice. Oba modely tak shodně předpokládají nezávislost pozorovatelných proměnných $x_1, \ldots, x_D$ za předpokladu, že jsou známy skryté proměnné $\boldsymbol{z}$.

4.3.1 Koncept supervektorů

V publikacích zabývajících se problematikou rozpoznávání mluvčích se pojmem *supervektor* zpravidla označuje vektor vytvořený zřetězením vektorů středních hodnot komponent GMM. Za předpokladu F-dimenzionálních příznakových vektorů a modelu s C komponentami bude rozměr *GMM supervektoru* CF. Kinnunen – Li (2010) pak nabízí širší interpretaci supervektoru jako způsobu reprezentace nahrávek různé délky prostřednictvím vektoru vysokého ale konstantního rozměru. Právě proměnná délka sekvence příznakových vektorů komplikuje aplikaci některých typů klasifikátorů. Snahy o rozpoznávání mluvčích na základě souboru statistik příznaků (typicky odvozených ze spektra signálu nebo základní frekvence) vyhodnocených přes celou délku nahrávky, tvořících tak reprezentaci nahrávky v podobě vektoru pevného rozměru, byly typické především pro rané studie zabývající se problematikou textově nezávislého rozpoznávání mluvčích (Bricker et al., 1971; Furui, 2009). Přesto lze i v relativně nedávné literatuře nalézt pokusy o tento přístup (Kinnunen et al., 2006). Výhodou těchto metod jsou jednoznačně nízké výpočetní nároky, avšak jak dokládá právě Kinnunen et al. (2006), úspěšnost těchto systémů značně zaostává za autory zvoleným UBM-GMM referenčním systémem využívajícím MFCC příznaky.

V nedávné době byl zájem o reprezentaci nahrávek pomocí vektorů konstantního rozměru motivován mimo jiné snahou o aplikaci SVM klasifikátorů (Fine et al., 2001; Campbell, 2002; Smith – Gales, 2002; Wan – Renals, 2005; Campbell et al., 2006b). Schopnost reprezentovat nahrávky libovolné délky jako bod ve společném prostoru však nabízí i další uplatnění. Řada metod pro kompenzaci variability akustických podmínek (Kenny, 2005; Burget et al., 2007; Vogt – Sridharan, 2008) je založena na předpokladu, že je možné provést dekompozici supervektoru $\boldsymbol{m}$, reprezentujícího danou nahrávku, mezi supervektor $\boldsymbol{s}$ specifický pro mluvčího, který je neměnný pro všechny jeho nahrávky, a supervektor $\boldsymbol{c}$, který odpovídá akustickým podmínkám dané nahrávky, tedy

$$\boldsymbol{m} = \boldsymbol{s} + \boldsymbol{c}. \tag{4.44}$$

4.3.2 Sdružená faktorová analýza

Model *sdružené faktorové analýzy* (JFA) (Kenny, 2005) představuje generativní model GMM supervektorů (tvořených zřetězením vektorů středních hodnot komponent GMM). Na rozdíl od modelu (4.39) zmíněného v úvodu kapitoly nepředstavují supervektory přímo pozorovatelné proměnné, ty jsou představovány sekvencemi F-rozměrných příznakových vektorů, které reprezentují nahrávky. V případě GMM modelu tvořeného C Gaussovskými komponentami bude rozměr GMM supervektoru CF (jeden z typických rozměrů akustických příznakových vektorů je 39, pro model s 1024 komponentami pak bude rozměr supervektoru 39 936). JFA model navíc na rozdíl od modelu (4.39) nepracuje pouze s jednou zátěžovou maticí, ale rozlišuje mezi faktory specifickými pro variabilitu hlasu mluvčích a faktory specifickými pro variabilitu akustických podmínek. Supervektor $\boldsymbol{m}_{r,s}$ příslušný r-té nahrávce mluvčího s je modelován generativním procesem

$$\boldsymbol{m}_{r,s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s + \boldsymbol{U}\boldsymbol{w}_{r,s}, \tag{4.45}$$

kde význam jednotlivých proměnných je následující:

- $\boldsymbol{\mu}$ je CF-dimenzionální supervektor, který je nezávislý na mluvčím i nahrávce,
- $\boldsymbol{y}_s$ je D_{spk}-dimenzionální vektor faktorů odpovídajících variabilitě mluvčích (D_{spk} je typicky v řádu stovek) a $\boldsymbol{V}$ je příslušná zátěžová matice rozměru $CF \times D_{spk}$,
- $\boldsymbol{z}_s$ je CF-dimenzionální vektor faktorů, které také odpovídají variabilitě mluvčích, příslušná zátěžová matice $\boldsymbol{D}$ má rozměr $CF \times CF$ a je diagonální,
- supervektor $\boldsymbol{s}_s$ charakterizující v souladu s (4.44) mluvčího s, nezávisle na konkrétní nahrávce, lze vyjádřit na základě uvedených členů jako

$$\boldsymbol{s}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s, \tag{4.46}$$

- $\boldsymbol{w}_{r,s}$ je D_{chan}-dimenzionální vektor faktorů odpovídajících variabilitě akustických podmínek (D_{chan} je typicky v řádu stovek) a $\boldsymbol{U}$ je příslušná zátěžová matice rozměru $CF \times D_{chan}$,
- supervektor $\boldsymbol{c}_{r,s}$, který dle dekompozice (4.44) reflektuje akustické podmínky dané nahrávky, je pak stanoven jako

$$\boldsymbol{c}_{r,s} = \boldsymbol{U}\boldsymbol{w}_{r,s}, \tag{4.47}$$

- apriorní rozložení všech faktorů jsou předpokládána jako standardní normální (s nulovou střední hodnotou a jednotkovou kovarianční maticí), platí tedy $P(\boldsymbol{y}_s) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $P(\boldsymbol{z}_s) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ a $P(\boldsymbol{w}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$,
- rozložení supervektorů $\boldsymbol{s}_s$ pak odpovídá normálnímu rozložení

$$P(\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{s}_s | \boldsymbol{\mu}, \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{D}^2) \tag{4.48}$$

a rozložení supervektorů $\boldsymbol{m}_{r,s}$ normálnímu rozložení

$$P(\boldsymbol{m}_{r,s}|\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{m}_{r,s}|\boldsymbol{s}_s, \boldsymbol{U}\boldsymbol{U}^T). \tag{4.49}$$

Model sdružené faktorové analýzy je souhrnně reprezentován pěticí hyperparametrů $\boldsymbol{\Lambda} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$. Dosud neuvedená matice $\boldsymbol{\Sigma}$ je diagonální kovarianční matice rozměru $CF \times CF$, která je tvořena diagonálními bloky $\boldsymbol{\Sigma}_c$, kde $c = 1, \ldots, C$. Význam této matice je následující. Jak již bylo uvedeno, supervektor $\boldsymbol{m}_{r,s}$ nepředstavuje pozorovatelná data. Pozorovatelná je sekvence příznakových vektorů $\boldsymbol{o}_t$ reprezentující nahrávku r mluvčího s, jejichž rozložení je stále uvažováno ve formě směsi Gaussových rozložení. Označíme-li část vektoru $\boldsymbol{m}_{r,s}$, která odpovídá komponentě c jako $\boldsymbol{m}_{r,s,c}$, je rozložení vektorů $\boldsymbol{o}_t$ touto komponentou reprezentováno normálním rozložením $\mathcal{N}(\boldsymbol{o}_t|\boldsymbol{m}_{r,s,c}, \boldsymbol{\Sigma}_c)$.

V literatuře zabývající se rozpoznáváním mluvčích se lze setkat s několika variantami modelu sdružené faktorové analýzy. Ty se od modelu (4.45) liší například absencí některých faktorů. Výkladem významu jednotlivých členů JFA modelu se detailněji zabývá podkapitola 4.3.5.

Schéma JFA systému

Existuje řada způsobů implementace JFA systému, pokusíme-li se však nalézt společné aspekty implementací vzhledem k základním etapám definovaným v podkapitole 1.2, bude jejich stručná charakteristika následovná:

- *Učení systému* – nejprve je natrénován UBM model, který definuje prostor GMM supervektorů. Dále je nutné stanovit hyperparametry JFA modelu $\boldsymbol{\Lambda}$ na dostatečně rozsáhlé databázi. Je nutné, aby bylo v trénovací databázi pro jednotlivé mluvčí k dispozici několik nahrávek. Odhad hyperparametrů je prováděn iterativním EM algoritmem, který zaručuje, že celková věrohodnost trénovacích dat se mezi jednotlivými iteracemi zvyšuje. Celková věrohodnost je rovna $\prod_s P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda})$, kde s je bráno přes všechny mluvčí v trénovací databázi a $\underline{\boldsymbol{O}}_s$ představuje všechny nahrávky mluvčího s.

- *Zapsání mluvčích* – lze uvažovat dva způsoby reprezentace mluvčích:

  1. pomocí JFA modelu s hyperparametry specifickými pro mluvčího, tj. $\boldsymbol{\Lambda}_s = \{\boldsymbol{\mu}_s, \boldsymbol{V}_s, \boldsymbol{D}_s, \boldsymbol{U}, \boldsymbol{\Sigma}\}$ (Kenny – Dumouchel, 2004a,b). Zatímco hyperparametry $\boldsymbol{\mu}$, $\boldsymbol{V}$ a $\boldsymbol{D}$ (nezávislé na mluvčím) slouží jako model rozdílů mezi mluvčími, úlohou $\boldsymbol{\mu}_s$, $\boldsymbol{V}_s$ a $\boldsymbol{D}_s$ je sloužit jako model posteriorního rozložení supervektoru $\boldsymbol{s}_s$ pro daného mluvčího s. Na základě trénovacích dat mluvčího s a hyperparametrů $\boldsymbol{\mu}$, $\boldsymbol{V}$ a $\boldsymbol{D}$ je nejprve stanoveno posteriorní rozložení $\boldsymbol{s}_s$ pro daného mluvčího. V případě tohoto posteriorního rozložení ovšem nemusí platit, že rozložení skrytých proměnných $\boldsymbol{y}_s$ a $\boldsymbol{z}_s$ jsou stále standardní normální. Následně je tak nalezeno rozložení, které odpovídá modelu ve tvaru $\boldsymbol{\mu}_s + \boldsymbol{V}_s\boldsymbol{y}_s + \boldsymbol{D}_s\boldsymbol{z}_s$, kde $\boldsymbol{y}_s$ a $\boldsymbol{z}_s$ mají standardní normální rozložení, a které je nejblíže

stanovenému posteriornímu rozložení $\boldsymbol{s}_s$ ve smyslu Kullback-Leiblerovy divergence. Potom $\boldsymbol{\mu}_s$ představuje odhad supervektoru specifického pro mluvčího a $\boldsymbol{V}_s$ a $\boldsymbol{D}_s$ reflektují nejistotu tohoto odhadu (Kenny et al., 2007a). Hyperparametry $\boldsymbol{U}$ a $\boldsymbol{\Sigma}$ zůstávají shodné s původním modelem nezávislým na mluvčím, protože zátěžové matice příslušené faktorům souvisejícím s variabilitou akustických podmínek by neměly být závislé na mluvčím.

2. stanovením bodového MAP odhadu supervektoru $\boldsymbol{s}_s$ na základě posteriorního rozložení skrytých proměnných (Kenny et al., 2007b, 2008b). Často je k dispozici omezené množství trénovacích dat a je tak vhodné uvažovat nejen očekávanou hodnotu supervektoru $\boldsymbol{s}_s$, tedy $\mathbb{E}[\boldsymbol{s}_s]$, ale také diagonální kovarianční matici $\mathrm{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$, která vychází z posteriorního rozložení skrytých proměnných pro trénovací data mluvčího s.

- *Rozpoznávání* – vyhodnocení skóre vyjadřujícího míru podpory hypotézy, že předložená nahrávka $\boldsymbol{O}$ byla skutečně vyslovena domnělým mluvčím s je závislá na použité reprezentaci mluvčích:

  1. v případě reprezentace prostřednictvím modelu $\boldsymbol{\Lambda}_s$ je provedeno vyhodnocení logaritmu poměru věrohodností v podobě

$$\frac{1}{T}\left(\log P(\boldsymbol{O}|\boldsymbol{\Lambda}_s) - \log P(\boldsymbol{O}|\boldsymbol{\Lambda})\right). \tag{4.50}$$

  2. v případě reprezentace modelu prostřednictvím supervektoru $\boldsymbol{s}_s$ (resp. $\mathbb{E}[\boldsymbol{s}_s]$ a $\mathrm{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$) se při vyhodnocení věrohodnosti pro testovanou nahrávku redukují skryté proměnné pouze na $\boldsymbol{w}_{r,s}$ (protože $\boldsymbol{y}_s$ a $\boldsymbol{z}_s$ jsou dané v rámci $\boldsymbol{s}_s$). Cílem je pak stanovit logaritmus poměru věrohodností

$$\frac{1}{T}\left(\log P(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{\Lambda}) - \log P(\boldsymbol{O}|\boldsymbol{\mu}, \boldsymbol{\Lambda})\right). \tag{4.51}$$

4.3.3 Metody odhadu hyperparametrů JFA modelu

Kenny (2005) navrhl dvě varianty EM algoritmu pro odhad hyperparametrů JFA modelu. První algoritmus aplikuje přístup založený na standardní metodě maximální věrohodnosti (ML) a druhý přístup založený na metodě minimální divergence (MD). Při odhadu hyperparametrů je zpravidla výhodné použití obou algoritmů. Nevýhodou ML algoritmu je totiž zpravidla pomalejší konvergence. MD algoritmus však předpokládá, že orientace prostorů specifických pro variabilitu mluvčích i akustických podmínek je známá a má být zachována, vyžaduje tak patřičnou inicializaci a ML odhad je v tomto případě vhodnou volbou.

Vhodné schéma odhadu hyperparametrů tak sestává z provedení několika iterací ML algoritmu a následně provedení alespoň jedné iterace MD algoritmu[16]. Vyhodnocením vlivu provedení MD odhadu s ohledem na celkovou věrohodnost se zabývá například Brummer (2010b) a porovnává také různé strategie postupné aplikace ML a MD algoritmů v průběhu trénování hyperparametrů.

[16] Vzhledem k rychlé konvergenci může být i jedna iterace dostatečná (Kenny et al., 2008b)

Nejlepších výsledků bylo dosaženo aplikací M kroku obou algoritmů (tj. provedení ML i MD odhadu v uvedeném pořadí) v každé iteraci EM algoritmu. Tato varianta je v následujícím textu označována jako ML-MD odhad.

Dříve než budou uvedeny výpočty prováděné v rámci těchto algoritmů, je vhodné provést odvození základních věrohodností a rozložení potřebných pro práci s JFA modelem.

4.3.4 Věrohodnost JFA modelu

Předpokládejme, že máme k dispozici soubor R nahrávek mluvčího s a k nim příslušné pozorovatelné proměnné $\underline{\boldsymbol{O}}_s = \{\boldsymbol{O}_{1,s}, \ldots, \boldsymbol{O}_{R,s}\}$, kde $\boldsymbol{O}_{r,s}$ reprezentuje sekvenci příznakových vektorů příslušnou nahrávce r. Skryté proměnné (faktory) pro tato data předpokládejme uspořádány ve formě vektoru $\underline{\boldsymbol{\kappa}}_s$ definovaného jako

$$\underline{\boldsymbol{\kappa}}_s = \begin{bmatrix} \boldsymbol{w}_{1,s} \\ \vdots \\ \boldsymbol{w}_{R,s} \\ \boldsymbol{y}_s \\ \boldsymbol{z}_s \end{bmatrix} \tag{4.52}$$

Pokud by byly hodnoty skrytých proměnných $\underline{\boldsymbol{\kappa}}_s$ známé, bylo by možné určit pro každou nahrávku supervektor $\boldsymbol{m}_{r,s}$ podle (4.45). Následně jednoduše vypočítat věrohodnost $\boldsymbol{O}_{r,s}$ a tím pádem i celkovou věrohodnost $\underline{\boldsymbol{O}}_s$. Hodnoty skrytých proměnných však nejsou známé a pro vyhodnocení věrohodnosti $\underline{\boldsymbol{O}}_s$ je tak nutné provést marginalizaci podmíněného rozložení $P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda})$ přes skryté proměnné $\underline{\boldsymbol{\kappa}}_s$. Tuto marginalizaci provedeme vyhodnocením integrálu

$$\begin{aligned} P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda}) = \int \Big(\int P(\boldsymbol{O}_{1,s}|\boldsymbol{y}_s, \boldsymbol{z}_s, \boldsymbol{w}_{1,s}, \boldsymbol{\Lambda}) \mathcal{N}(\boldsymbol{w}_{1,s}|\boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{w}_{1,s} \ldots \\ \int P(\boldsymbol{O}_{R,s}|\boldsymbol{y}_s, \boldsymbol{z}_s, \boldsymbol{w}_{R,s}, \boldsymbol{\Lambda}) \mathcal{N}(\boldsymbol{w}_{R,s}|\boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{w}_{R,s} \Big) \mathcal{N}(\boldsymbol{y}_s|\boldsymbol{0}, \boldsymbol{I}) \mathcal{N}(\boldsymbol{z}_s|\boldsymbol{0}, \boldsymbol{I}) d\boldsymbol{y}_s d\boldsymbol{z}_s. \end{aligned} \tag{4.53}$$

Což lze na základě (4.52) souhrnně vyjádřit jako

$$P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda}) = \int P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda}) \mathcal{N}(\underline{\boldsymbol{\kappa}}_s|\underline{\boldsymbol{0}}, \underline{\boldsymbol{I}}) d\underline{\boldsymbol{\kappa}}_s. \tag{4.54}$$

Nyní předpokládejme, že máme pro každou nahrávku r postačující statistiky $N_{r,s,c}$, $\tilde{\boldsymbol{F}}_{r,s,c}(\boldsymbol{\mu}_c)$ a $\tilde{\boldsymbol{S}}_{r,s,c}(\boldsymbol{\mu}_c)$ vyhodnocené podle (4.23a) a (4.27) na základě UBM modelu ($c = 1, \ldots, C$, kde C je počet komponent tohoto modelu). Statistiky $\tilde{\boldsymbol{S}}_{r,s,c}(\boldsymbol{\mu}_c)$ jsou přitom uvažovány pouze diagonální[17]. Definujme na základě těchto statistik následující proměnné v prostoru supervektorů. Nechť $\boldsymbol{N}_{r,s}$ je diagonální matice $CF \times CF$ jejíž diagonála je tvořena bloky $N_{r,s,c}\boldsymbol{I}$, kde $\boldsymbol{I}$ je jednotková matice rozměru $F \times F$. Dále nechť $\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{\mu})$ je CF-dimenzionální vektor vytvořený zřetězením vektorů $\tilde{\boldsymbol{F}}_{r,s,c}(\boldsymbol{\mu}_c)$ a konečně $\tilde{\boldsymbol{S}}_{r,s}(\boldsymbol{\mu})$ je diagonální matice rozměru $CF \times CF$ jejíž diagonála je tvořena bloky $\tilde{\boldsymbol{S}}_{r,s,c}(\boldsymbol{\mu}_c)$.

[17] $\tilde{\boldsymbol{S}}_c(\boldsymbol{\mu}_c) = \operatorname{diag}\left(\sum_{t=1}^{T} \gamma_{t,c}(\boldsymbol{o}_t - \boldsymbol{\mu}_c)(\boldsymbol{o}_t - \boldsymbol{\mu}_c)^T\right)$

Pro soubor nahrávek dále definujme obdobným způsobem proměnné $\underline{\boldsymbol{N}}_s, \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}), \underline{\tilde{\boldsymbol{S}}}_s(\boldsymbol{\mu})$. Nechť $\underline{\boldsymbol{N}}_s$ je $RCF \times RCF$ diagonální matice jejíž diagonála je tvořena bloky $\boldsymbol{N}_{r,s}$. Nechť $\underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu})$ je RCF-dimenzionální vektor vytvořený zřetězením vektorů $\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{\mu})$ a $\underline{\tilde{\boldsymbol{S}}}_s(\boldsymbol{\mu})$ je diagonální matice rozměru $RCF \times RCF$ jejíž diagonála je tvořena bloky $\tilde{\boldsymbol{S}}_{r,s}(\boldsymbol{\mu})$. Obdobně je definována $RCF \times RCF$ diagonální matice $\underline{\boldsymbol{\Sigma}}$ na základě $\boldsymbol{\Sigma}$. Dále definujme matici $\underline{\boldsymbol{W}}$ rozměru $RCF \times (RD_{chan} + D_{spk} + CF)$ jako

$$\underline{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{U} & & & \boldsymbol{V} & \boldsymbol{D} \\ & \ddots & & \vdots & \vdots \\ & & \boldsymbol{U} & \boldsymbol{V} & \boldsymbol{D} \end{pmatrix}. \tag{4.55}$$

Podmíněné rozložení pozorovatelných dat při daných skrytých proměnných

Logaritmus podmíněného rozložení $P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda})$ lze na základě definovaných proměnných vyjádřit jako (Kenny, 2005)

$$\begin{aligned} \log P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda}) =& \sum_{r=1}^{R} \left(\sum_{c=1}^{C} N_{r,s,c} \log \frac{1}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{S}}_{r,s}(\boldsymbol{\mu}) \right) \right) \\ &+ \underline{\boldsymbol{\kappa}}_s^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}) - \frac{1}{2} \underline{\boldsymbol{\kappa}}_s^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{N}}_s \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{W}} \underline{\boldsymbol{\kappa}}_s \\ =& G_s(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \underline{\boldsymbol{\kappa}}_s^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}) - \frac{1}{2} \underline{\boldsymbol{\kappa}}_s^T \underline{\boldsymbol{W}}^T \underline{\boldsymbol{N}}_s \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{W}} \underline{\boldsymbol{\kappa}}_s, \end{aligned} \tag{4.56}$$

kde bylo pro větší přehlednost dalšího zápisu zavedeno

$$G_s(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{r=1}^{R} \left(\sum_{c=1}^{C} N_{r,s,c} \log \frac{1}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{S}}_{r,s}(\boldsymbol{\mu}) \right) \right). \tag{4.57}$$

Pro vyjádření $\log P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda})$ podle (4.56) je nutné předpokládat, že každý příznakový vektor byl přiřazen při výpočtu postačujících statistik $N_{r,s,c}$, $\boldsymbol{F}_{r,s,c}$ a $\boldsymbol{S}_{r,s,c}$ vždy právě k jedné komponentě UBM modelu, tedy ve smyslu Viterbiho algoritmu. Správně by však měl být proveden součet přes všechna možná přiřazení příznakových vektorů ke komponentám UBM modelu. Výpočet věrohodnosti modelu založeného na faktorové analýze by byl ale v takovém případě nezvládnutelný (Kenny et al., 2007a). V praxi se však často namísto statistik vyplývajících z aplikace Viterbiho algoritmu používají statistiky odpovídající Baum-Welchovu algoritmu, to jsou statistiky definované dle (4.23).

Marginální rozložení pozorovatelných dat

Logaritmus celkové věrohodnosti lze po dosazení (4.56) do (4.54) vyjádřit v uzavřené formě jako (Kenny, 2005; Kenny et al., 2007b)

$$\log P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda}) = G_s(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \frac{1}{2} \log |\underline{\boldsymbol{L}}_s| + \frac{1}{2} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu})^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{W}} \underline{\boldsymbol{L}}_s^{-1} \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}) \tag{4.58}$$

$$= G_s(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \frac{1}{2} \log |\underline{\boldsymbol{L}}_s| + \frac{1}{2} \left\| \underline{\boldsymbol{L}}_s^{-1/2} \underline{\boldsymbol{W}} \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}) \right\|^2, \tag{4.59}$$

kde $\|.\|$ je eukleidovská norma vektoru[18], $\underline{\boldsymbol{L}}_s$ je symetrická pozitivně definitní matice rozměru $(RD_{chan} + D_{spk} + CF) \times (RD_{chan} + D_{spk} + CF)$ definovaná jako

$$\underline{\boldsymbol{L}}_s = \underline{\boldsymbol{I}} + \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\boldsymbol{N}}_s \underline{\boldsymbol{W}} \tag{4.60}$$

a $\underline{\boldsymbol{L}}_s^{-1/2}$ je inverze $\underline{\boldsymbol{L}}_s^{1/2}$, dolní trojúhelníkové matice získané Choleskyho dekompozicí $\underline{\boldsymbol{L}}_s$. Platí tedy $\underline{\boldsymbol{L}}_s = \underline{\boldsymbol{L}}_s^{1/2}(\underline{\boldsymbol{L}}_s^{1/2})^T$. Vyjádření (4.59) je praktické z implementačního hlediska jednak protože není nutné počítat přímo $\underline{\boldsymbol{L}}_s^{-1}$ a dále lze na základě $\underline{\boldsymbol{L}}_s^{1/2}$ jednoduše stanovit hodnotu determinantu [19] $|\underline{\boldsymbol{L}}_s|$.

Posteriorní rozložení skrytých proměnných při daných pozorovatelných datech

Posteriorní rozložení skrytých proměnných pro daná pozorovatelná data $P(\underline{\boldsymbol{\kappa}}_s|\underline{\boldsymbol{O}}_s, \boldsymbol{\Lambda})$ odpovídá (Kenny, 2005) normálnímu rozložení se střední hodnotou

$$\mathbb{E}[\underline{\boldsymbol{\kappa}}_s] = \underline{\boldsymbol{L}}_s^{-1} \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}) \tag{4.61}$$

a kovarianční maticí

$$\mathrm{cov}(\underline{\boldsymbol{\kappa}}_s, \underline{\boldsymbol{\kappa}}_s) = \underline{\boldsymbol{L}}_s^{-1}, \tag{4.62}$$

kde $\underline{\boldsymbol{L}}_s$ odpovídá (4.60). Platí tedy $P(\underline{\boldsymbol{\kappa}}_s|\underline{\boldsymbol{O}}_s, \boldsymbol{\Lambda}) = \mathcal{N}(\underline{\boldsymbol{\kappa}}_s|\underline{\boldsymbol{L}}_s^{-1} \underline{\boldsymbol{W}}^T \underline{\boldsymbol{\Sigma}}^{-1} \underline{\tilde{\boldsymbol{F}}}_s(\boldsymbol{\mu}), \underline{\boldsymbol{L}}_s^{-1})$.

Zatímco skryté proměnné jsou vzhledem k apriorním rozložením předpokládány jako nezávislé, pro posteriorní rozložení již tento předpoklad neplatí. Matice $\underline{\boldsymbol{L}}_s$ je řídká, ale matice $\underline{\boldsymbol{L}}_s^{-1}$ již není (Kenny, 2005). Rozměr matice $\underline{\boldsymbol{L}}_s$ roste s počtem nahrávek R a v případě velkého počtu nahrávek značně narůstá výpočetní náročnost inverze. Protože jsou však faktory $\boldsymbol{y}_s$ a $\boldsymbol{z}_s$ uvažovány jako společné pro všechny nahrávky, není možné provést nezávislé vyhodnocení posteriorních rozložení skrytých proměnných pro jednotlivé nahrávky odděleně.

4.3.5 Význam složek JFA modelu

Na první pohled nemusí být zřejmý důvod uvažování dvou druhů faktorů specifických pro variabilitu mluvčích, důležitý je ale rozdíl ve struktuře jim příslušných zátěžových matic. Zaměřme se tedy nyní na význam jednotlivých členů v modelu (4.45).

Uvažujme nejprve model, pro který je pouze $\boldsymbol{D} \neq \boldsymbol{0}$, zatímco $\boldsymbol{V} = \boldsymbol{0}$ a $\boldsymbol{U} = \boldsymbol{0}$. Jak uvadí (Kenny et al., 2008b), model (4.45) za této situace odpovídá klasickému MAP odhadu parametrů (Gauvain – Lee, 1994). Výhodou klasického MAP odhadu je, že s rostoucím množstvím dat dostupných pro adaptaci modelu se odhad asymptoticky blíží ML odhadu. Kovarianční matice supervektorů je však v tomto případě diagonální a nereflektuje tak korelaci mezi jednotlivými komponentami GMM. V případě omezeného množství dat tak dochází pouze k adaptaci v těch dimenzích supervektoru, které odpovídají komponentám, kterým byla přiřazena nějaká data.

[18] $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^T\boldsymbol{x}$

[19] Platí $|\boldsymbol{AB}| = |\boldsymbol{A}||\boldsymbol{B}|$ a tedy $|\underline{\boldsymbol{L}}_s| = |\underline{\boldsymbol{L}}_s^{1/2}||(\underline{\boldsymbol{L}}_s^{1/2})^T|$, přitom determinant $\underline{\boldsymbol{L}}_s^{1/2}$ se stanoví jednoduše jako součin prvků na diagonále.

Ve speciálním případě, kdy $\boldsymbol{D}^2 = \frac{1}{\tau}\boldsymbol{\Sigma}$, kde τ je relevanční faktor, odpovídá příslušný model přímo relevančnímu MAP odhadu. Platí totiž

$$\begin{aligned} \boldsymbol{L}_s &= \boldsymbol{I} + \boldsymbol{D}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{D} \\ &= \boldsymbol{I} + \frac{1}{\tau}\boldsymbol{N}_s, \end{aligned} \tag{4.63}$$

$$\begin{aligned} \mathbb{E}[\boldsymbol{s}_s] = \boldsymbol{\mu} + \boldsymbol{D}\mathbb{E}[\boldsymbol{z}_s] &= \boldsymbol{\mu} + \boldsymbol{D}\boldsymbol{L}_s^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}_s(\boldsymbol{\mu}) \\ &= (\tau\boldsymbol{I} + \boldsymbol{N}_s)^{-1}(\tau\boldsymbol{\mu} + \boldsymbol{F}_s), \end{aligned} \tag{4.64}$$

což, vyjádřeno pro jednotlivé komponenty GMM, přesně odpovídá (4.32).

Na druhé straně lze předpokládat, že většina variability hlasů mluvčích je obsažena pouze v nízkodimenzionálním podprostoru supervektorového prostoru. Tento „prostor mluvčích“ je definován maticí $\boldsymbol{V}$, která představuje zátěžovou matici faktorů mluvčích. Výhodou modelu, pro který je $\boldsymbol{V} \neq \boldsymbol{0}$, je plná kovarianční matice rozložení $\boldsymbol{s}_s$ v prostoru supervektorů[20]. Je tím pádem vyjádřena vzájemná korelace složek supervektoru i napříč komponentami GMM a i v případě malého množství dat dostupného pro zapsání mluvčího jsou tak adaptovány všechny složky. Přitom je nutné stanovit hodnoty pouze pro malý počet volných parametrů představovaných faktory mluvčích. Supervektor specifický pro každého mluvčího, který má být zapsán do systému, je pak ale vždy (nezávisle na množství dostupných trénovacích dat) hledán pouze ve vymezeném prostoru mluvčích a to bez ohledu na to, zda se v tomto prostoru skutečně nachází či nikoliv. Efekt $\boldsymbol{V}$ a $\boldsymbol{D}$ je tak do značné míry komplementární a je výhodné zahrnout oba členy $\boldsymbol{V}\boldsymbol{y}_s$ a $\boldsymbol{D}\boldsymbol{z}_s$ do modelu supervektoru $\boldsymbol{s}_s$.

Význam zahrnutí členu $\boldsymbol{U}\boldsymbol{w}$ v (4.45) souvisí se schopností přizpůsobení modelu mluvčího akustickým podmínkám testované nahrávky a také schopností stanovení modelu mluvčího nezávislého na akustických podmínkách trénovacích dat. Předpoklad, že variabilita akustických podmínek je obsažena v nízkodimenzionálním podprostoru supervektorového prostoru, je detailně zdůvodněn například v (Kenny et al., 2007a). Z uvedených argumentů zmiňme jeden podstatný – pokud by byla kovarianční matice odpovídající variabilitě akustických podmínek plného řádu (analogicky k $\boldsymbol{D}$), znamenalo by to, že hlas jednoho mluvčího může znít jako hlas libovolného jiného mluvčího příslušnou změnou akustických podmínek a to by bylo z pohledu úlohy rozpoznávání mluvčích velmi nepříznivé.

Všechny faktory zahrnuté v JFA modelu tedy mají svůj význam a jak dokládá například (Burget et al., 2009), úplný model JFA poskytuje vyšší úspěšnost rozpoznávání mluvčích v porovnání se zjednodušenými variantami. V praxi se však přesto při implementaci JFA modelu často provádí několik zjednodušení s cílem omezení výpočetní náročnosti a to jak v souvislosti s odhadem hyperparametrů, tak s vyhodnocením věrohodnosti při rozpoznávání. Jedná se především o *oddělený odhad* hyperparametrů $\boldsymbol{V}$, $\boldsymbol{D}$ a $\boldsymbol{U}$. Vzájemná korelace skrytých proměnných v posteriorních rozloženích je totiž podstatným zdrojem nárůstu výpočetní náročnosti v případě *společného odhadu* všech hyperparametrů. Oddělený odhad $\boldsymbol{V}$ a $\boldsymbol{D}$ má však vedle snížení výpočetní a implementační

[20] $\boldsymbol{V}\boldsymbol{V}^T$

náročnosti i další důvod. Experimenty (Kenny et al., 2008b) prokázaly, že v případě společného odhadu je příspěvek $\boldsymbol{D}$ k celkové věrohodnosti trénovacích dat oproti příspěvku $\boldsymbol{V}$ zanedbatelný[21]. Většina variability mezi mluvčími je reprezentována členem $\boldsymbol{Vy}$ a v důsledku velmi malého vlivu ztrácí člen $\boldsymbol{Dz}$ na významu. Autoři (Kenny et al., 2008b) přisuzují příčinu tohoto problému charakteru ML odhadu. Počet volných parametrů $\boldsymbol{V}$ ($D_{spk}CF$ parametrů) totiž výrazně převyšuje počet volných parametrů $\boldsymbol{D}$ (CF parametrů). Kenny et al. (2008b) proto navrhují řešení v podobě rozdělení trénovacích dat do dvou sad, větší část trénovací databáze je použita pro odhad $\boldsymbol{V}$ a zbylá data jsou pak použita pro oddělený odhad $\boldsymbol{D}$.

4.3.6 Odhad hyperparametrů metodou maximální věrohodnosti

Je možné ukázat (Kenny, 2005), že celková věrohodnost modelu $\boldsymbol{\Lambda}$ pro trénovací data $P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda})$ bude vyšší oproti věrohodnosti původního modelu $P(\underline{\boldsymbol{O}}_s|\hat{\boldsymbol{\Lambda}})$, bude-li v rámci M kroku iterace EM algoritmu maximalizována hodnota pomocné funkce

$$Q(\boldsymbol{\Lambda}|\hat{\boldsymbol{\Lambda}}) = \sum_s \mathbb{E}_{\underline{\boldsymbol{\kappa}}_s|\underline{\boldsymbol{O}}_s,\hat{\boldsymbol{\Lambda}}}[\log P(\underline{\boldsymbol{O}}_s|\underline{\boldsymbol{\kappa}}_s, \boldsymbol{\Lambda})]. \tag{4.65}$$

Položením derivací této pomocné funkce podle jednotlivých hyperparametrů rovných nule získáme vztahy pro výpočet nových hodnot hyperparametrů.

Uvažujme pro jednoduchost odhad hyperparametrů pro model pouze s $\boldsymbol{V} \neq \boldsymbol{0}$, jehož skryté parametry tedy představují pouze faktory $\boldsymbol{y}_s$. Později bude ukázáno, jak může být popsaný postup uplatněn i k odhadu ostatních hyperparametrů.

Nejprve uveďme ML odhad hyperparametrů $\boldsymbol{V}$ a $\boldsymbol{\Sigma}$. Pro inicializaci hyperparametrů $\boldsymbol{\mu}$ je vhodnou volbou supervektor vytvořený zřetězením vektorů středních hodnot komponent UBM modelu $\boldsymbol{\mu}_c$. Obdobně kovarianční matice komponent UBM modelu $\boldsymbol{\Sigma}_c$ jsou použity pro inicializaci $\boldsymbol{\Sigma}$ (resp. jejích diagonálních bloků). Zátěžová matice $\boldsymbol{V}$ je inicializována náhodně.

Jak již bylo uvedeno, pro stanovení hyperparametrů JFA modelu je vyžadována trénovací databáze, ve které je od každého mluvčího k dispozici několik nahrávek. V rámci hledání podprostoru specifického pro variabilitu mluvčích je možné dosáhnout zjednodušení výpočtů jednoduchou úvahou, že sečtením postačujících statistik stanovených pro jednotlivé nahrávky mluvčího dojde (za předpokladu, že jsou data vhodně vyvážená) k zprůměrování a tím pádem vyrušení efektu různých akustických podmínek odpovídajících jednotlivým nahrávkám. Vstupem algoritmu jsou tedy statistiky $\boldsymbol{N}_s = \sum_r \boldsymbol{N}_{r,s}$, $\boldsymbol{F}_s = \sum_r \boldsymbol{F}_{r,s}$ a $\boldsymbol{S}_s = \sum_r \boldsymbol{S}_{r,s}$, kde r je bráno přes všechny nahrávky mluvčího s. Dle (Kenny et al., 2005) dochází pouze k nevýraznému poklesu úspěšnosti rozpoznávání mluvčích při použití tohoto přístupu v trénování hyperparametrů ve srovnání s odhadem pracujícím s nahrávkami mluvčích samostatně. Bez ohledu na použitý přístup však platí, že hodnost odhadované zátěžové matice $\boldsymbol{V}$ nesmí překročit počet mluvčích v trénovací databázi.

[21] Vliv $\boldsymbol{D}$ a $\boldsymbol{V}$ lze relativně porovnat na základě očekávaných hodnot $\|\boldsymbol{Vy}\|^2$ a $\|\boldsymbol{Dz}\|^2$. Protože apriorní rozložení $\boldsymbol{y}$ i $\boldsymbol{z}$ je standardní normální, je možné určit uvedené očekávané hodnoty jako $\mathrm{tr}(\boldsymbol{VV}^T)$ a $\mathrm{tr}(\boldsymbol{D}^2)$.

Před uvedením vztahů pro výpočet nového ML odhadu hyperparametrů je třeba definovat několik statistik akumulovaných v průběhu E kroku EM algoritmu přes všechny mluvčí s, jsou to:

$$\mathfrak{A}_c = \sum_s N_{s,c}\mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T] \tag{4.66a}$$

$$\mathfrak{B} = \sum_s \tilde{\boldsymbol{F}}_s(\hat{\boldsymbol{\mu}})\mathbb{E}[\boldsymbol{y}_s^T] \tag{4.66b}$$

$$\boldsymbol{N} = \sum_s \boldsymbol{N}_s \tag{4.66c}$$

kde očekávané hodnoty skrytých proměnných $\mathbb{E}[.]$ jsou určeny na základě jejich posteriorního rozložení. To je stanoveno s využitím původního odhadu hyperparametrů modelu $\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{\Sigma}}\}$ analogicky k (4.61) a (4.62). Platí tedy $\mathbb{E}[\boldsymbol{y}_s] = \boldsymbol{L}_s^{-1}\hat{\boldsymbol{V}}^T\hat{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{F}}_s(\hat{\boldsymbol{\mu}})$ a $\mathrm{cov}(\boldsymbol{y}_s, \boldsymbol{y}_s) = \boldsymbol{L}_s^{-1}$, kde $\boldsymbol{L}_s = \boldsymbol{I} + \hat{\boldsymbol{V}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{N}_s\hat{\boldsymbol{V}}$. Pro $\mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T]$ platí

$$\mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T] = \mathbb{E}[\boldsymbol{y}_s]\mathbb{E}[\boldsymbol{y}_s^T] + \mathrm{cov}(\boldsymbol{y}_s, \boldsymbol{y}_s). \tag{4.67}$$

Rozdělíme-li zátěžovou matici $\boldsymbol{V}$ na bloky $\boldsymbol{V}_c$ ($c = 1, \ldots, C$) rozměru $F \times D_{spk}$, takže

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_C \end{pmatrix}, \tag{4.68}$$

a obdobně rozdělíme $\mathfrak{B}$ na bloky $\mathfrak{B}_c$, je odhad nových hodnot pro blok $\boldsymbol{V}_c$ stanoven na základě řešení rovnice

$$\boldsymbol{V}_c\mathfrak{A}_c = \mathfrak{B}_c. \tag{4.69}$$

ML odhad $\boldsymbol{\Sigma}$ je stanoven jako

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\left(\sum_s \tilde{\boldsymbol{S}}_s(\hat{\boldsymbol{\mu}}) - \mathrm{diag}(\mathfrak{B}\boldsymbol{V}^T)\right). \tag{4.70}$$

ML odhad $\boldsymbol{\mu}$ lze provést dvěma způsoby (Kenny, 2005). První způsob využívá výše uvedeného odhadu s tím, že zátěžová matice je rozšířena o supervektor $\boldsymbol{\mu}$ a vektor skrytých proměnných je rozšířen o jeden konstantní člen, tedy

$$\acute{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{V} & \boldsymbol{\mu} \end{pmatrix}, \tag{4.71}$$

$$\acute{\boldsymbol{y}}_s = \begin{bmatrix} \boldsymbol{y}_s \\ 1 \end{bmatrix}. \tag{4.72}$$

Je přitom zřejmé, že platí $\acute{\boldsymbol{V}}\acute{\boldsymbol{y}}_s = \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{\mu}$.

Druhý způsob ML odhadu $\boldsymbol{\mu}$ spočívá v přímé optimalizaci pomocné funkce podle $\boldsymbol{\mu}$ za předpokladu, že ostatní hyperparametry jsou pevné. Počáteční odhad hyperparametrů je v rámci E kroku použit k výpočtu statistiky $\mathfrak{C}$ definované jako

$$\mathfrak{C} = \sum_s \boldsymbol{N}_s\hat{\boldsymbol{V}}\mathbb{E}[\boldsymbol{y}_s] \tag{4.73}$$

a ML odhad nových hodnot koeficientů $\boldsymbol{\mu}$ je stanoven jako

$$\boldsymbol{\mu} = \boldsymbol{N}^{-1}\left(\sum_s \tilde{\boldsymbol{F}}_s(\mathbf{0}) - \mathfrak{C}\right). \tag{4.74}$$

Pro získání odhadu všech hyperparametrů lze pak střídavě provádět odhad $\boldsymbol{V}$ a $\boldsymbol{\Sigma}$ v jedné iteraci a odhad $\boldsymbol{\mu}$ v následující.

Jak již bylo zmíněno, nevýhodou ML odhadu je pomalá konvergence. Jak uvádí (Kenny, 2005), řešení nalezené ML algoritmem však není ideální i z dalšího hlediska. V určité fázi EM algoritmu by totiž měla pro posteriorní rozložení skrytých proměnných platit podmínka[22]

$$\frac{1}{S}\sum_s \mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T] = \boldsymbol{I}, \tag{4.75}$$

kde S je počet mluvčích v trénovací databázi. Zkušenost autora (Kenny, 2005) však ukazuje, že bez ohledu na počet provedených iterací ML algoritmu tato podmínka není nikdy splněna a diagonální členy výrazu na levé straně (4.75) jsou vždy výrazně menší než 1. Řešení tohoto problému představuje odhad parametrů metodou minimální divergence popsaný v následující kapitole.

4.3.7 Odhad hyperparametrů metodou minimální divergence

Opět uvažujme model, pro který je pouze $\boldsymbol{V} \neq \mathbf{0}$. V případě odhadu založeného na metodě minimální divergence (MD) je předpokládáno, že orientace prostoru mluvčích je známá[23] (stanovená například ML odhadem) a cílem je minimalizovat divergenci posteriorního a požadovaného apriorního rozložení skrytých proměnných. Je možné ukazát, že snížením Kullback-Leiblerovy divergence těchto rozložení, s ohledem na omezení dané požadavkem zachování orientace prostoru mluvčích, dojde ke zvýšení celkové věrohodnosti pro trénovací data $P(\underline{\boldsymbol{O}}_s|\boldsymbol{\Lambda})$.

Před uvedením vztahů pro MD odhad hyperparametrů opět definujme několik (normalizovaných) akumulačních statistik, které jsou vypočtené v rámci E kroku, těmi jsou

$$\boldsymbol{\mu}_{\boldsymbol{y}} = \frac{1}{S}\sum_s \mathbb{E}[\boldsymbol{y}_s] \tag{4.76a}$$

$$\boldsymbol{\mathfrak{D}}_{\boldsymbol{yy}} = \frac{1}{S}\sum_s \mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T] \tag{4.76b}$$

$$\boldsymbol{N} = \sum_s \boldsymbol{N}_s \tag{4.76c}$$

Nový MD odhad hyperparametrů je pak definován na základě vztahů

$$\boldsymbol{\mu} = \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{V}}\boldsymbol{\mu}_{\boldsymbol{y}} \tag{4.77}$$

$$\boldsymbol{V} = \hat{\boldsymbol{V}}\boldsymbol{K}_{\boldsymbol{yy}}^{1/2} \tag{4.78}$$

[22] Tato podmínka vyplývá z apriorního rozložení skrytých proměnných, které je předpokládáno jako standardní normální rozložení, tedy s nulovou střední hodnotou a jednotkovou kovarianční maticí.

[23] Prostor je vymezen rozsahem $\hat{\boldsymbol{V}}\hat{\boldsymbol{V}}^T$ posunutým o $\boldsymbol{\mu}$. Zdůrazněme, že v případě ML odhadu není činěn žádný takový předpoklad.

kde $\boldsymbol{K}_{yy}^{1/2}$ je dolní trojúhelníková matice, pro kterou platí[24]

$$\boldsymbol{K}_{yy}^{1/2}(\boldsymbol{K}_{yy}^{1/2})^T = \boldsymbol{\mathfrak{D}}_{yy} - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^T. \tag{4.79}$$

S ohledem na zvýšení celkové věrohodnosti trénovacích dat je v rámci MD algoritmu proveden nový odhad $\boldsymbol{\Sigma}$ následovně[25]:

$$\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\sum_s \left(\tilde{\boldsymbol{S}}_s(\hat{\boldsymbol{\mu}}) - 2\,\mathrm{diag}\left(\tilde{\boldsymbol{F}}_s(\hat{\boldsymbol{\mu}})\mathbb{E}[\boldsymbol{y}_s^T]\hat{\boldsymbol{V}}^T\right) + \mathrm{diag}\left(\hat{\boldsymbol{V}}\mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T]\hat{\boldsymbol{V}}^T\boldsymbol{N}_s\right)\right). \tag{4.80}$$

Veškerá odvození uvedených vztahů lze nalézt v (Kenny, 2005).

4.3.8 Praktický návrh JFA systému

V rámci praktické implementace systémů je přirozenou snahou zajištění co nejnižší výpočetní náročnosti, pokud možno s co nejmenším dopadem na úspěšnost rozpoznávání. Jeden ze způsobů snížení výpočetní náročnosti představuje vypuštění některé ze složek JFA modelu (typickou volbou je především člen $\boldsymbol{Dz}$). Pokud chceme zachovat úplný model, protože vliv jednotlivých složek se mění v souvislosti s podmínkami (např. množstvím dat dostupných pro stanovení modelu mluvčího, viz 4.3.5), které nejsou předem známé a navíc je nelze považovat za neměnné, lze dosáhnout snížení výpočetní a implementační náročnosti odhadu hyperparametrů, trénování modelů mluvčích a vyhodnocení předložených verifikačních soudů níže popsanými způsoby.

Odhad hyperparametrů

V rámci odhadu hyperparametrů JFA modelu jsou stanoveny matice $\boldsymbol{V}$, $\boldsymbol{U}$ a $\boldsymbol{D}$ odděleně. Motivace pro nezávislý odhad parametrů $\boldsymbol{V}$ a $\boldsymbol{D}$ spočívající ve snaze potlačení dominantního vlivu $\boldsymbol{V}$ na úkor $\boldsymbol{D}$ již byla uvedena (viz podkapitola 4.3.5). Poznamenejme, že prostory definované maticemi $\boldsymbol{V}$ a $\boldsymbol{U}$ jsou nízkodimenzionálními podprostory ve vysokodimenzionálním prostoru, lze je tak celkem oprávněně považovat za nezávislé, protínající se pouze v počátku původního prostoru (Kenny et al., 2007a). Níže popsaný postup odděleného odhadu hyperparametrů JFA modelu je jednou z možných variant. V literatuře je možné nalézt odlišná schémata, která se liší například pořadím odhadu jednotlivých zátěžových matic. Zde uvedené pořadí bylo úspěšně použito autory (Burget et al., 2008), zatímco například autoři (Kenny et al., 2008a) popisují opačné pořadí odhadu $\boldsymbol{U}$ a $\boldsymbol{D}$. Uvedený postup včetně příslušných výpočtů shrnuje příloha A.

Nejprve je proveden odhad zátěžové matice $\boldsymbol{V}$. Pro tento okamžik je uvažován model GMM supervektorů $\boldsymbol{m}_{r,s}$ pouze ve tvaru $\boldsymbol{m}_{r,s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s$. Pro každého mluvčího s z trénovací databáze jsou sečteny postačující statistiky stanovené pro jemu příslušné nahrávky. Dostáváme tedy $\boldsymbol{N}_s = \sum_r \boldsymbol{N}_{r,s}$, $\boldsymbol{F}_s = \sum_r \boldsymbol{F}_{r,s}$ a $\boldsymbol{S}_s = \sum_r \boldsymbol{S}_{r,s}$, kde r je bráno přes všechny nahrávky mluvčího s. Statistiky

[24] Tzn. získaná Choleskyho dekompozicí.

[25] Odhad $\boldsymbol{\Sigma}$ není nijak ovlivněn požadavkem zachování orientace prostoru mluvčích definovaného původním modelem.

$\boldsymbol{F}_s$ a $\boldsymbol{S}_s$ jsou dále vycentrovány kolem supervektoru $\boldsymbol{\mu}$ a dostáváme tak $\tilde{\boldsymbol{F}}_s(\boldsymbol{\mu})$ a $\tilde{\boldsymbol{S}}_s(\boldsymbol{\mu})$. Hodnoty parametrů $\boldsymbol{V}$ je pak možné stanovit algoritmy popsanými v podkapitolách 4.3.6 a 4.3.7. Způsob popsaný v tab. A.1 přílohy A odpovídá kombinovanému ML-MD odhadu.

Následně je proveden odhad matice $\boldsymbol{U}$. U trénovacích nahrávek je nejprve pro každého mluvčího s na základě posteriorního rozložení faktorů $\boldsymbol{y}_s$ stanoven odhad $\mathbb{E}[\boldsymbol{y}_s]$. Pro tento účel je provedeno sečtení statistik vyhodnocených pro jednotlivé nahrávky daného mluvčího. Následně je možné provést vyjádření odhadu supervektoru specifického pro mluvčího s ve tvaru $\mathbb{E}[\boldsymbol{s}_s] = \boldsymbol{\mu} + \boldsymbol{V}\mathbb{E}[\boldsymbol{y}_s]$. Modelu supervektorů $\boldsymbol{m}_{r,s}$ je nyní uvažován ve tvaru $\boldsymbol{m}_{r,s} = \mathbb{E}[\boldsymbol{s}_s] + \boldsymbol{U}\boldsymbol{w}_{r,s}$. Je patrné, že tento model je analogický k výše uvažovanému modelu v případě odhadu matice $\boldsymbol{V}$. Tím pádem je možné použít shodné nástroje pro odhad obou typů hyperparametrů. V rámci odhadu $\boldsymbol{U}$ jsou však statistiky vyhodnocené pro jednotlivé nahrávky uvažovány samostatně a jsou vycentrované kolem supervektoru $\mathbb{E}[\boldsymbol{s}_s]$. Pracujeme tedy se statistikami $\tilde{\boldsymbol{F}}_{r,s}(\mathbb{E}[\boldsymbol{s}_s])$ a $\tilde{\boldsymbol{S}}_{r,s}(\mathbb{E}[\boldsymbol{s}_s])$, kde r je opět bráno přes všechny nahrávky mluvčího s.

Nakonec je stanovena diagonální matice $\boldsymbol{D}$. Pro nahrávky z trénovací databáze je nejprve určeno posteriorní rozložení faktorů $\boldsymbol{y}_s$ a $\boldsymbol{w}_{r,s}$ následujícím způsobem. Pro každého mluvčího s je na základě statistik sečtených přes všechny jeho nahrávky stanoven odhad supervektoru $\mathbb{E}[\boldsymbol{s}_s]$. Dále je zvlášť pro každou nahrávku provedeno vyhodnocení posteriorního rozložení $\boldsymbol{w}_{r,s}$ na základě postačujících statistik vycentrovaných kolem $\mathbb{E}[\boldsymbol{s}_s]$. Pro každou nahrávku je tak možné vyjádřit odhad supervektoru $\mathbb{E}[\boldsymbol{h}_{r,s}] = \mathbb{E}[\boldsymbol{s}_s] + \boldsymbol{U}\mathbb{E}[\boldsymbol{w}_{r,s}]$. Model supervektorů $\boldsymbol{m}_{r,s}$ pak lze vyjádřit v podobě $\boldsymbol{m}_{r,s} = \mathbb{E}[\boldsymbol{h}_{r,s}] + \boldsymbol{D}\boldsymbol{z}_s$. Postačující statistiky mluvčího s vyhodnocené pro jednotlivé nahrávky jsou pak vycentrovány kolem $\mathbb{E}[\boldsymbol{h}_{r,s}]$ a sečteny dohromady, dostáváme tedy $\boldsymbol{N}_s = \sum_r \boldsymbol{N}_{r,s}$, $\tilde{\boldsymbol{F}}_s(\mathbb{E}[\boldsymbol{.}, \boldsymbol{h}_s]) = \sum_r \tilde{\boldsymbol{F}}_{r,s}(\mathbb{E}[\boldsymbol{h}_{r,s}])$ a $\tilde{\boldsymbol{S}}_s(\mathbb{E}[\boldsymbol{.}, \boldsymbol{h}_s]) = \sum_r \tilde{\boldsymbol{S}}_{r,s}(\mathbb{E}[\boldsymbol{h}_{r,s}])$. Hodnoty parametrů $\boldsymbol{D}$ jsou pak stanoveny opět s využitím kombinovaného ML-MD odhadu. Postup se v principu nijak neliší od odhadu podprostorů definovaných $\boldsymbol{V}$ a $\boldsymbol{U}$ s tím rozdílem, že výpočty využívají faktu, že matice $\boldsymbol{D}$ je diagonální (viz tab. A.3 přílohy A).

Upozorněme, že pokud v rámci odhadu hyperparametrů nedochází k reestimaci matice $\boldsymbol{\Sigma}$ (tzn. že je ponechán počáteční odhad, typicky vycházející z parametrů UBM modelu), nejsou nutné postačující statistiky druhého řádu $\boldsymbol{S}$.

Trénování modelů mluvčích

Uvažujme modely mluvčích reprezentované parametry (očekávanou hodnotou a kovarianční maticí) posteriorního rozložení supervektoru $\boldsymbol{s}_s$. Toto rozložení je stanoveno na základě předložených trénovacích dat mluvčího s.

Nejprve jsou sečteny postačující statistiky vyhodnocené pro všechny trénovací nahrávky mluvčího dohromady. Na základě statistik $\boldsymbol{N}_s$, $\tilde{\boldsymbol{F}}_s(\boldsymbol{\mu})$ a $\tilde{\boldsymbol{S}}_s(\boldsymbol{\mu})$ je současně stanoveno posteriorní rozložení faktorů $\boldsymbol{w}$ a $\boldsymbol{y}_s$. Přestože by bylo možné argumentovat, že sečtením statistik přes několik nahrávek již došlo k vyrušení vlivu akustických podmínek a není proto nutné provádět odhad faktorů $\boldsymbol{w}$, je nutné brát v úvahu také častou situaci, kdy je k dispozici pouze jediná trénovací nahrávka. Faktory

jsou při společném odhadu uspořádány ve formě vektoru $\boldsymbol{\kappa}_s = [\boldsymbol{w}^T, \boldsymbol{y}_s^T]^T$ a zátěžové matice ve formě matice $\boldsymbol{W} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \end{pmatrix}$. Parametry posteriorního rozložení jsou stanoveny vhodným dosazením do (4.61) a (4.62). Upozorněme, že v tomto případě je možné použít naprosto shodný výpočet, který by byl použit pro stanovení samotných faktorů $\boldsymbol{y}_s$.

Následně je použit bodový odhad $\boldsymbol{w}$ a $\boldsymbol{y}_s$ pro vycentrování statistik $\boldsymbol{N}_s$, $\boldsymbol{F}_s$ a $\boldsymbol{S}_s$ kolem bodu $\boldsymbol{\mu}+\boldsymbol{V}\mathbb{E}[\boldsymbol{y}_s]+\boldsymbol{U}\mathbb{E}[\boldsymbol{w}]$ a výsledné statistiky použity pro stanovení posteriorního rozložení $\boldsymbol{d}_s$. Faktory $\boldsymbol{w}$ stanovené pro trénovací data nemají další využití, protože (ideálně) nenesou žádnou informaci související s identitou mluvčího.

Parametry posteriorního rozložení supervektoru $\boldsymbol{s}_s$ jsou pak stanoveny následovně:

$$\mathbb{E}[\boldsymbol{s}_s] = \boldsymbol{\mu} + \boldsymbol{V}\mathbb{E}[\boldsymbol{y}_s] + \boldsymbol{D}\mathbb{E}[\boldsymbol{z}_s] \tag{4.81}$$

$$\operatorname{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s) \cong \operatorname{diag}(\boldsymbol{V}\operatorname{cov}(\boldsymbol{y}_s, \boldsymbol{y}_s)\boldsymbol{V}^T) + \boldsymbol{D}\operatorname{cov}(\boldsymbol{z}_s, \boldsymbol{z}_s)\boldsymbol{D} \tag{4.82}$$

Vyjádření $\operatorname{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$ podle (4.82) není přesné, protože nezahrnuje korelaci mezi $\boldsymbol{y}_s$ a $\boldsymbol{z}_s$. Kenny et al. (2007b) ale uvádí, že tato aproximace poskytuje dobré výsledky. Model mluvčího je tedy reprezentován prostřednictvím $\mathbb{E}[\boldsymbol{s}_s]$ a $\operatorname{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$.

Vyhodnocení věrohodnosti

Věrohodnost JFA modelu je definována vztahem (4.54). Budeme-li uvažovat faktory specifické pro mluvčího jako dané a jeho model reprezentovaný supervektorem $\boldsymbol{s}_s$, zůstávají skrytými proměnnými pouze faktory specifické pro variabilitu akustických podmínek $\boldsymbol{w}$. Přes ty je tedy nutné provést marginalizaci, můžeme zapsat

$$\log P(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{\Lambda}) = \log \int P(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{w}, \boldsymbol{\Lambda})\mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \boldsymbol{I})d\boldsymbol{w}. \tag{4.83}$$

Na základě (4.59) pak dostáváme vyjádření výpočtu věrohodnosti ve tvaru (Kenny et al., 2007b)

$$\begin{aligned} \log P(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{\Lambda}) = &\sum_{c=1}^{C} N_c \log \frac{1}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} - \frac{1}{2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)) \\ &- \frac{1}{2}\log|\boldsymbol{L}| + \frac{1}{2}\left\|\boldsymbol{L}^{-1/2}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)\right\|^2, \end{aligned} \tag{4.84}$$

kde $\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{U}$. Postačující statistiky jsou tedy centrovány kolem bodu $\boldsymbol{s}_s$. V praxi však tento supervektor není přesně znám a je k dispozici pouze jeho odhad $\mathbb{E}[\boldsymbol{s}_s]$ stanovený na základě předložených trénovacích dat mluvčího. Zejména v situacích, kdy je k dispozici malé množství dat, lze očekávat, že bodový odhad není dobrou reprezentací skutečného supervektoru $\boldsymbol{s}_s$ příslušného mluvčímu s. Tuto nejistotu lze promítnout do výpočtu (4.84) nahrazením postačujících statistik $\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)$ a $\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)$ za jejich očekávané hodnoty $\mathbb{E}[\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)]$ a $\mathbb{E}[\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)]$, které jsou stanoveny takto

$$\mathbb{E}[\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)] = \boldsymbol{F} - \boldsymbol{N}\mathbb{E}[\boldsymbol{s}_s] \tag{4.85a}$$

$$\mathbb{E}[\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)] = \boldsymbol{S} - 2\operatorname{diag}(\boldsymbol{F}\mathbb{E}[\boldsymbol{s}_s^T]) + \operatorname{diag}\left(\boldsymbol{N}\left(\mathbb{E}[\boldsymbol{s}_s]\mathbb{E}[\boldsymbol{s}_s^T] + \operatorname{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)\right)\right), \tag{4.85b}$$

kde $\mathbb{E}[\boldsymbol{s}_s]$ a $\mathrm{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$ jsou stanoveny dle (4.81) a (4.82).

Vyhodnocení (4.84) závisí na modelu daného mluvčího s pouze prostřednictvím statistik $\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)$ a $\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)$, resp. $\mathbb{E}[\tilde{\boldsymbol{F}}(\boldsymbol{s}_s)]$ a $\mathbb{E}[\tilde{\boldsymbol{S}}(\boldsymbol{s}_s)]$, jejichž výpočet není náročný, protože odhad $\boldsymbol{s}_s$, resp. $\mathbb{E}[\boldsymbol{s}_s]$ a $\mathrm{cov}(\boldsymbol{s}_s, \boldsymbol{s}_s)$, je proveden v rámci trénování modelu mluvčího. V rámci vyhodnocení (4.84) je tak naročný především výpočet $\boldsymbol{L}^{-1/2}$ (hodnotu determinantu $|\boldsymbol{L}|$ lze na základě znalosti $\boldsymbol{L}^{-1/2}$ určit snadno), který je však nezávislý na mluvčím. To je významné při vyhodnocení věrohodnosti pro více testovaných mluvčích, typicky např. při aplikaci T-norm normalizace. Dále poznamenejme, že hodnost matice $\boldsymbol{V}$ ovlivňuje výpočetní náročnost odhadu $\boldsymbol{s}_s$ (tedy proces trénování modelu mluvčích), nikoliv však vyhodnocení věrohodnosti podle (4.84) (Kenny et al., 2007b).

V případě většího množství testovaných modelů (např. při aplikaci zmíněné T-norm normalizace) lze dosáhnout snížení výpočetní náročnosti vyhodnocení věrohodnosti použitím shodného bodového odhadu faktorů $\boldsymbol{w}$ (stanoveného na základě univerzálního modelu) pro všechny testované modely mluvčích. Na základě tohoto přístupu lze aplikací Taylorova rozvoje prvního řádu odvodit velmi efektivní aproximaci logaritmu poměru věrohodností pro testovaný model (Glembek et al., 2009).

4.4 I-vektory

Autoři Dehak et al. (2011) navrhují aplikaci jednoduchého modelu faktorové analýzy pro redukci rozměru CF-dimenzionálních GMM supervektorů. Pro supervektor $\boldsymbol{m}_{r,s}$ odpovídající nahrávce r mluvčího s je v tomto případě předpokládán model

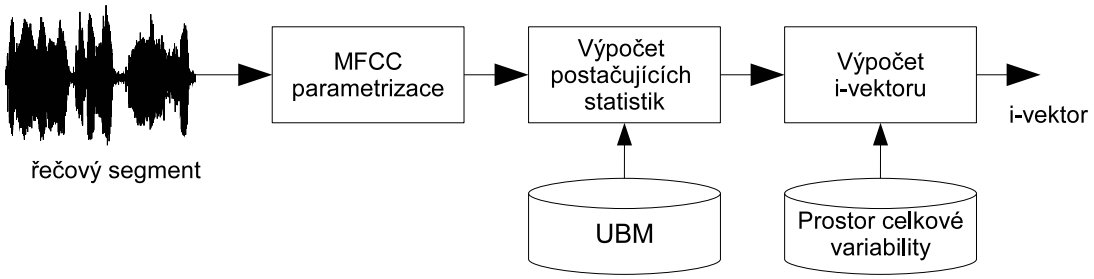
$$\boldsymbol{m}_{r,s} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{x}_{r,s}, \tag{4.86}$$

kde $\boldsymbol{\mu}$ je supervektor, který je nezávislý na mluvčím i nahrávce, $\boldsymbol{T}$ je matice rozměru $CF \times D_{ivec}$ a $\boldsymbol{x}_{r,s}$ je D_{ivec}-dimenzionální[26] náhodný vektor s rozložením $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Vektor $\boldsymbol{x}_{r,s}$ je označován jako *i-vektor*[27]. Model (4.86) nerozlišuje variabilitu specifickou pro hlasy mluvčích a akustické podmínky. Matice $\boldsymbol{T}$ definuje prostor celkové variability (total variability space) a prvky vektoru $\boldsymbol{x}_{r,s}$ představují faktory odpovídající této variabilitě. Hyperparametry $\boldsymbol{T}$ jsou stanoveny shodně jako hyperparametry $\boldsymbol{U}$ v případě JFA modelu (viz podkapitola 4.3.8) s tím rozdílem, že postačující statistiky nejsou centrovány kolem MAP odhadu supervektoru specifického pro mluvčího $\boldsymbol{s}_s$ ale jsou centrovány kolem $\boldsymbol{\mu}$.

Rozložení GMM supervektorů $\boldsymbol{m}_{r,s}$ je na základě modelu (4.86) předpokládáno jako normální se střední hodnotou $\boldsymbol{\mu}$ a kovarianční maticí $\boldsymbol{T}\boldsymbol{T}^T$. I-vektor představuje D_{ivec}-dimenzionální reprezentaci nahrávky ve formě MAP odhadu $\boldsymbol{x}_{r,s}$ stanoveného pro tuto nahrávku (viz podkapitola 4.3.4).

[26] $D_{ivec} \ll CF$ a je typicky v řádu stovek

[27] Autor této práce se setkal s několika interpretacemi termínu i-vektor. Dehak et al. (2011) uvádí termín i-vektor jako zkratku označení „identitní vektor“. Brummer et al. (2010) vysvětluje původ označení ve středním (intermediate) rozměru těchto vektorů, který je vyšší než rozměr akustických příznakových vektorů a nižší než rozměr supervektorů. Lze se však setkat i s dalšími výklady.



Obrázek 4.1: *Proces extrakce i-vektorů lze uvažovat jako součást parametrizace řečového signálu*

Proces výpočtu i-vektorů lze pak interpretovat jako součást parametrizace řečového signálu, jak ilustruje obr. 4.1.

4.4.1 Klasifikace pomocí i-vektorů

Otázkou je nyní způsob klasifikace pomocí i-vektorů. Dehak et al. (2011) provádí mimo jiné klasifikaci založenou na využití SVM klasifikátoru (o SVM a dalších diskriminativních klasifikátorech pojednává kapitola 5), pro který volí na základě předchozí zkušenosti (Dehak et al., 2009) jako jádrovou funkci kosinovou vzdálenost. Jádrová funkce pro dvojici i-vektorů $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ má tak podobu

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \, \|\boldsymbol{x}_2\|}, \tag{4.87}$$

kde $\langle ., . \rangle$[28] představuje skalární součin. Lepších výsledků však bylo v (Dehak et al., 2011) dosaženo aplikací kosinové vzdálenosti přímo pro výpočet skóre hodnotícího hypotézu, že mluvčí v nahrávce reprezentované i-vektorem $\boldsymbol{x}_1$ je shodný s mluvčím v nahrávce reprezentované i-vektorem $\boldsymbol{x}_2$, tj.

$$s_{CDS}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \, \|\boldsymbol{x}_2\|}. \tag{4.88}$$

Pro potvrzení hypotézy je toto skóre (po případném provedení kalibrace) porovnáno s rozhodovacím prahem. Výhodou tohoto přístupu je nižší výpočetní náročnost a s tím související rychlost vyhodnocení.

Další možností je aplikace pravděpodobnostní lineární diskriminační analýzy (PLDA), která bude popsána v podkapitole 4.5.

4.4.2 Kompenzace variability akustických podmínek

Zatímco v případě JFA modelu je variabilita akustických podmínek explicitně modelována, v případě modelu (4.86) tomu tak není. Dehak et al. (2011) se zabývá několika způsoby potlačení vlivu variability akustických podmínek aplikovanými namísto prostoru supervektorů až v prostoru i-vektorů. Konkrétně se jedná o metodu normalizace kovariance uvnitř tříd (WCCN), metodu projekce rušivých atributů (NAP) a lineární diskriminační analýzu (LDA). Metody WCCN a NAP byly navrženy v souvislosti s aplikací SVM klasifikátorů a budou popsány v podkapitole 5.3. Tyto

[28] $\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \boldsymbol{x}_1^T \boldsymbol{x}_2$

metody zachovávají na rozdíl od LDA rozměr vstupních vektorů. Nejlepší úspěšnosti rozpoznávání mluvčích bylo v případě (Dehak et al., 2011) dosaženo použitím kombinace metody WCCN a LDA.

4.4.3 Lineární diskriminační analýza

Lineární diskriminační analýza[29] (LDA) usiluje o nalezení podprostoru, ve kterém by byly třídy (v našem případě mluvčí) co nejlépe rozlišitelné. Intuitivně lze odvodit požadavek zajištění velkého rozptylu mezi třídami a malého rozptylu uvnitř tříd v nově definovaném podprostoru. Hledána je projekce vektorů $\boldsymbol{x}$ z původního n-dimenzionálního prostoru do prostoru nižší dimenze m v podobě lineární transformace $\boldsymbol{x}' = \boldsymbol{A}^T\boldsymbol{x}$, kde $\boldsymbol{A}$ je matice rozměru $n \times m$.

Transformační matice je stanovena na základě trénovacích dat následujícím způsobem (Bishop, 2006). Předpokládejme, že máme k dispozici data pro S tříd (mluvčích), na základě kterých je určena celková kovariance uvnitř tříd jako

$$\boldsymbol{\Sigma}_W = \sum_s \boldsymbol{\Sigma}_s, \tag{4.89}$$

kde $\boldsymbol{\Sigma}_s$ je kovarianční matice třídy s, tedy

$$\boldsymbol{\Sigma}_s = \sum_{r=1}^{R_s} (\boldsymbol{x}_r - \boldsymbol{\mu}_s)(\boldsymbol{x}_r - \boldsymbol{\mu}_s)^T \tag{4.90}$$

a

$$\boldsymbol{\mu}_s = \frac{1}{R_s} \sum_{r=1}^{R_s} \boldsymbol{x}_r. \tag{4.91}$$

R_s je počet pozorování dostupných pro třídu s (počet dostupných nahrávek od daného mluvčího). Kovariance mezi třídami $\boldsymbol{\Sigma}_B$ je pak definována jako[30]

$$\boldsymbol{\Sigma}_B = \sum_s R_s(\boldsymbol{\mu}_s - \boldsymbol{\mu}_T)(\boldsymbol{\mu}_s - \boldsymbol{\mu}_T)^T, \tag{4.92}$$

kde

$$\boldsymbol{\mu}_T = \frac{1}{R} \sum_{r=1}^{R} \boldsymbol{x}_r = \frac{1}{R} \sum_s R_s \boldsymbol{\mu}_s \tag{4.93}$$

a $R = \sum_{s=1}^{S} R_s$ je celkový počet všech dostupných pozorování. Upozorněme, že v případě i-vektorů je $\boldsymbol{\mu}_T = \boldsymbol{0}$, protože, jak bylo uvedeno v souvislosti s modelem (4.86), je předpokládáno standardní normální rozložení i-vektorů $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Uvedené kovarianční matice jsou definovány v původním n-dimenzionálním prostoru $\boldsymbol{x}$, obdobně je možné definovat matice $\boldsymbol{\Omega}_W$ a $\boldsymbol{\Omega}_B$ v m-dimenzionálním prostoru $\boldsymbol{x}'$. Optimalizační kritérium je s ohledem na uvedené požadavky definováno jako

$$J(\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{\Omega}_W^{-1}\boldsymbol{\Omega}_B), \tag{4.94}$$

[29] Někdy označovaná také jako Fisherův lineární diskriminant.

[30] Kovarianci mezi třídami lze odvodit na základě dekompozice celkové kovariance $\boldsymbol{\Sigma}_T = \sum_{r=1}^{R}(\boldsymbol{x}_r - \boldsymbol{\mu}_T)(\boldsymbol{x}_r - \boldsymbol{\mu}_T)^T$, platí totiž $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B$.

což je možné přepsat jako funkci transformační matice $\boldsymbol{A}$ v následující podobě

$$J(\boldsymbol{A}) = \mathrm{tr}\left((\boldsymbol{A}^T\boldsymbol{\Sigma}_W\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{\Sigma}_B\boldsymbol{A})\right). \tag{4.95}$$

Maximalizace tohoto kritéria bude dosažena (Bishop, 2006), pokud budou sloupce transformační matice $\boldsymbol{A}$ tvořeny vlastními vektory příslušnými m největším vlastním číslům matice $\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\Sigma}_B$.

Kovarianční matice $\boldsymbol{\Sigma}_B$ je definována jako součet S matic, z nichž je každá výsledkem vnějšího součinu[31] dvou vektorů a tím pádem mají hodnost 1. V důsledku toho platí, že maximálně $(S-1)$ z těchto matic je nezávislých. Hodnost matice $\boldsymbol{\Sigma}_B$ je rovna nejvýše $(S-1)$ a tím pádem je maximálně $(S-1)$ vlastních čísel matice nenulových. Jinými slovy, pro stanovení m rozměrného podprostoru je tak třeba, aby trénovací databáze obsahovala nejméně $(m+1)$ tříd.

4.5 Pravděpodobnostní lineární diskriminační analýza

Pravděpodobnostní lineární diskriminační analýza (PLDA) představuje metodu založenou na faktorové analýze, navrženou původně pro úlohu rozpoznávání tváří (Prince – Elder, 2007). V úloze rozpoznávání mluvčích je v současné době typicky aplikována pro klasifikaci v prostoru i-vektorů. PLDA definuje generativní model i-vektoru $\boldsymbol{x}_{r,s}$ reprezentujícího r-tou nahrávku mluvčího s jako

$$\boldsymbol{x}_{r,s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{w}_{r,s} + \boldsymbol{\epsilon}_{r,s}. \tag{4.96}$$

Tento model je v souladu s dekompozicí (4.44) uvažován složený ze dvou složek. První část $\boldsymbol{s}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s$ charakterizuje mluvčího a je předpokládána jako neměnná pro všechny jeho nahrávky. Druhá část $\boldsymbol{c}_{r,s} = \boldsymbol{U}\boldsymbol{w}_{r,s} + \boldsymbol{\epsilon}_{r,s}$ charakterizuje akustické podmínky, které jsou pro každou nahrávku specifické. Význam zátěžových matic $\boldsymbol{V}$ a $\boldsymbol{U}$ je analogický k JFA modelu (4.45). Sloupce matice $\boldsymbol{V}$ definují báze prostoru charakteristického pro variabilitu hlasů mluvčích a sloupce matice $\boldsymbol{U}$ pak báze prostoru charakteristického pro variabilitu akustických podmínek. Pro skryté proměnné (faktory) je opět předpokládáno standardní normální apriorní rozložení, tj. $P(\boldsymbol{y}_s) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ a $P(\boldsymbol{w}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Vektor $\boldsymbol{\epsilon}_{r,s}$ pak reprezentuje reziduální variabilitu nezachycenou ostatními členy, pro kterou je předpokládáno normální rozložení $P(\boldsymbol{\epsilon}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, kde $\boldsymbol{\Sigma}$ je $D_{ivec} \times D_{ivec}$ diagonální kovarianční matice. Platí tedy $P(\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V}\boldsymbol{V}^T)$, $P(\boldsymbol{c}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{\Sigma})$ a $P(\boldsymbol{m}_{r,s}|\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{s}_s, \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{\Sigma})$. PLDA model je tedy souhrnně reprezentován čtveřicí parametrů $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$.

4.5.1 Trénování hyperparametrů PLDA modelu

Postup odhadu hyperparametrů PLDA modelu je analogický k odhadu hyperparametrů JFA modelu. Stejně jako v případě trénování hyperparametrů JFA modelu je vyžadováno, aby bylo pro jednotlivé mluvčí z trénovací databáze k dispozici několik nahrávek. Na rozdíl od JFA modelu ale například neplatí tvrzení, že rozměr podprostorů definovaných maticemi $\boldsymbol{V}$ a $\boldsymbol{U}$ je výrazně nižší

[31] $\boldsymbol{x} \otimes \boldsymbol{x} = \boldsymbol{x}\boldsymbol{x}^T$

ve srovnání s rozměrem původního prostoru, tedy prostoru i-vektorů. V důsledku toho je poněkud sporné tvrzení o nezávislosti podprostorů $\boldsymbol{V}$ a $\boldsymbol{U}$, a tím pádem předpoklad, že se protínají pouze v počátku původního prostoru. Proto je preferován společný odhad hyperparametrů.

Předpokládejme, že máme k dispozici R nahrávek od mluvčího s, reprezentovaných i-vektory $\boldsymbol{x}_{r,s}$, a PLDA model $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$. Pro ML i MD variantu EM algoritmu, který provádí odhad hyperparametrů PLDA modelu, je nejprve nutné stanovit posteriorní rozložení skrytých proměnných[32] $\boldsymbol{y}_s$ a $\boldsymbol{w}_{r,s}$, viz podkapitoly 4.3.6 a 4.3.7.

Model předpokládající, že všechny nahrávky pocházejí od shodného řečníka (a tedy sdílí faktory $\boldsymbol{y}_s$), lze formulovat na základě kombinace generativních modelů (4.96) pro jednotlivé nahrávky následovně:

$$\begin{bmatrix} \boldsymbol{x}_{1,s} \\ \vdots \\ \boldsymbol{x}_{R,s} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu} \\ \vdots \\ \boldsymbol{\mu} \end{bmatrix} + \begin{pmatrix} \boldsymbol{U} & & \boldsymbol{V} \\ & \ddots & \vdots \\ & \boldsymbol{U} & \boldsymbol{V} \end{pmatrix} \begin{bmatrix} \boldsymbol{w}_{1,s} \\ \vdots \\ \boldsymbol{w}_{R,s} \\ \boldsymbol{y}_s \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_{1,s} \\ \vdots \\ \boldsymbol{\epsilon}_{R,s} \end{bmatrix}. \tag{4.97}$$

Použijeme-li souhrnné vyjádření skrytých proměnných po vzoru (4.52) a hyperparametrů podle (4.55), lze model vyjádřit ve tvaru

$$\underline{\boldsymbol{x}}_s = \underline{\boldsymbol{\mu}} + \underline{\boldsymbol{W}}\boldsymbol{\kappa}_s + \underline{\boldsymbol{\epsilon}}_s. \tag{4.98}$$

Položme dále

$$\tilde{\boldsymbol{x}}_{r,s} = \boldsymbol{x}_{r,s} - \boldsymbol{\mu} \tag{4.99}$$

a

$$\underline{\tilde{\boldsymbol{x}}}_s = \underline{\boldsymbol{x}}_s - \underline{\boldsymbol{\mu}} \tag{4.100}$$

Požadované parametry posteriorního rozložení $\underline{\boldsymbol{\kappa}}_s$ jsou definovány v souladu s (4.61) a (4.62) následovně:

$$\mathbb{E}[\underline{\boldsymbol{\kappa}}_s] = \underline{\boldsymbol{L}}_s^{-1}\underline{\boldsymbol{W}}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\tilde{\boldsymbol{x}}}_s \tag{4.101}$$

$$\mathrm{cov}(\underline{\boldsymbol{\kappa}}_s, \underline{\boldsymbol{\kappa}}_s) = \underline{\boldsymbol{L}}_s^{-1} \tag{4.102}$$

kde $\underline{\boldsymbol{L}}_s = \underline{\boldsymbol{I}} + \underline{\boldsymbol{W}}^T\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{W}}$.

Podobně jako v případě JFA modelu, je výpočetně náročné především provedení inverze $\underline{\boldsymbol{L}}_s$, protože rozměr $\underline{\boldsymbol{L}}_s$ se mění v závislosti na počtu nahrávek R. Jednou z možností je aplikace blokového Choleskyho algoritmu (Kenny, 2005). Další možností je přímé vyjádření $\underline{\boldsymbol{L}}_s^{-1}$ a $\mathbb{E}[\underline{\boldsymbol{\kappa}}_s]$, které je provedeno v příloze B.

[32] Resp. očekávanou hodnotu $\mathbb{E}[.]$ a kovariační matici $\mathrm{cov}(.,.)$.

Definujme nyní pro každého mluvčího s a jeho nahrávku r následující proměnné:

$$\boldsymbol{\kappa}_{r,s} = \begin{bmatrix} \boldsymbol{w}_{r,s} \\ \boldsymbol{y}_s \end{bmatrix} \qquad \boldsymbol{W} = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{V} \end{pmatrix} \tag{4.103}$$

$$\acute{\boldsymbol{\kappa}}_{r,s} = \begin{bmatrix} \boldsymbol{\kappa}_{r,s} \\ 1 \end{bmatrix} \qquad \acute{\boldsymbol{W}} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{\mu} \end{pmatrix} \tag{4.104}$$

Kombinovaný ML-MD odhad hyperparametrů provedeme následovně. Na základě očekávaných hodnot $\boldsymbol{\kappa}_{r,s}$ jsou nejprve stanoveny následující statistiky a proměnné:

$$N = \sum_s R_s \tag{4.105}$$

$$\mathfrak{A} = \sum_s \sum_r \mathbb{E}[\acute{\boldsymbol{\kappa}}_{r,s} \acute{\boldsymbol{\kappa}}_{r,s}^T] \tag{4.106}$$

$$\mathfrak{B} = \sum_s \sum_r \boldsymbol{x}_{r,s} \mathbb{E}[\acute{\boldsymbol{\kappa}}_{r,s}^T] \tag{4.107}$$

$$\boldsymbol{\mu}_{\boldsymbol{\kappa}} = \frac{1}{N} \sum_s \sum_r \mathbb{E}[\boldsymbol{\kappa}_{r,s}] \tag{4.108}$$

$$\boldsymbol{K}_{\mathcal{X}\mathcal{X}}^{1/2} (\boldsymbol{K}_{\mathcal{X}\mathcal{X}}^{1/2})^T = \frac{1}{N} \sum_s \sum_r \mathbb{E}[\boldsymbol{\kappa}_{r,s} \boldsymbol{\kappa}_{r,s}^T] - \boldsymbol{\mu}_{\boldsymbol{\kappa}} \boldsymbol{\mu}_{\boldsymbol{\kappa}}^T \tag{4.109}$$

kde R_s je počet nahrávek od mluvčího s, S je celkový počet mluvčích v trénovací databázi a

$$\mathbb{E}[\acute{\boldsymbol{\kappa}}_{r,s} \acute{\boldsymbol{\kappa}}_{r,s}^T] = \begin{pmatrix} \mathbb{E}[\boldsymbol{\kappa}_{r,s} \boldsymbol{\kappa}_{r,s}^T] & \mathbb{E}[\boldsymbol{\kappa}_{r,s}] \\ \mathbb{E}[\boldsymbol{\kappa}_{r,s}^T] & 1 \end{pmatrix} \tag{4.110}$$

Pro nový ML odhad pak platí

$$\acute{\boldsymbol{W}}^{ML} \mathfrak{A} = \mathfrak{B} \tag{4.111}$$

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_s \sum_r \operatorname{diag}\left(\boldsymbol{x}_{r,s} \boldsymbol{x}_{r,s}^T \right) - \frac{1}{N} \operatorname{diag}\left(\mathfrak{B} (\acute{\boldsymbol{W}}^{ML})^T \right) \tag{4.112}$$

a na základě následně provedeného MD odhadu (v téže iteraci EM algoritmu) jsou nové hodnoty aktualizovány takto

$$\boldsymbol{W} = \boldsymbol{W}^{ML} \boldsymbol{K}_{\mathcal{X}\mathcal{X}}^{1/2} \tag{4.113}$$

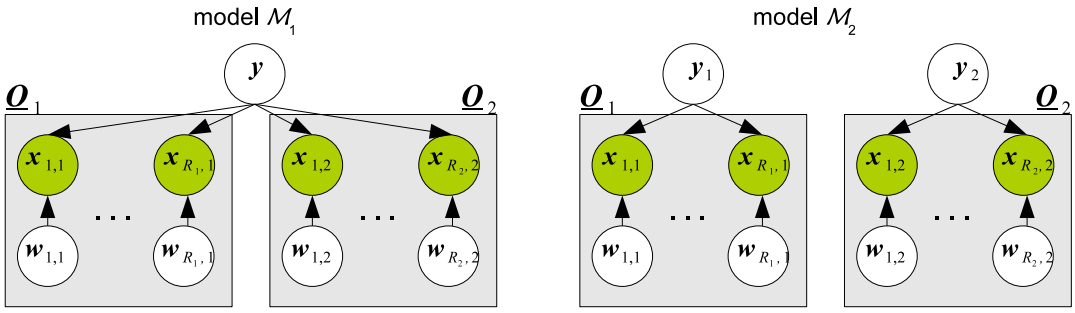
$$\boldsymbol{\mu} = \boldsymbol{\mu}^{ML} + \boldsymbol{W}^{ML} \boldsymbol{\mu}_{\boldsymbol{\kappa}} \tag{4.114}$$

kde $\boldsymbol{W}^{ML}$ a $\boldsymbol{\mu}^{ML}$ odpovídají rozkladu $\acute{\boldsymbol{W}}^{ML}$ podle (4.104). Nové hodnoty parametrů $\boldsymbol{V}$ a $\boldsymbol{U}$ pak odpovídají příslušným hodnotám $\boldsymbol{W}$ podle (4.103).

Autoři Burget et al. (2011) se zabývají návrhem diskriminativního trénování hyperparametrů PLDA modelu a dokládají výrazné zlepšení úspěšnosti rozpoznávání dosažené jeho aplikací.

4.5.2 Rozpoznávání

Proces rozpoznávání je možné formulovat jako Bayesovský výběr modelu (Prince – Elder, 2007). Konkrétně pro úlohu detekce mluvčích předpokládejme, že máme k dispozici dvě sady nahrávek $\underline{\boldsymbol{O}}_1$



Obrázek 4.2: *Modely uvažované při verifikaci mluvčích. Oba modely reprezentují odlišný vztah mezi skrytými proměnnými $\boldsymbol{y}$, které identifikují mluvčího, a pozorovatelnými proměnnými $\boldsymbol{x}$. V případě modelu $\mathcal{M}_1$ je mluvčí obou sad nahrávek totožný, zatímco v případě modelu $\mathcal{M}_2$ nikoliv*

a $\underline{\boldsymbol{O}}_2$, o kterých víme, že nahrávky v rámci každé sady pocházejí od jednoho mluvčího. Naším cílem je rozhodnout, zda je tento mluvčí totožný pro obě sady. Zdůrazněme, že v tomto případě není vůbec rozlišováno mezi trénovacími a testovanými daty, vyhodnocení je tedy symetrické. Uvažujme tedy model $\mathcal{M}_1$ reprezentující situaci, kdy všechny nahrávky pocházejí od téhož mluvčího (a tím pádem všechny sdílejí faktory $\boldsymbol{y}$), a model $\mathcal{M}_2$ reprezentující situaci, kdy mluvčí nahrávek v obou sadách je odlišný (první sadě nahrávek odpovídají faktory $\boldsymbol{y}_1$ a druhé sadě $\boldsymbol{y}_2$). Protože jsou v modelu $\mathcal{M}_2$ skryté proměnné (faktory) příslušné variabilitě mluvčích nezávislé (viz obr. 4.2), lze věrohodnost tohoto modelu rozepsat jako[33]

$$P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_2) = P(\underline{\boldsymbol{O}}_1|\mathcal{M}_2)P(\underline{\boldsymbol{O}}_2|\mathcal{M}_2). \tag{4.115}$$

Výstup detektoru mluvčích pak odpovídá vyhodnocení logaritmu poměru věrohodností těchto modelů, tj.

$$\begin{aligned} s_{PLDA} &= \log \frac{P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_1)}{P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_2)} \\ &= \log P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_1) - \log P(\underline{\boldsymbol{O}}_1|\mathcal{M}_2) - \log P(\underline{\boldsymbol{O}}_2|\mathcal{M}_2). \end{aligned} \tag{4.116}$$

Nyní se zaměřme na vyhodnocení věrohodnosti $P(\underline{\boldsymbol{O}}|\mathcal{M})$. Každá nahrávka r je reprezentována i-vektorem $\boldsymbol{x}_r$, který představuje pozorovatelná data. Předpokládáme-li, že všechny nahrávky v sadě $\underline{\boldsymbol{O}}$ pocházejí od shodného řečníka, platí model (4.97). Věrohodnost $\underline{\boldsymbol{O}}$ je tedy stanovena na základě marginálního rozložení $P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta})$. Protože podmíněné rozložení $\underline{\boldsymbol{x}}$ při daných skrytých proměnných odpovídá normálnímu rozložení $\mathcal{N}(\underline{\boldsymbol{x}}|\underline{\boldsymbol{\mu}} + \underline{\boldsymbol{W}}\boldsymbol{\kappa}, \underline{\boldsymbol{\Sigma}})$ a apriorní rozložení skrytých proměnných je předpokládáno jako standardní normální rozložení, je rozložení $P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta})$ opět normální, konkrétně

$$\mathcal{N}(\underline{\boldsymbol{x}}|\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}). \tag{4.117}$$

Logaritmus věrohodnosti modelu je tedy možné vyjádřit jako

$$\log P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta}) = -\frac{1}{2}RD\log(2\pi) - \frac{1}{2}\log\left|\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}\right| - \frac{1}{2}\tilde{\underline{\boldsymbol{x}}}^T(\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}})^{-1}\tilde{\underline{\boldsymbol{x}}}, \tag{4.118}$$

[33] Pro zjednodušení zápisu je vynechána reference na PLDA model $\boldsymbol{\Theta}$.

kde D je rozměr i-vektorů a $\underline{\tilde{\boldsymbol{x}}}$ je v souladu s (4.100) stanoveno jako $\underline{\tilde{\boldsymbol{x}}} = \underline{\boldsymbol{x}} - \underline{\boldsymbol{\mu}}$. Na základě znalosti struktury matice $\underline{\boldsymbol{W}}$ lze pak provést vyjádření umožňující efektivní výpočet věrohodnosti ve tvaru, který neobsahuje žádné členy, jejichž rozměr by se měnil v závislosti na počtu nahrávek R. Po provedení potřebných úprav dostáváme následující výraz:

$$\begin{aligned}\log P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta}) = &-\frac{1}{2}RD\log(2\pi) - \frac{1}{2}R\log|\boldsymbol{\Upsilon}| - \frac{1}{2}\log|\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}| \\ &-\frac{1}{2}\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r \\ &+\frac{1}{2}\left(\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)\left(\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)^{-1}\left(\sum_{r=1}^{R}\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r\right),\end{aligned} \tag{4.119}$$

kde $\boldsymbol{\Upsilon} = \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{\Sigma}$. Rovnost (4.118) a (4.119) dokládá příloha C.

Kapitola 5

Diskriminativní klasifikátory

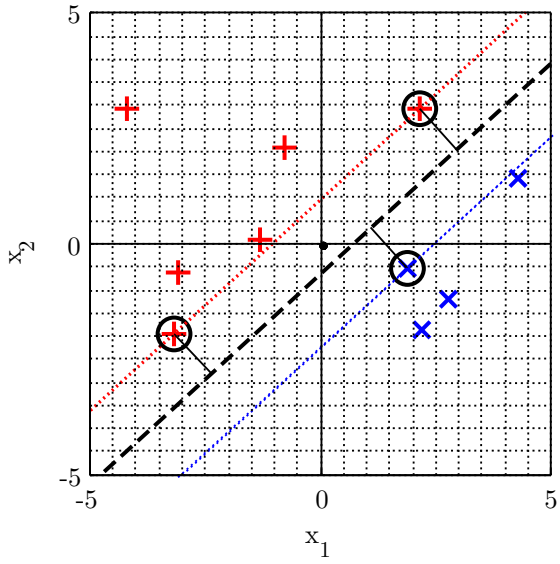
Diskriminativní klasifikátory jsou založeny na přímém učení posteriorních pravděpodobností tříd $P(y|\boldsymbol{x})$, na jejichž základě je provedeno rozhodnutí o příslušnosti vstupních dat k jedné ze tříd, nebo na učení diskriminačních funkcí provádějících přímé zobrazení vstupních dat $\boldsymbol{x}$ na označení tříd. Cílem klasifikátorů je přitom provést právě zobrazení $\boldsymbol{x} \rightarrow y$. Existuje však řada důvodů, pro které je výhodné znát hodnoty posteriorních pravděpodobností $P(y|\boldsymbol{x})$, i když jsou následně použity pro rozhodnutí o příslušnosti k jedné ze tříd (Bishop, 2006). Nicméně hledání sdružené pravděpodobnosti $P(\boldsymbol{x}, y)$, ke kterému dochází v případě generativních klasifikátorů, není pro klasifikaci nezbytné a může představovat zbytečné zvýšení výpočetních nároků nebo klást nepřiměřené požadavky na množství trénovacích dat. Získání spolehlivého odhadu tohoto rozložení může být komplikované zejména pokud mají vstupní data velkou dimenzi. Jak uvádí (Vapnik, 1998):

> Snahou by mělo být hledání přímého řešení problému a nikdy by neměl být řešen obecnější problém jako mezikrok. Je možné, že dostupná data jsou dostatečná pro nalezení přímého řešení, ale jsou nedostatečná pro řešení obecnějšího vnořeného problému.

Hledání sdruženého rozložení $P(\boldsymbol{x}, y)$ představuje právě takovýto vnořený problém. Obecně je diskriminativním klasifikátorům přisuzována lepší zobecňovací schopnost oproti generativním klasifikátorům.

Převážně používaným typem diskriminativních klasifikátorů v úloze rozpoznávání mluvčích jsou SVM (Support Vector Machine) klasifikátory[1]. Obecným úvodem do problematiky SVM se zabývá (Burges, 1998). V úloze rozpoznávání mluvčích však byly aplikovány i další typy diskriminativních klasifikátorů. Katz (2008) se zabývá metodami založenými na nelineární logistické regresi s využitím jádrových funkcí. Chen – Tang (2005) se zabývají použitím RVM (Relevance Vector Machine) klasifikátorů (Bishop, 2006). Ty mají mnoho společných rysů s SVM ale řeší některé z jejich nedostatků. Výhodou ve srovnání s SVM je řidší model (vedoucí k rychlejšímu vyhodnocení testovaných dat) a výstup odpovídající posteriorní pravděpodobnosti oproti binárnímu rozhodnutí o klasifikaci.

[1] V české literatuře se lze setkat s označením metoda podpůrných vektorů, řada autorů však zachovává původní anglické označení.



Obrázek 5.1: *Hraniční nadroviny SVM klasifikátoru pro případ lineárně separovatelných dat, podpůrné vektory jsou znázorněny v kroužku*

Nevýhodou je především dlouhý trénovací proces.

5.1 SVM klasifikátory

Společně s metodami založenými na faktorové analýze představují systémy založené na SVM v současné době hlavní přístupy používané v rámci návrhu moderních systémů pro rozpoznávání mluvčích. Jejich předností ve srovnání s generativními klasifikátory je již uvedená vyšší zobecňovací schopnost. Následující výklad se postupně věnuje návrhu klasifikátoru pro nejsnazší případ lineárně separovatelných dat, následně pro případ dat, která nejsou lineárně separovatelná, ale s ohledem na jejich uspořádání je lineární hranice vhodnou separující nadrovinou pro klasifikaci, a konečně pro případ dat, která nejsou svou povahou lineárně separovatelná. Důležitou součástí SVM klasifikátorů je *jádrová funkce*, která provádí transformaci z původního prostoru příznakových vektorů do prostoru transformovaných příznaků (typicky vyšší dimenze) a umožňuje převést původně lineárně neseparovatelnou úlohu na úlohu lineárně separovatelnou.

5.1.1 Lineárně separovatelná data

SVM klasifikátory představují metodu strojového učení určenou pro binární klasifikaci dat. Cílem SVM je nalezení rozdělující (separující) nadroviny, která představuje rozhodovací hranici mezi dvěma třídami. V úloze detekce mluvčích je jedna třída představována trénovacími vektory zapsaného mluvčího (tuto třídu označíme jako $+1$) a druhá vektory ostatních mluvčích, kteří představují populaci možných neoprávněných mluvčích (třída -1), viz obr. 5.1. V případě, že jsou data lineárně separabilní, je možné umístit v prostoru nadrovinu takovým způsobem, že všechna trénovací data

budou klasifikována správně. Tuto nadrovinu je možné reprezentovat váhovým vektorem $\boldsymbol{u}$ (který představuje normálový vektor nadroviny) a posunutím ρ ($\frac{\rho}{\|\boldsymbol{u}\|}$ je vzdálenost nadroviny od počátku). Testovaný vektor je pak klasifikován na základě hodnoty funkce

$$f(\boldsymbol{x}) = \text{sign}(\langle \boldsymbol{u}, \boldsymbol{x} \rangle + \rho). \tag{5.1}$$

V případě lineárně separabilních dat dokonce existuje obvykle nekonečně mnoho takových nadrovin. Pro nalezení jednoznačného řešení je proto nutné rozšířit požadavky. Trénovací algoritmus SVM proto hledá takovou rozdělující nadrovinu, která zajišťuje maximální šířku pásma mezi třídami. Předpokládejme, že máme k dispozici trénovací data tvořená dvojicemi $(\boldsymbol{x}_i, y_i)$, kde y_i je známá klasifikace vektoru $\boldsymbol{x}_i$ a $i = 1, \ldots, N$. Požadavky pro hledání optimální rozdělující nadroviny SVM klasifikátoru jsou formulovány následovně:

$$\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle + \rho \geq +1 \quad \text{pro} \quad y_i = +1, \tag{5.2}$$

$$\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle + \rho \leq -1 \quad \text{pro} \quad y_i = -1. \tag{5.3}$$

Tyto podmínky lze souhrnně vyjádřit jedinou podmínkou

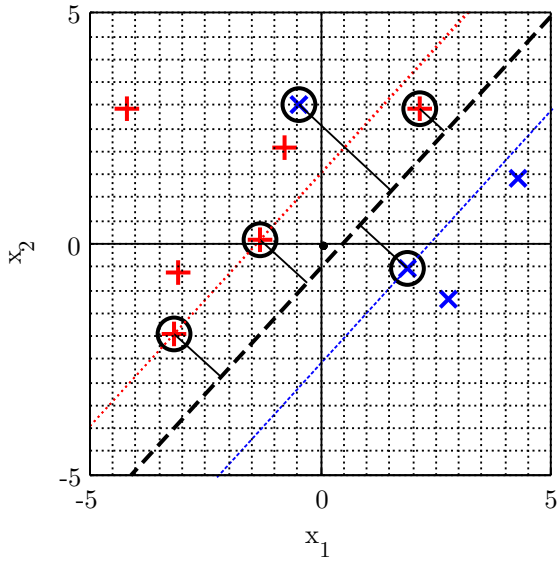
$$y_i(\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle + \rho) \geq 1, \tag{5.4}$$

která musí platit pro všechna data v trénovací databázi. Pro body ležící na separující nadrovině pak platí rovnost $\langle \boldsymbol{u}, \boldsymbol{x} \rangle + \rho = 0$. Všechna trénovací data spadající do třídy $+1$ leží za nadrovinou nebo na nadrovině, pro kterou platí rovnost $\langle \boldsymbol{u}, \boldsymbol{x}_+ \rangle + \rho = +1$. Tato nadrovina se označuje jako hraniční pro třídu $+1$. Analogická podmínka platí pro data ze třídy -1. Vzdálenost obou hraničních nadrovin se pak označuje jako *margin*. Je možné ukázat, že velikost tohoto marginu je rovna $\frac{2}{\|\boldsymbol{u}\|}$. Hraniční nadroviny jsou tak hledány s ohledem na minimalizaci účelové funkce $\boldsymbol{u}^T\boldsymbol{u}$ při zachování podmínky (5.4) (Burges, 1998). Tento optimalizační problém odpovídá *primární formulaci* úlohy *kvadratického programování* (QP). V průběhu optimalizačního procesu jsou nalezeny v trénovací databázi body $\boldsymbol{x}_+$ a $\boldsymbol{x}_-$, které leží na hraničních nadrovinách příslušných tříd. Tyto body se označují jako *podpůrné vektory*. Jejich vyjmutí z trénovací databáze by změnilo nalezené řešení.

5.1.2 Nelineárně separovatelná data

V důsledku různých rušivých vlivů se mohou stát snímaná data příslušná oběma třídám lineárně neseparovatelná i v případě, že lineární hranice jinak představuje vhodnou separující nadrovinu pro klasifikaci. Tuto situaci ilustruje obr. 5.2. Dosud uvedený postup pro nalezení separující nadroviny však funguje pouze pokud jsou všechna trénovací data lineárně separovatelná, protože požadavek maximalizace marginu nepřipouští chybně klasifikovaná data z trénovací množiny. Aby bylo možné provést trénování SVM za popsané situace, je provedeno zmírnění podmínek (5.4) zavedením doplňkových proměnných (tzv. slack variables) ξ_i následujícím způsobem (Burges, 1998):

$$y_i(\langle \boldsymbol{u}, \boldsymbol{x}_i \rangle + \rho) \geq 1 - \xi_i, \tag{5.5}$$



Obrázek 5.2: *Hraniční nadroviny SVM klasifikátoru pro případ lineárně neseparovatelných dat, podpůrné vektory jsou znázorněny v kroužku*

kde $\xi_i \geq 0$, pro $i = 1, \ldots, N$. Zavedením ξ_i je přidána tolerance pro chybnou klasifikaci některých trénovacích vzorků, hodnoty ξ_i představují míru vychýlení vzorku mimo požadovanou oblast. Pro data splňující podmínku (5.4) je $\xi_i = 0$.

Ve snaze zabránit zvyšování marginu na úkor chybné klasifikace dat z trénovací množiny je účelová funkce rozšířena o penalizaci odpovídající chybné klasifikaci trénovacích dat a je definována jako $\boldsymbol{u}^T\boldsymbol{u} + W\sum_i \xi_i$, kde parametr W řídí váhu požadavků dosažení maximální hodnoty marginu a minimální míry chybné klasifikace trénovacích dat. Pro nízké hodnoty W bude preferována velká hodnota marginu za cenu více chyb klasifikace trénovacích dat, naopak pro velké hodnoty W bude za cenu snížení marginu dosaženo nižší míry chybné klasifikace trénovacích dat.

Problém minimalizace $\boldsymbol{u}^T\boldsymbol{u} + W\sum_i \xi_i$ podle $\boldsymbol{u}$ a ρ s ohledem na podmínky (5.5) opět odpovídá primární formě úlohy QP. Použitím metody Lagrangeových multiplikátorů, lze ukázat, že duální formulace QP odpovídající tomuto optimalizačnímu problému spočívá v maximalizaci

$$\sum_i \alpha_i - \frac{1}{2}\sum_i\sum_j \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j \tag{5.6}$$

podle multiplikátorů α_i $(i = 1, \ldots, N)$ při splnění podmínek

$$0 \leq \alpha_i \leq W, \tag{5.7}$$

$$\sum_i \alpha_i y_i = 0. \tag{5.8}$$

V duální formulaci je separující nadrovina definována hodnotami Lagrangeových multiplikátorů α_i. Každý z multiplikátorů α_i souvisí s jedním vzorkem z množiny trénovacích dat a určuje vliv tohoto vzorku na umístění separující nadroviny. Na základě stanovených hodnot α_i je možné vyjádřit

hledané řešení jako

$$\boldsymbol{u} = \sum_i \alpha_i y_i \boldsymbol{x}_i, \tag{5.9}$$

kde i je bráno přes všechna trénovací data, a

$$\rho = y_k - \boldsymbol{u}^T \boldsymbol{x}_k \tag{5.10}$$

pro libovolné k, pro které je $0 < \alpha_k < W$ (příp. je možné namísto náhodně zvoleného k použít střední hodnotu na základě všech ρ stanovených pro data, pro která je splněna uvedená podmínka). Nalezené řešení musí vyhovovat podmínkám označovaným jako Karush-Kuhn-Tuckerovy (KKT) podmínky (Burges, 1998). Na základě těchto podmínek lze ukázat, že pro data ležící na správné straně hraniční nadroviny je $\alpha_i = 0$ (přitom je $\xi_i = 0$), pro data ležící na hraniční nadrovině příslušné třídy je $0 < \alpha_i < W$ (přitom je $\xi_i = 0$) a konečně pro data ležící na nesprávné straně hraniční nadroviny třídy je $\alpha_i = W$ (přitom pokud je $\xi_i \leq 1$ jsou data ještě klasifikována správně a pokud je $\xi_i > 1$ dochází k chybné klasifikaci). Součet (5.9) má význam uvažovat pouze přes data, pro která je $\alpha_i > 0$. Ty pak představují podpůrné vektory. Optimální rozhodovací nadrovinu je možné nalézt jak řešením primární, tak duální formulace standardními metodami QP (Longworth, 2010). Duální formulace je přitom výhodnější. Existuje však několik algoritmů navržených speciálně pro řešení úloh souvisejících s trénováním SVM klasifikátorů. Jedním z nejčastěji odkazovaných algoritmů je SMO (Sequential Minimal Optimization) algoritmus (Platt, 1999).

Klasifikace testovaného vektoru $\boldsymbol{x}$ do třídy $+1$ nebo -1 je pak provedena na základě hodnoty funkce

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{L} \alpha_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + \rho\right), \tag{5.11}$$

kde L je počet podpůrných vektorů modelu.

5.1.3 Nelineární SVM založené na jádrových funkcích

Výše uvedený postup umožňuje nalezení separující nadroviny v případě, kdy trénovací data příslušná oběma třídám nejsou lineárně separovatelná ale s ohledem na uspořádání dat v prostoru představuje lineární klasifikátor vhodnou volbu. Nicméně v řadě praktických úloh nejsou data svou povahou lineárně separovatelná a je nutné, aby hranice mezi třídami měla složitější nelineární podobu. K řešení tohoto problému je obvykle možné použít zobrazení dat nelineární funkcí $\Phi(\boldsymbol{x})$. Z původního prostoru jsou data transformována do nového *obrazového* prostoru (typicky výrazně vyššího rozměru), ve kterém jsou již data lineárně separovatelná. Lineární separující hranice stanovená v tomto prostoru pak odpovídá nelineární hranici v původním prostoru. Explicitní zobrazení do vysokodimenzionálního prostoru však přináší komplikace v podobě zvýšení výpočetní náročnosti trénovacího i klasifikačního procesu.

Důležitou společnou vlastností účelové funkce duální trénovací úlohy (5.6) a klasifikační funkce (5.11) je, že v nich data vystupují vždy ve formě skalárního součinu. Metoda označovaná jako *kernel trik* spočívá v nahrazení těchto skalárních součinů *jádrovou funkcí* $K(.,.)$ takovou, že

platí $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle$, kde $\Phi(\boldsymbol{x}_i)$ představuje explicitní zobrazení do obrazového prostoru. Běžným typem jádrové funkce je například homogenní polynomiální funkce řádu k, která má podobu $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T \boldsymbol{x}_j)^k$. Dalšími standardními jádrovými funkcemi jsou nehomogenní polynomiální funkce, radiální bázová funkce (RBF) nebo sigmodiální funkce (Burges, 1998). Lineární jádrová funkce je představována skalárním součinem v původním prostoru, tj. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$.

Jádrová funkce implicitně provádí skalární součin v obrazovém prostoru, který může být i nekonečné dimenze. Výpočetní náročnost trénování modelu a klasifikace se přitom nijak výrazně neliší ve srovnání s prováděním skalárního součinu v původním prostoru. Protože data nikdy nevystupují pro účely trénování modelů a klasifikace jinak než ve formě skalárního součinu, není nutné provádět explicitní zobrazení do obrazového prostoru a vždy je možné použít jádrovou funkci.

Nahrazením skalárního součinu jádrovou funkcí je možné přepsat klasifikační funkci (5.11) SVM jako

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{L} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \rho\right). \tag{5.12}$$

Obdobně je účelová funkce duální trénovací úlohy (5.6) přepsána jako

$$\sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{5.13}$$

Podmínky (5.7) a (5.8) přitom zůstávají zachovány beze změny.

Aby byla funkce $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ platnou jádrovou funkcí, musí odpovídat nějakému obrazovému prostoru. Zobrazení do tohoto prostoru však nemusí být explicitně vyjádřeno a tento prostor může mít i nekonečnou dimenzi. Platná jádrová funkce musí vyhovovat Mercerově podmínce (Burges, 1998). Nutnou a postačující podmínkou, kterou musí funkce $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ splňovat je, že Gramova matice, jejíž prvky jsou rovny $K(\boldsymbol{x}_m, \boldsymbol{x}_n)$ (kde m je index řádku a n index sloupce), je pozitivně semidefinitní pro všechny možné výběry množiny $\{\boldsymbol{x}_n\}$ (Bishop, 2006).

5.2 Systémy založené na SVM v úloze rozpoznávání mluvčích

První snahy o využití SVM v systémech rozpoznávání mluvčích se soustředily na aplikaci SVM přímo na jednotlivé akustické příznakové vektory a využívaly standardní jádrové funkce, např. polynomiální (Schmidt – Gish, 1996) nebo radiální bázové funkce (Wan – Campbell, 2000). Později byly představeny jádrové funkce navržené přímo s cílem porovnání sekvencí příznakových vektorů libovolné délky (Campbell, 2002; Fine et al., 2001; Smith – Gales, 2002). Rozsáhlý přehled metod pro rozpoznávání mluvčích založených na jádrových funkcích lze nalézt v (Longworth, 2010). Princip několika vybraných SVM systémů je popsán v následujících podkapitolách.

S ohledem na etapy definované v podkapitole 1.2 je poměrně komplikované nalézt společné rysy níže popsaných systémů pro etapu označovanou jako *učení systému*. Parametry systémů jsou závislé na typu použité jádrové funkce a metody pro kompenzaci variability akustických podmínek. Princip ostatních etap je ale pro všechny systémy založené na SVM shodný:

- *Zapsání mluvčích* – spočívá ve stanovení SVM modelu na základě předložených trénovacích dat. Trénovacímu algoritmu jsou předložena data od zapisovaného mluvčího s, kterým je přiřazena pozitivní klasifikace ($y_{i,s} = +1$), a data od mluvčích určených k reprezentaci populace možných neoprávněných mluvčích, kterým je přiřazena negativní klasifikace ($y_{i,s} = -1$). Model mluvčího je reprezentován L_s podpůrnými vektory $\boldsymbol{x}_{i,s}$, jim příslušnými Lagrangeovými multiplikátory $\alpha_{i,s}$ a proměnnou ρ_s. Musí být přitom splněny podmínky $\alpha_{i,s} > 0$ a $\sum_{i=1}^{L_s} \alpha_{i,s} y_{i,s} = 0$. K trénování modelů lze využít volně dostupných nástrojů, autor této práce používal při implementaci SVM systémů program libsvm (Chang – Lin, 2005). V souvislosti s omezeným množstvím trénovacích dat dostupných obvykle pro mluvčího s je poměr negativně a pozitivně klasifikovaných dat dostupných pro trénování zpravidla značně nevyvážený. Možné řešení tohoto problému spočívá ve stanovení rozdílných hodnot penalizací W_+ a W_- pro chybně klasifikovaná data ze třídy $+1$ a -1 (Osuna et al., 1997).

- *Rozpoznávání* – testovaná promluva reprezentovaná vektorem $\boldsymbol{x}$ je klasifikována v závislosti na poloze vůči separující nadrovině odpovídající modelu mluvčího s v prostoru transformovaných příznaků. S ohledem na možnost provedení následné kalibrace je upřednostňován výstup SVM klasifikátoru formou skóre ve tvaru

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \alpha_{i,s} y_{i,s} K(\boldsymbol{x}, \boldsymbol{x}_{i,s}) + \rho_s \tag{5.14}$$

  před binární klasifikací podle (5.12). Kalibrace skóre (5.14) je nezbytná, protože toto skóre nemá žádnou pravděpodobnostní/věrohodnostní interpretaci.

5.2.1 GLDS jádrová funkce

Jednou z nejjednodušších metod založených na SVM klasifikátorech, která umožňuje porovnání sekvencí různé délky, je metoda využívající GLDS (Generalized Linear Discriminant Sequence) jádrovou funkci (Campbell, 2002), ta má podobu

$$K_{GLDS}(\bar{\boldsymbol{b}}_1, \bar{\boldsymbol{b}}_2) = \bar{\boldsymbol{b}}_1^T \boldsymbol{R}^{-1} \bar{\boldsymbol{b}}_2, \tag{5.15}$$

kde $\bar{\boldsymbol{b}}$ je výsledkem zobrazení sekvence příznakových vektorů do SVM obrazového prostoru provedeným jako

$$\bar{\boldsymbol{b}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{b}(\boldsymbol{o}_t), \tag{5.16}$$

kde $\boldsymbol{b}(.)$ představuje expanzi z původního příznakového prostoru do prostoru vyššího řádu. V případě (Campbell, 2002) jsou složky $\boldsymbol{b}(\boldsymbol{o})$ tvořeny všemi monomy[2] až do řádu k, které lze vytvořit

[2]Monom je definován jako součin nezáporných mocnim proměnných, libovolný polynom lze pak uvažovat jako lineární kombinaci monomů. Stupeň monomu odpovídá součtu mocnin jednotlivých proměnných v monomu a stupeň polynomu pak odpovídá nejvyššímu stupni monomů tvořících polynom. Pro příznakový vektor $\boldsymbol{o} = [o_1, o_2]^T$ a maximální stupeň monomů $k = 2$ je vektor $\boldsymbol{b}(\boldsymbol{o}) = [1, o_1, o_2, o_1^2, o_1 o_2, o_2^2]^T$.

ze složek příznakového vektoru $\boldsymbol{o}$. Přesné odvození matice $\boldsymbol{R}$ v (5.15) lze nalézt v (Campbell et al., 2006a). Za běžně platného předpokladu, že počet dat od anti-mluvčích N_{bkg} převyšuje výrazně při trénování modelu mluvčího počet pro něho dostupných dat N_{tar}, je možné stanovit matici $\boldsymbol{R}$ jako

$$\boldsymbol{R} = \frac{1}{N_{bkg}} \sum_{i \in bkg} \bar{\boldsymbol{b}}_i \bar{\boldsymbol{b}}_i^T. \tag{5.17}$$

Pokud provedeme Choleskyho dekompozici $\boldsymbol{R}^{-1} = \boldsymbol{P}^T\boldsymbol{P}$, je možné vyjádřit jádrovou funkci (5.15) jako

$$K_{GLDS}(\bar{\boldsymbol{b}}_1, \bar{\boldsymbol{b}}_2) = (\boldsymbol{P}\bar{\boldsymbol{b}}_1)^T(\boldsymbol{P}\bar{\boldsymbol{b}}_2) = \dot{\boldsymbol{b}}_1^T\dot{\boldsymbol{b}}_2, \tag{5.18}$$

kde $\dot{\boldsymbol{b}}_i = \boldsymbol{P}\bar{\boldsymbol{b}}_i$. Po provedení transformace vektorů $\bar{\boldsymbol{b}}_i$ tak lze provést rychlé vyhodnocení jádrové funkce jako skalárního součinu. Pro snížení výpočetní náročnosti lze uvažovat matici $\boldsymbol{R}^{-1}$ pouze jako diagonální.

Rozhodovací skóre SVM klasifikátoru založeného na GLDS jádrové funkci je pro testovanou nahrávku $\boldsymbol{O}$ reprezentovanou vektorem $\bar{\boldsymbol{b}}$ a cílového mluvčího s definováno jako

$$s_{GLDS-SVM}(\boldsymbol{O}, s) = \sum_{i=1}^{L} \alpha_{i,s} y_{i,s} K_{GLDS}(\bar{\boldsymbol{b}}, \bar{\boldsymbol{b}}_{i,s}) + \rho_s. \tag{5.19}$$

Užitečnou vlastností jádrové funkce (5.15) je, že umožňuje kompaktní reprezentaci SVM modelu (Campbell, 2002). Výpočet rozhodovacího skóre pak nabývá podoby

$$s_{GLDS-SVM}(\boldsymbol{O}, s) = \left(\sum_{i=1}^{L} \alpha_{i,s} y_{i,s} \dot{\boldsymbol{b}}_{i,s} \right)^T \dot{\boldsymbol{b}} + \rho_s = \boldsymbol{u}_s^T \dot{\boldsymbol{b}} + \rho_s, \tag{5.20}$$

kde $\boldsymbol{u}_s$ je vážený součet podpůrných vektorů $\dot{\boldsymbol{b}}_{i,s}$ definujících SVM model. Výhodou tohoto vyjádření je, že pro určení hodnoty výstupního skóre stačí spočítat jediný skalární součin.

5.2.2 Jádrová funkce odvozená od Kullback-Leiblerovy divergence GMM

Běžný způsob vyhodnocení rozdílu mezi dvěma rozloženími spočívá ve stanovení Kullback-Leiblerovy (KL) divergence. Uvažujme rozložení reprezentované modely λ_1 a λ_2. Platí, že tato divergence $D_{KL}(\lambda_1 \,||\, \lambda_2)$ je nulová pouze v případě, že jsou obě rozložení shodná, v našem případě tedy pokud platí $\lambda_1 = \lambda_2$, jinak je hodnota divergence vždy kladná. Zatímco KL divergenci dvou vícerozměrných Gaussových rozložení je možné vyjádřit v uzavřené formě, pro směsi Gaussových rozložení řešení v uzavřené formě neexistuje. Navíc přímé použití KL divergence GMM modelů jako jádrové funkce není možné, protože nevyhovuje Mercerově podmínce (Purdy et al., 2003). V případě GMM modelů je pro stanovení KL divergence možné použít metody Monte Carlo nebo využít aproximace skutečné hodnoty KL divergence pomocí její horní meze (Do, 2003). Ta je totiž vyjádřitelná v uzavřené formě. Druhý uvedený způsob vede na jádrovou funkci vyhovující Mercerově podmínce. Hodnota této aproximace je závislá na vzájemném přiřazení komponent GMM modelů a s permutací komponent jednoho modelu dojde ke změně hodnoty. Proto je vhodné, aby bylo

přiřazení mezi komponentami obou GMM jasně definované. Toho lze dosáhnout například, jsou-li oba modely výsledkem adaptace stejného původního modelu. Uvažujme tedy GMM modely získané pro jednotlivé nahrávky MAP adaptací UBM způsobem popsaným v sekci 4.2.4. Protože mají oba GMM modely shodný počet komponent a liší se pouze vektory středních hodnot jednotlivých komponent (váhy a kovarianční matice komponent nejsou adaptovány), platí

$$D_{KL}(\lambda_1 \mid\mid \lambda_2) \leq \sum_{c=1}^{C} w_c D_{KL}(\mathcal{N}(\boldsymbol{\mu}_{1,c}, \boldsymbol{\Sigma}_c) \mid\mid \mathcal{N}(\boldsymbol{\mu}_{2,c}, \boldsymbol{\Sigma}_c)), \tag{5.21}$$

kde $\boldsymbol{\mu}_{1,c}$ a $\boldsymbol{\mu}_{2,c}$ představují adaptované vektory středních hodnot jednotlivých komponent c modelů λ_1 a λ_2. Pravou stranu (5.21) představující za daných podmínek aproximaci KL divergence ve formě horní meze lze pak vyjádřit ve tvaru

$$d_{KL}(\lambda_a \mid\mid \lambda_b) = \frac{1}{2} \sum_{c=1}^{C} w_c (\boldsymbol{\mu}_{1,c} - \boldsymbol{\mu}_{2,c})^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{\mu}_{1,c} - \boldsymbol{\mu}_{2,c}). \tag{5.22}$$

KL divergence $D_{KL}(\lambda_1 \mid\mid \lambda_2)$ je nesymetrickou mírou rozdílu mezi dvěma rozloženími, symetrickou míru rozdílu získáme jednoduše jako součet $D_{KL}(\lambda_1 \mid\mid \lambda_2) + D_{KL}(\lambda_2 \mid\mid \lambda_1)$. Jádrová funkce odvozená na základě symetrické KL divergence GMM (získaných MAP adaptací UBM) má pak podobu (Campbell et al., 2006b)

$$\begin{aligned} K_{KL-GMM}(\boldsymbol{O}_1, \boldsymbol{O}_2) &= \sum_{c=1}^{C} w_c \boldsymbol{\mu}_{1,c}^T \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_{2,c} \\ &= \sum_{c=1}^{C} \left(\sqrt{w_c} \boldsymbol{\Sigma}_c^{-\frac{1}{2}} \boldsymbol{\mu}_{1,c} \right)^T \left(\sqrt{w_c} \boldsymbol{\Sigma}_c^{-\frac{1}{2}} \boldsymbol{\mu}_{2,c} \right) \\ &= \sum_{c=1}^{C} \dot{\boldsymbol{\mu}}_{1,c}^T \dot{\boldsymbol{\mu}}_{2,c}, \end{aligned} \tag{5.23}$$

kde $\dot{\boldsymbol{\mu}}_c = \sqrt{w_c} \boldsymbol{\Sigma}_c^{-1/2} \boldsymbol{\mu}_c$ a $\boldsymbol{\Sigma}_c^{-1/2}$ je výsledkem Choleskyho dekompozice $\boldsymbol{\Sigma}_c^{-1} = (\boldsymbol{\Sigma}_c^{-1/2})^T \boldsymbol{\Sigma}_c^{-1/2}$. Zpravidla jsou pro komponenty uvažovány diagonální kovarianční matice a výpočet se tím pádem zjednodušuje. Složky vektoru $\dot{\boldsymbol{\mu}}_c$ pak odpovídají složkám $\boldsymbol{\mu}_c$ normalizovaným příslušnou směrodatnou odchylkou a odmocninou váhy dané komponenty. Zřetězením vektorů $\dot{\boldsymbol{\mu}}_c$ pro $c = 1, \ldots, C$ pak získáme supervektor $\dot{\boldsymbol{\mu}}$, na základě kterého lze jádrovou funkci (5.23) jednoduše vyjádřit jako

$$K_{KL-GMM}(\boldsymbol{O}_1, \boldsymbol{O}_2) = \dot{\boldsymbol{\mu}}_1^T \dot{\boldsymbol{\mu}}_2. \tag{5.24}$$

Jádrová funkce K_{KL-GMM} také umožňuje kompaktní formu reprezentace SVM modelu (obdobně k (5.20)) a provedení výpočtu výstupního skóre ve formě skalárního součinu jednoho supervektoru $\dot{\boldsymbol{\mu}}$ reprezentujícího testovanou promluvu $\boldsymbol{O}$ a druhého reprezentujícího model cílového mluvčího s, tedy

$$s_{GMM-SVM}(\boldsymbol{O}, s) = \sum_{i=1}^{L} \alpha_{i,s} y_{i,s} K_{KL-GMM}(\dot{\boldsymbol{\mu}}_{i,s}, \dot{\boldsymbol{\mu}}) + \rho_s = \boldsymbol{u}_s^T \dot{\boldsymbol{\mu}} + \rho_s, \tag{5.25}$$

kde $\boldsymbol{u}_s$ je vážený součet podpůrných vektorů definujících SVM model.

5.2.3 SVM s využitím MLLR adaptačních koeficientů

Problémem akustických příznaků počítaných přes krátké časové intervaly (typicky desítky ms) je mimo jiné závislost jejich rozložení na fonetickém obsahu. Metoda navržená autory (Stolcke et al., 2005) usiluje právě o potlačení tohoto vlivu.

Moderní systémy rozpoznávání řeči využívají často adaptační metody pro přizpůsobení univerzálních (nezávislých na mluvčím) akustických modelů, reprezentovaných skrytými Markovovými modely (HMM), pro mluvčího rozpoznávané promluvy. Metoda maximálně věrohodné lineární regrese (MLLR) (Leggetter – Woodland, 1995) spočívá v aplikaci afinní transformace na vektory středních hodnot komponent *řečových* akustických modelů. Vztah pro transformaci vektoru středních hodnot $\hat{\boldsymbol{\mu}}_c$ původního modelu příslušného komponentě c je možné vyjádřit jako

$$\boldsymbol{\mu}_c = \boldsymbol{A}_k \hat{\boldsymbol{\mu}}_c + \boldsymbol{d}_k, \tag{5.26}$$

kde k je regresní třída přiřazená komponentě c a matice $\boldsymbol{A}_k$ a vektor $\boldsymbol{d}_k$ představují adaptační parametry stanovené pro tuto třídu. Parametry $\boldsymbol{A}_k$ a $\boldsymbol{d}_k$ $(k = 1, \ldots, K)$ jsou stanoveny s ohledem na zajištění maximální hodnoty celkové věrohodnosti rozpoznávané promluvy pro adaptovaný model a dostupný přepis.

Pro účely rozpoznávání mluvčích jsou pak jednotlivé nahrávky reprezentovány vektory $\boldsymbol{b}(\boldsymbol{O})$ tvořenými všemi koeficienty transformačních matic $\boldsymbol{A}_k$ a vektorů $\boldsymbol{d}_k$, kde $k = 1, \ldots, K$. Rozměr vektorů $\boldsymbol{b}(\boldsymbol{O})$ je ovlivněn rozměrem příznakových vektorů používaných rozpoznávačem řeči a zvoleným počtem regresních tříd. Motivace pro využití MLLR transformačních parametrů jako příznaků pro rozpoznávání mluvčích vychází z představy, že transformační koeficienty nejsou ovlivněny fonetickým obsahem promluv. Rozpoznávání mluvčích s využitím MLLR transformačních parametrů spočívá ve vytváření modelu rozdílu mezi univerzálním a adaptovaným akustickým modelem namísto vytváření modelu distribuce akustických příznaků. Tento rozdíl je pak reprezentován koeficienty MLLR transformace.

Pro stanovení parametrů MLLR transformace je nutná znalost předběžného přepisu promluvy, proto je nutné zařazení modulu pro rozpoznávání řeči. Z pohledu systému rozpoznávání mluvčích způsobuje modul rozpoznávání řeči výrazné zvýšení výpočetní náročnosti. Na druhé straně, v komplexních systémech provádějících dvouprůchodové rozpoznávání řeči s využitím adaptovaných modelů (Červa, 2007), jsou MLLR transformační parametry vedlejším produktem adaptačního modulu a celková výpočetní náročnost systému není nijak ovlivněna. V systémech zaměřených výhradně na rozpoznávání mluvčích je však praktičtějším řešením (navíc jazykově nezávislým) nahrazení rozpoznávače řeči založeného na HMM jednodušším vyhodnocením věrohodnosti GMM (Karam – Campbell, 2007) pro předloženou sekvenci příznaků podle (4.13). Cílem je pak obdobně nalézt MLLR transformaci přinášející maximální věrohodnost adaptovaného GMM. Ferras et al. (2007) používají namísto MLLR transformace jí blízkou metodu CMLLR (Constrained MLLR), při té jsou adaptovány vektory středních hodnot a kovarianční matice komponent HMM/GMM shodnou transformací.

Rozhodovací skóre MLLR-SVM systému pro testovanou nahrávku $\boldsymbol{O}$ reprezentovanou vektorem $\boldsymbol{b}$ a model domnělého mluvčího s je při aplikaci lineární jádrové funkce možné vyjádřit jako

$$s_{MLLR-SVM}(\boldsymbol{O}, s) = \sum_{i=1}^{L} \alpha_{i,s} y_{i,s} \boldsymbol{b}_{i,s}^T \boldsymbol{b} + \rho_s = \boldsymbol{u}_s^T \boldsymbol{b} + \rho_s, \tag{5.27}$$

kde $\boldsymbol{u}_s$ je vážený součet podpůrných vektorů definujících SVM model.

5.3 Metody kompenzace variability akustických podmínek pro SVM klasifikátory

Dosud nebyla věnována žádná pozornost metodám pro potlačení vlivu variability akustických podmínek v systémech založených na SVM klasifikátorech, přestože i v případě těchto systémů představuje tato variabilita významnou příčinu poklesu úspěšnosti rozpoznávání mluvčích. Dvěma základními metodami používanými pro potlačení vlivu variability akustických podmínek v systémech založených na SVM jsou metody označované jako projekce rušivých atributů (NAP) a normalizace kovariance uvnitř tříd (WCCN).

5.3.1 Metoda projekce rušivých atributů

Základní myšlenkou metody projekce rušivých atributů (NAP) (Solomonoff et al., 2005) je odstranění informací nesouvisejících s identitou mluvčího z vektorů používaných pro SVM klasifikaci. Toho je dosaženo odebráním D_{chan}-rozměrného podprostoru asociovaného s nežádanou variabilitou akustických podmínek z obrazového prostoru SVM klasifikátoru. Konkrétně jsou vektory transformovány dle vztahu

$$\boldsymbol{b}_{NAP}(\boldsymbol{O}) = \boldsymbol{b}(\boldsymbol{O}) - \boldsymbol{U}(\boldsymbol{U}^T \boldsymbol{b}(\boldsymbol{O})), \tag{5.28}$$

kde $\boldsymbol{b}(\boldsymbol{O})$ je vektor v obrazovém prostoru SVM klasifikátoru příslušný nahrávce $\boldsymbol{O}$ a $\boldsymbol{U}$ je obdélníková matice tvořená D_{chan} ortonormálními sloupci, která definuje NAP podprostor charakteristický pro variabilitu akustických podmínek. Na základě transformace (5.28) lze definovat projekci[3] $\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^T$. Campbell et al. (2006b) navrhuje dva způsoby odhadu parametrů matice $\boldsymbol{U}$. První způsob vyžaduje anotaci akustických podmínek pro jednotlivé nahrávky v trénovací databázi, například ve formě informace o typu mikrofonu použitého pro záznam. Druhý způsob vyžaduje trénovací databázi, ve které je pro jednotlivé mluvčí k dispozici několik nahrávek. Cílem je v tomto případě minimalizace vzájemné vzdálenosti vektorů (transformovaných dle (5.28)) příslušných všem nahrávkám daného mluvčího. Účelová funkce je pak formulována jako souhrnné vyhodnocení přes všechny mluvčí a parametry matice $\boldsymbol{U}$ jsou stanoveny na základě kritéria

$$\boldsymbol{U}^* = \underset{\boldsymbol{U}}{\operatorname{argmin}} \sum_s \sum_{i,j}^{R_s} \left\| \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_i) - \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_j) \right\|^2. \tag{5.29}$$

[3]Projekce je idempotentní lineární transformace, tedy taková pro kterou platí $\boldsymbol{P}^2 = \boldsymbol{P}$. Idempotentní je obecně operace, která při opětovné aplikaci nemění výsledek.

Nyní definujme matici $\boldsymbol{R}$ jako

$$\boldsymbol{R} = \left(\tilde{\boldsymbol{b}}(\boldsymbol{O}_{1,1}) \dots \tilde{\boldsymbol{b}}(\boldsymbol{O}_{R_1,1}) \dots\dots \tilde{\boldsymbol{b}}(\boldsymbol{O}_{1,S}) \dots \tilde{\boldsymbol{b}}(\boldsymbol{O}_{R_S,S})\right), \tag{5.30}$$

kde S je celkový počet mluvčích v trénovací databázi, R_s je počet nahrávek dostupných pro mluvčího s a

$$\tilde{\boldsymbol{b}}(\boldsymbol{O}_{r,s}) = \boldsymbol{b}(\boldsymbol{O}_{r,s}) - \bar{\boldsymbol{b}}_s, \tag{5.31}$$

kde

$$\bar{\boldsymbol{b}}_s = \frac{1}{R_s} \sum_r \boldsymbol{b}(\boldsymbol{O}_{r,s}). \tag{5.32}$$

S ohledem na kritérium (5.29) je optimální projekční matice $\boldsymbol{P}$ tvořena vlastními vektory příslušnými největším vlastním číslům matice $\frac{1}{R}\boldsymbol{R}\boldsymbol{R}^T$ (Campbell et al., 2006b), kde $R = \sum_s R_s$.

Poznamenejme ještě, že platí

$$\boldsymbol{R}\boldsymbol{R}^T = \sum_{r=1}^{R} \tilde{\boldsymbol{b}}(\boldsymbol{O}_{r,s})^T \tilde{\boldsymbol{b}}(\boldsymbol{O}_{r,s}). \tag{5.33}$$

Pro D-dimenzionální vektory $\boldsymbol{b}$ má matice $\boldsymbol{R}$ rozměr $D \times R$. Pro velký rozměr D (např. rozměr supervektorů) může být přímý výpočet vlastních vektorů matice $\boldsymbol{R}\boldsymbol{R}^T$ (rozměru $D \times D$) nezvládnutelný. Obvykle ale platí, že $R \ll D$ a namísto výpočtu vlastních vektorů matice $\boldsymbol{R}\boldsymbol{R}^T$ je výhodné hledat vlastní vektory matice $\boldsymbol{R}^T\boldsymbol{R}$ (tato matice má rozměr $R \times R$) (Brummer et al., 2007). Ty po vynásobení maticí $\boldsymbol{R}$ odpovídají[4] vlastním vektorům matice $\boldsymbol{R}\boldsymbol{R}^T$. Tedy s rozdílem, že nejsou ortonormální, a proto musí být jejich délka normalizována na rovnou jedné.

Lineární jádrovou funkci zahrnující NAP transformaci SVM příznakových vektorů lze vyjádřit jako

$$K(\boldsymbol{O}_1, \boldsymbol{O}_2) = \langle \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_1), \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_2) \rangle \tag{5.34}$$

$$= (\boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_1))^T \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_2) \tag{5.35}$$

$$= \boldsymbol{b}(\boldsymbol{O}_1)^T \boldsymbol{P}\boldsymbol{b}(\boldsymbol{O}_2) \tag{5.36}$$

kde rovnost (5.35) a (5.36) plyne z idempotence transformace. Uvedený vztah však není úplně jednoznačný ve smyslu, jak má být NAP transformace správně aplikována. Z porovnání provedeného v (Brummer et al., 2007) vyplývá, že je důležité provést NAP transformaci vektorů použitých pro trénování SVM modelů. NAP transformace vektorů při rozpoznávání pak není v důsledku idempotence nutná. Použití NAP transformace pouze pro testovaná data však není účinné, stejně jako aplikace NAP transformace na podpůrné vektory natrénovaných modelů mluvčích.

[4] Protože je-li vektor $\boldsymbol{v}$ vlastním vektorem příslušným vlastnímu číslu λ matice $\boldsymbol{R}^T\boldsymbol{R}$, musí platit $\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{v} = \lambda\boldsymbol{v}$. Je zřejmé, že platí $\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{v}) = \boldsymbol{R}(\lambda\boldsymbol{v})$, což lze přepsat jako $(\boldsymbol{R}\boldsymbol{R}^T)(\boldsymbol{R}\boldsymbol{v}) = \lambda(\boldsymbol{R}\boldsymbol{v})$. V čehož je patrné, že vektor $\boldsymbol{R}\boldsymbol{v}$ je vlastním vektorem matice $\boldsymbol{R}\boldsymbol{R}^T$.

5.3.2 Metoda normalizace kovariance uvnitř tříd

Metoda normalizace kovariance uvnitř tříd (WCCN) nebyla primárně navržena s ohledem na omezení vlivu variability akustických podmínek, ale minimalizace horní meze míry chybného přijetí a zamítnutí (Hatch – Stolcke, 2006). Metoda předpokládá zobecněnou lineární jádrovou funkci, tj.

$$K(\boldsymbol{O}_1, \boldsymbol{O}_2) = \boldsymbol{b}(\boldsymbol{O}_1)^T \boldsymbol{R}_{WCCN} \boldsymbol{b}(\boldsymbol{O}_2), \tag{5.37}$$

kde $\boldsymbol{b}(\boldsymbol{O})$ je vektor v obrazovém prostoru SVM klasifikátoru příslušný nahrávce $\boldsymbol{O}$ a $\boldsymbol{R}_{WCCN}$ je pozitivně semidefinitní matice. S ohledem na uvedené kritérium minimalizace horních chybových mezí je tato matice stanovena jako $\boldsymbol{R}_{WCCN} = \boldsymbol{W}^{-1}$, kde $\boldsymbol{W}$ je očekávaná kovarianční matice tříd. Třídy jsou v našem případě reprezentovány mluvčími v trénovací databázi. Platí tedy

$$\boldsymbol{W} = \sum_s P(s) \boldsymbol{\Sigma}_s, \tag{5.38}$$

kde $P(s)$ je apriorní pravděpodobnost třídy (mluvčího) s a $\boldsymbol{\Sigma}_s$ je kovarianční matice příslušná této třídě, platí tedy

$$\boldsymbol{\Sigma}_s = \frac{1}{R_s} \sum_r (\boldsymbol{b}(\boldsymbol{O}_{r,s}) - \bar{\boldsymbol{b}}_s)(\boldsymbol{b}(\boldsymbol{O}_{r,s}) - \bar{\boldsymbol{b}}_s)^T \tag{5.39}$$

a $\bar{\boldsymbol{b}}_s = \frac{1}{R_s} \sum_r \boldsymbol{b}(\boldsymbol{O}_{r,s})$. Pro odhad parametrů matice $\boldsymbol{R}_{WCCN}$ je tedy opět vyžadována trénovací databáze, ve které je pro jednotlivé mluvčí dostupných několik nahrávek.

Pokud je množství trénovacích dat pro stanovení parametrů $\boldsymbol{W}$ malé, odhad nemusí být příliš robustní. (Hatch – Stolcke, 2006) navrhuje řešení tohoto problému provedením „vyhlazení“ na základě interpolace s jednotkovou maticí ve formě

$$\boldsymbol{W}_s = (1 - \alpha)\boldsymbol{W} + \alpha \boldsymbol{I}, \tag{5.40}$$

kde α je volitelný parametr, který musí splňovat podmínku $0 \geq \alpha \geq 1$, řídící váhu stanoveného odhadu a jednotkové matice.

Další možnou komplikací při odhadu parametrů $\boldsymbol{R}_{WCCN}$ může být příliš vysoký rozměr příznakových vektorů, v důsledku kterého může být inverze $\boldsymbol{W}$ neproveditelná. Řešením tohoto problému se zabývá (Hatch et al., 2006), princip spočívá v rozdělení původního příznakového prostoru do dvou podprostorů – nízkodimenzionálního prostoru nalezeného metodou analýzy hlavních komponent (PCA) a doplňkového prostoru. Metoda WCCN je pak aplikována pouze v PCA podprostoru.

Kapitola 6

Rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů

Pozornost Laboratoře počítačového zpracování řeči působící na Technické univerzitě v Liberci (TUL) je zaměřena na řadu praktických aplikací a jedním z řešených problémů je přepis záznamů televizních a rozhlasových pořadů (TVR) (Nouza et al., 2006) a tvorba multimediálních archivů (Zdansky et al., 2008) umožňujících efektivní vyhledávání v těchto záznamech. Významnou součástí přepisů je nejen informace o tom co bylo řečeno, ale také kým. Rozpoznávání mluvčích v těchto záznamech přitom představuje komplikovanou úlohu. Zejména ve zpravodajských pořadech dochází k častému střídání mluvčích a prostředí, je využíváno velké množství různých mikrofonů a přenosových kanálů. Délka souvislých promluv mluvčích je typicky krátká a proměnná v poměrně velkém rozsahu (od několika vteřin po desítky vteřin). Vysokou variabilitu vykazuje také množství trénovacích dat dostupné pro zapsání jednotlivých mluvčích (od desítek vteřin až po desítky minut řeči).

Práce autora se značnou měrou soustředila právě na návrh systémů pro rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů. Specifický charakter řešené úlohy byl popsán společně s návrhem systému pro komplexní klasifikaci audio segmentů v článku (Silovsky – Nouza, 2006). V rámci tohoto systému bylo později analyzováno několik základních řešení založených jak na generativním, tak diskriminativním přístupu (Silovsky et al., 2009). Konkrétně byly implementovány tři systémy:

- UBM-GMM systém – využívající modely odvozené MAP adaptací UBM
- GMM-SVM systém – založený na SVM klasifikaci supervektorů středních hodnot GMM modelů odvozených MAP adaptací UBM
- MLLR-SVM systém – založený na SVM klasifikaci koeficientů MLLR transformačních matic.

Jedním z přínosů této části práce je návrh vlastní evaluační databáze sloužící k jednotnému porovnání implementovaných řešení. V předešlém textu byl opakovaně zdůrazněn význam evaluací

pořádaných americkým úřadem NIST. Navzdory rostoucímu množství záznamů a rozšiřování evaluačních dat o nové typy záznamů jsou některé vlastnosti použitých nahrávek příliš jednotvárné a neodpovídají tak reálným aplikacím. S výjimkou posledního ročníku je jednou z těchto vlastností například délka promluv používaných pro trénování modelů mluvčích a obdobně délka rozpoznávaných promluv. Základní (povinné) evaluační zadání (tzv. core condition) zahrnuje nahrávky, které obsahují přibližně 2,5 minuty řeči (čisté řeči, délka nahrávek je v závislosti na typu záznamu zhruba 3 až 5 minut) pro trénování a rozpoznávání. V praxi však existuje řada aplikací, kde je zejména pro rozpoznávání dostupných pouze několik vteřin řeči a i množství trénovacích dat může být značně omezené. V NIST evaluacích je definováno volitelné zadání zabývající se vyhodnocením systémů pracujících s omezeným množstvím dat. Typicky je dostupných 10 vteřin řeči pro trénování a 10 vteřin řeči pro rozpoznávání. Další evaluační zadání definovaná v závislosti na množství dostupných dat umožňují vyhodnotit přínos rostoucího množství dat na výkon systémů. Předpoklad shodného množství trénovacích dat dostupného pro trénování všech modelů a shodné délky všech rozpoznávaných promluv, tak jak je aplikován v rámci jednotlivých zadání, je však značně nereálný. Reálným aplikacím vzdálené je ale také množství dat dostupných pro vývoj systémů, protože v každém ročníku smějí účastníci evaluace použít data z předešlých ročníků jako vývojová. V reálných aplikacích však často není takový objem anotovaných dat k dispozici a tomu je nutné podřídit výběr metod.

6.1 Specifikace evaluační databáze a vyhodnocení systémů

Evaluační databáze byla vytvořena na základě rozsáhlého korpusu českých televizních a rozhlasových pořadů vytvářeného na TUL v průběhu více než pěti let. Záznamy byly ručně rozděleny do segmentů, ve kterých nedochází k žádným změnám mluvčích. Délka těchto segmentů je nejčastěji v intervalu 5 až 15 vteřin.

6.1.1 Evaluační úloha

Evaluační úloha byla formulována jako verifikační. Z uvedeného korpusu bylo vybráno celkem 229 mluvčích (131 mužů a 98 žen), pro které bylo k dispozici alespoň 30 vteřin dat dostupných pro trénování jejich modelu a dalších 30 vteřin pro testování (bez ohledu na příslušný počet segmentů). Podmínkou bylo, aby trénovací a testovací data jednotlivých mluvčích pocházela z odlišných záznamů pořadů. Takto vytvořená testovací sada obsahovala celkem 4 245 nahrávek. Dále byla rozšířena o 2 436 nahrávek pocházejících od pro systém neznámých (nezapsaných) mluvčích a obsahovala tedy celkem 6 681 nahrávek. Následně byla testovací sada rozdělena do dvou sad tak, že žádný mluvčí nebyl přítomný v obou sadách, a experimenty byly provedeny vzájemnou validací (první sada byla použita pro kalibraci systému a druhá pro vyhodnocení a naopak). Do první sady bylo zařazeno 3 332 testovacích nahrávek od 111 (56 mužů a 55 žen) zapsaných a 151 nezapsaných mluvčích (97 mužů a 54 žen) o celkové délce 6,71 hod. V rámci pre-evaluačních experimentů bylo pro každou testovanou nahrávku dvěma systémy (UBM-GMM a SVM-GMM) identifikováno pět

nejpravděpodobnějších kandidátů ze zapsaných mluvčích. Evaluační soudy pak byly pro každou nahrávku provedeny pro tyto mluvčí. V závislosti na shodě obou použitých systémů tak bylo pro nahrávky provedeno 5 až 10 verifikačních soudů. Celkem bylo v rámci první sady provedeno 24 966 verifikačních soudů, z toho bylo 2 067 soudů oprávněných. Druhá sada pak sestávala z 3 349 nahrávek pocházejících od 118 (75 mužů a 43 žen) zapsaných a 144 (102 mužů a 42 žen) nezapsaných mluvčích o celkové délce 6,56 hod. V rámci druhé sady bylo provedeno 24 671 verifikačních soudů, z toho 2 099 oprávněných.

6.1.2 Evaluační metriky

Pro vyhodnocení úspěšnosti systémů v úloze verifikace byly použity dvě metriky. Standardně používaná metrika EER reflektující rozlišovací schopnosti systému a metrika C_{llr} zohledňující jak rozlišovací schopnosti systému, tak kvalitu jeho kalibrace. Na závěr je provedeno porovnání systémů na základě DET charakteristik.

6.2 Experimentální vyhodnocení systémů

6.2.1 Využití trénovacích dat

Data potřebná pro účely trénování UBM, metod pro kompenzaci variability akustických podmínek a dále pro reprezentaci neoprávněných mluvčích při trénování SVM modelů byla vybrána z ostatních nahrávek dostupných v uvedené databázi a pocházejících od mluvčích, kteří nebyly zapsáni do systému.

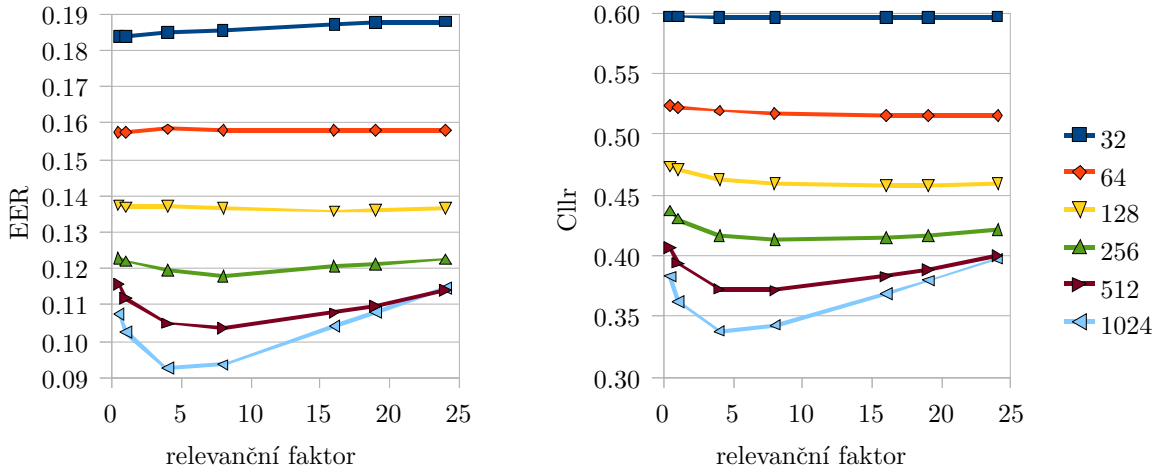
Všechny vyhodnocované systémy byly navrženy jako nezávislé na pohlaví. K natrénování UBM modelu bylo použito celkem 3 668 nahrávek od 432 mužských mluvčích v celkové délce přibližně 8 hod. a 1 686 nahrávek od 245 ženských mluvčích v celkové délce 3,3 hod.

Aplikované metody pro kompenzaci variability akustických podmínek vyžadují trénovací databázi, ve které je pro jednotlivé mluvčí k dispozici několik nahrávek. V našem případě byla tato databáze tvořena 1 516 nahrávkami od 199 mluvčích (137 mužů a 62 žen) a celkové délce přibližně 4 hod.

Dále je nutné vyčlenit nahrávky, které slouží pro vytvoření negativních příkladů při trénování SVM modelů. Pro tento účel byla vytvořena sada 834 nahrávek pocházejících od 674 mluvčích (430 mužů a 244 žen) o celkové délce 2,2 hod.

6.2.2 Parametrizace řečového signálu

V následujících experimentech byly jako akustické příznaky použity standardní Mel-frekvenční kepstrální koeficienty (MFCC). Příznakové vektory byly tvořeny celkem 39 příznaky (12 MFCC + c0 a jejich první a druhou derivací). Na příznakové vektory byla aplikována metoda odečítání kepstrálního průměru (CMS).



Obrázek 6.1: *Vliv počtu komponent a hodnoty relevančního faktoru na úspěšnost rozpoznávání v případě UBM-GMM systému*

6.2.3 UBM-GMM systém

Základními parametry UBM-GMM systému (viz podkapitola 4.2.4) jsou a) počet Gaussovských komponent použitých GMM modelů, b) hodnota relevančního faktoru pro MAP adaptaci a c) v případě systémů využívajících metodu adaptace modelů mluvčích s využitím vlastních kanálů (ECA) pak počet vlastních kanálů. Tyto parametry ovlivňují nejen úspěšnost rozpoznávání, ale s výjimkou hodnoty relevančního faktoru také výpočetní náročnost.

Nejprve byl analyzován vliv hodnoty relevančního faktoru a počtu Gaussovských komponent. Mnoho autorů uvádí, že hodnota relevančního faktoru nemá žádný vliv na úspěšnost rozpoznávání (Reynolds et al., 2000; Kenny et al., 2007b). Výsledky našich experimentů potvrzují tento závěr v případě systémů pracujících s nižším počtem komponent, zatímco pro systémy využívající modely s více než 256 komponentami se již vliv hodnoty relevančního faktoru projevuje. Převažující vliv na úspěšnost rozpoznávání má však počet Gaussovských komponent. Obr. 6.1 zobrazuje průměrné hodnoty EER a C_{llr} vyhodnocené pro obě evaluační sady. Přestože EER na rozdíl od C_{llr} nehodnotí kvalitu kalibrace systémů, vykazují obě sledované metriky obdobný průběh. Zdá se tak, že pro všechny systémy zůstává příslušný rozdíl C_{llr} a C_{llr}^{min}, který reflektuje výhradně kvalitu kalibrace, přibližně stejný. Pro kalibraci skóre byla použita logistická regrese. Při porovnání vah LLR transformačních funkcí stanovených na první a druhé evaluační sadě bylo zjištěno, že jejich hodnoty jsou zpravidla téměř totožné. Lze to přisuzovat shodnému charakteru dat v obou testovacích sadách (vytvořených rozdělením jedné databáze), které umožňují provedení dobré vzájemné kalibrace. Hodnota C_{llr} je tak určována především rozlišovacími schopnostmi systému a blíží se hodnotě C_{llr}^{min}.

Metoda adaptace modelů mluvčích s využitím vlastních kanálů

Metoda adaptace s využitím vlastních kanálů (ECA) umožňuje přizpůsobení GMM modelu mluvčího, trénovaného za libovolných akustických podmínek, odlišným akustickým podmínkám rozpoznávané promluvy. GMM supervektor $\boldsymbol{s}_{r,s}$, tvořený zřetězením vektorů středních hodnot komponent modelu, které jsou navíc *normalizovány* příslušnými směrodatnými odchylkami, je adaptován dle vztahu

$$\boldsymbol{m}_{r,s} = \boldsymbol{s}_s + \boldsymbol{U}\boldsymbol{w}_{r,s}, \tag{6.1}$$

kde $\boldsymbol{U}$ je matice definující prostor *vlastních kanálů*[1], $\boldsymbol{w}_{r,s}$ je váhový vektor specifický pro nahrávku r mluvčího s, $\boldsymbol{s}_s$ je supervektor příslušný natrénovanému modelu mluvčího a konečně $\boldsymbol{m}_{r,s}$ je supervektor modelu adaptovaného na akustické podmínky testované promluvy. Sloupce matice $\boldsymbol{U}$ jsou označovány jako vlastní kanály (eigenchannels) a představují směry, ve kterých se v největší míře projevuje v prostoru supervektorů variabilita akustických podmínek. Počet vlastních kanálů je v následujícím výkladu označen D_{chan}. Složky vektoru $\boldsymbol{w}_{r,s}$ jsou označovány jako kanálové faktory a jsou stanoveny na základě maximalizace věrohodnosti

$$P(\boldsymbol{O}_{r,s}|\boldsymbol{s}_s + \boldsymbol{U}\boldsymbol{w}_{r,s})P(\boldsymbol{w}_{r,s}). \tag{6.2}$$

Vztah (6.1) odpovídá dekompozici supervektoru $\boldsymbol{m}_{r,s}$ v souladu s (4.44) na složku specifickou pro mluvčího a složku specifickou pro akustické podmínky. Podobně jako v případě JFA modelu (4.45) je předpokládáno, že variabilita akustických podmínek je obsažena v nízkodimenzionálním podprostoru supervektorového prostoru. Rozdíly ve srovnání s JFA systémem však spočívají ve způsobu stanovení matice $\boldsymbol{U}$, stanovení supervektoru odpovídajícího modelu mluvčího a také vyhodnocení věrohodnosti pro testovanou nahrávku.

Vlastní kanály jsou v tomto případě stanoveny metodou analýzy hlavních komponent (PCA) obdobně jako v případě metody NAP, která byla popsána v podkapitole 5.3.1 (Brummer et al., 2007). Opět je tedy pro trénování potřeba databáze, ve které je od jednotlivých mluvčích dostupných několik nahrávek. Jednotlivé vlastní kanály, stanovené jako vlastní vektory způsobem popsaným v sekci 5.3.1 jsou v případě metody ECA normalizovány hodnotou[2] $\sqrt{2e_j}$, kde e_j je vlastní číslo příslušné vlastnímu vektoru j. Ve směrech, ve kterých je nižší variabilita akustických podmínek, je tak prováděna adaptace v nižším rozsahu.

Apriorní rozložení faktorů $\boldsymbol{w}_{r,s}$ v (6.2) je předpokládáno jako standardní normální, tedy $P(\boldsymbol{w}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Kanálové faktory jsou pak pro testovanou nahrávku stanoveny na základě vztahu

$$\boldsymbol{w}_{r,s} = (\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{N}_{r,s}\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{s}_s), \tag{6.3}$$

[1] S ohledem na název metody je převzato označení vlastní kanály, i když je zde uvažován podprostor odpovídající veškeré variabilitě, která způsobuje rozdílnost supervektorů příslušných nahrávkám jednoho mluvčího.

[2] Tato normalizace odpovídá MAP variantě metody ECA, kdy je předpokládáno pro faktory $\boldsymbol{w}_{r,s}$ jako apriorní standardní normální rozložení. V případě ML varianty není předpokládáno žádné apriorní rozložení a není ani prováděna žádná normalizace délky vlastních kanálů.

kde $\boldsymbol{N}_{r,s}$ a $\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{s}_s)$ jsou stejné postačující statistiky jaké byly definovány v sekci 4.3.4 a $\boldsymbol{\Sigma}^{1/2}$ je diagonální matice vytvořená zřetězením vektorů směrodatných odchylek komponent UBM modelu.

V případě, že dochází k vyhodnocení, a tím pádem i adaptaci, více než jednoho modelu mluvčích (například v úloze identifikace nebo v případě T-norm normalizace), je možné dosáhnout snížení výpočetní náročnosti na základě předpokladu, že okupační pravděpodobnosti $\gamma_{t,c}$ vyhodnocené na základě UBM modelu lze v důsledku odvození modelů mluvčích MAP adaptací uvažovat jako shodné pro všechny modely. Protože modely mluvčích se od UBM modelu liší pouze vektory středních hodnot komponent, jediným členem, který je pak v rovnici (6.3) závislý na mluvčím, je člen $\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{s}_s)$. Součin ostatních členů tak stačí spočítat pouze jednou pro testovanou promluvu, nezávisle na počtu mluvčích.

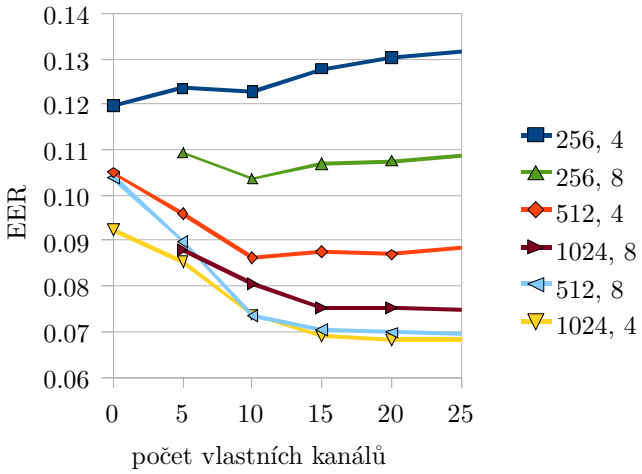
Dále lze dosáhnout snížení výpočetní náročnosti na základě již popsaného principu, vycházejícího z pozorování, že pouze omezený počet k komponent přispívá významnou měrou k celkové hodnotě věrohodnosti GMM modelu pro jednotlivé příznakové vektory, a proto lze příspěvek ostatních komponent považovat za nulový. Na základě UBM modelu je tak možné stanovit pro každý příznakový vektor sekvence k nejvýznamnějších komponent a ve výpočtu $\tilde{\boldsymbol{F}}_{r,s}(\boldsymbol{s}_s)$ pro model mluvčího s následně použít pouze tyto komponenty.

V okamžiku, kdy je znám faktorový vektor $\boldsymbol{w}_{r,s}$ je možné provést adaptaci supervektoru modelu mluvčího na základě vztahu (6.1). Je důležité upozornit, že supervektory jsou pro účely metody ECA normalizovány příslušnými směrodatnými odchylkami a adaptovaný supervektor $\boldsymbol{m}_{r,s}$ proto musí být „denormalizován“ dříve než je použit pro výpočet věrohodnosti adaptovaného modelu na základě vztahu (4.13).

Metoda ECA byla aplikována pro systémy s 256, 512 a 1024 komponentami. Na základě předchozích experimentů lze usoudit, že optimální hodnota relevančního faktoru leží pro systémy s více než 256 komponentami v intervalu mezi 4 až 8 (viz obr. 6.1). Metoda ECA byla proto vyhodnocena pro systémy pracující s těmito nastaveními. Vyhodnocené hodnoty EER (opět se jedná o průměrné hodnoty vyhodnocené pro obě evaluační sady) shrnuje obr. 6.2, nulový počet vlastních kanálů odpovídá systému bez aplikace metody ECA.

Z výsledků vyplývá, že aplikace metody ECA jednoznačně přináší lepší výsledky pro systémy s 512 a 1024 komponentami. Úspěšnost rozpoznávání systému s 1024 komponentami a relevančním faktorem 4 se s rostoucím počtem vlastních kanálů (v testovaném rozsahu jejich počtu) stále zvyšuje, nicméně pro více než 30 vlastních kanálů se jedná již o nevýznamné změny. Toto chování odpovídá efektu normalizace délky vlastních kanálů příslušnými vlastními čísly (seřazenými v sestupném pořadí). Systém s 1024 komponentami a relevančním faktorem 8 dosahuje mírně horších výsledků.

Výraznější vliv relevačního faktoru byl pozorován v případě systému pracujícího s modely s 512 komponentami. Pro tento systém byly lepší výsledky dosaženy při použití vyšší hodnoty relevančního faktoru. Úspěšnost systému s 512 komponentami a relevančním faktorem 8 je srovnatelná se systémem s 1024 komponentami a relevančním faktorem 4, přitom nižší počet komponent vede k výraznému snížení výpočetní náročnosti. Výpočetní náročnost systému s 512 komponentami je

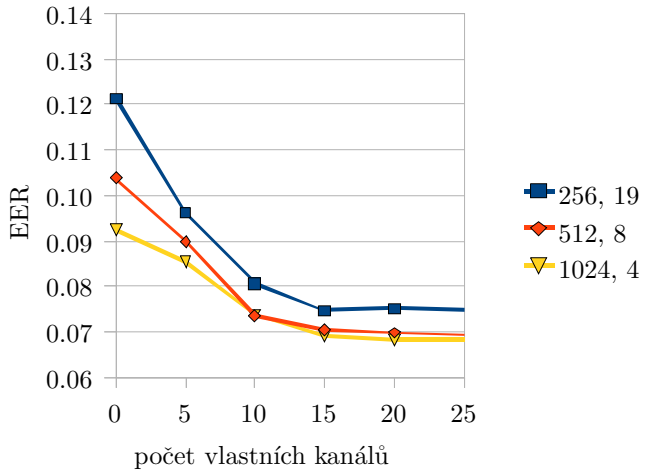


Obrázek 6.2: *Vliv počtu vlastních kanálů při aplikaci metody ECA v případě UBM-GMM systému*

ve srovnání se systémem s 1024 komponentami nižší o 60 % (v případě 30 vlastních kanálů).

V případě systému s 512 komponentami a relevančním faktorem 4 dochází při použití více než 20 vlastních kanálů k poklesu úspěšnosti rozpoznávání. Vzhledem k tomu, že (v důsledku normalizace délky) poslední vlastní kanály ovlivňují metodu ECA v menším rozsahu je přitom toto pozorování v rozporu s očekáváním. Pokusme se tedy o možné vysvětlení. Připomeňme nejprve efekt hodnoty relevančního faktoru v případě MAP odhadu. Zatímco vyšší hodnota relevančního faktoru způsobí, že modely budou bližší původnímu modelu (UBM), nižší hodnota ponechává MAP adaptaci více volnosti. Při daném množství dat a hodnotě relevančního faktoru dojde v případě modelů s nižším počtem komponent k větší výchylce supervektorů příslušných jednotlivým nahrávkám od střední hodnoty odpovídající danému mluvčímu. Výsledkem jsou pak i vyšší hodnoty vlastních čísel, které jsou použity pro normalizaci délky vlastních kanálů. V důsledku toho mohou být supervektory adaptovány ve větším rozsahu než v případě systému využívajícího vyšší hodnotu relevančního faktoru. V případě systému s 256 komponentami a relevančním faktorem 4 došlo po aplikaci metody ECA (bez ohledu na počet použitých vlastních kanálů) dokonce ke snížení úspěšnosti rozpoznávání. Tyto výsledky naznačují, že v případě systémů s nižším počtem komponent a nižší hodnotou relevančního faktoru, lze separovat variabilitu odpovídající hlasu mluvčích a akustickým podmínkám obtížněji. Současně jsou také u těchto systémů supervektory odpovídající modelům mluvčích vzdálenější od supervektoru příslušného UBM a předpoklad, že okupační pravděpodobnosti stanovené na základě UBM lze použít i pro modely mluvčích, nemusí být zcela oprávněný.

Protože je z výsledků patrné, že u systémů s nižším počtem komponent dochází v důsledku snižování hodnoty relevančního faktoru k poklesu úspěšnosti rozpoznávání, byly provedeny experimenty s vyššími hodnotami relevančního faktoru. Obr. 6.3 shrnuje nejlepší dosažené výsledky vzhledem k testovaným hodnotám relevančního faktoru pro systémy s různým počtem komponent. Pro systém s 256 komponentami došlo při použití metody ECA a relevančního faktoru 19 k výraznému zvýšení úspěšnosti rozpoznávání.



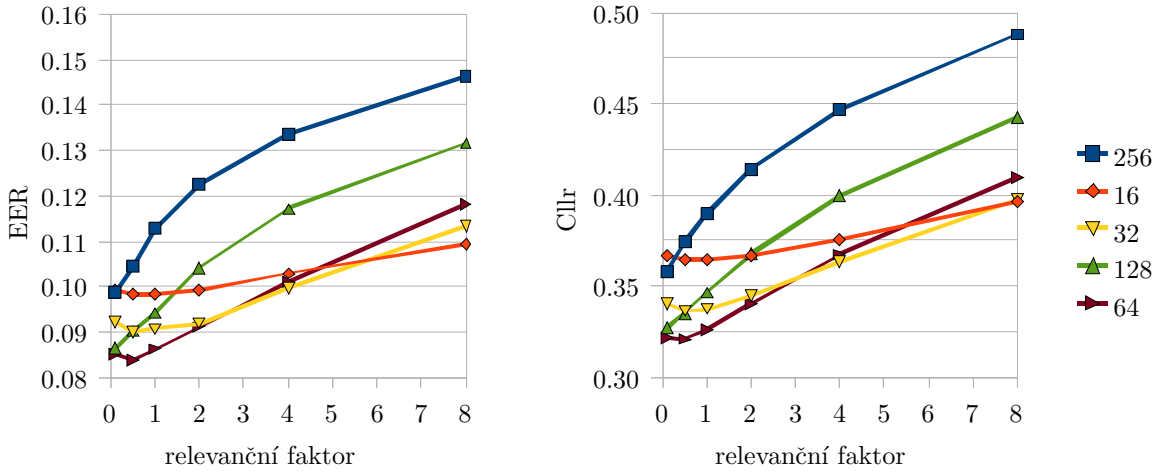
Obrázek 6.3: *Vliv počtu vlastních kanálů při aplikaci metody ECA v případě UBM-GMM systému*

6.2.4 GMM-SVM systém

Tento SVM systém využívá jádrovou funkci odvozenou na základě KL divergence GMM modelů (viz podkapitola 5.2.2). Základní parametry GMM-SVM systému jsou shodné jako v případě UBM-GMM systému, přestože princip obou systémů je odlišný. Parametry jsou tedy a) počet Gaussovských komponent, b) hodnota relevančního faktoru pro MAP adaptaci a c) dimenze NAP podprostoru (analogicky k počtu vlastních kanálů v případě metody ECA).

Nejprve byl opět analyzován vliv hodnoty relevančního faktoru a počtu Gaussovských komponent pro systém nevyužívající kompenzaci variability akustických podmínek. V případě UBM-GMM systému není žádný důvod pro pokles úspěšnosti rozpoznávání u systémů využívajících GMM modely s velkým počtem komponent, protože nepozměněné komponenty (z důvodu nedostatečného množství dat) nemají žádný efekt na poměr věrohodností. V případě GMM-SVM systému vektory středních hodnot nepozměněných komponent vytváří příznakový supervektor stejně jako ostatní komponenty a mohou tak ovlivňovat klasifikaci.

Obr. 6.4 shrnuje vyhodnocení vlivu počtu komponent a hodnoty relevančního faktoru. Opět se jedná o průměrné hodnoty EER a C_{llr} vyhodnocené přes obě evaluační sady. Stejně jako v případě UBM-GMM systému vykazují obě sledované metriky obdobný průběh a potvrzují tak blízkost obou evaluačních sad, umožňující provést dobrou vzájemnou kalibraci. Z výsledků je patrný velký vliv jak počtu Gaussovských komponent, tak hodnoty relevančního faktoru. Přitom rostoucí počet komponent nevede nutně ke zvýšení úspěšnosti rozpoznávání. Podobně jako v případě UBM-GMM systémů lze sledovat vyšší vliv hodnoty relevančního faktoru na systémy pracující s modely s vyšším počtem komponent. V případě GMM-SVM systémů je však tento vliv výraznější a je pozorovatelný i pro systémy, které pracují s nízkým počtem komponent. Zajímavý je ale především interval hodnot relevančního faktoru, pro které dosahují systémy nejlepších výsledků. Systémy, které používají 128 a 256 komponent, dosáhly nejlepší úspěšnosti pro hodnotu relevančního faktoru 0,1, která se

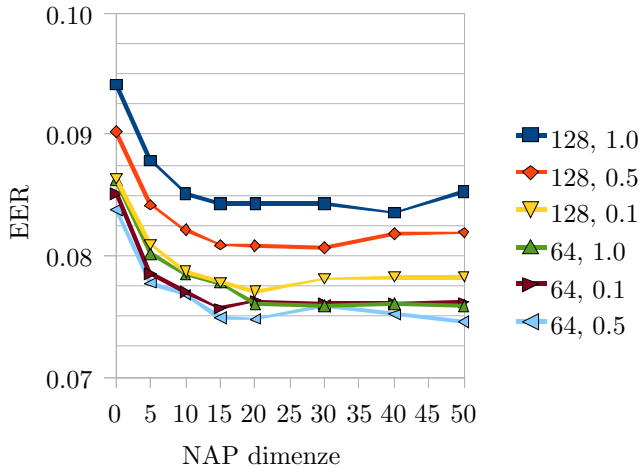


Obrázek 6.4: *Vliv počtu komponent a hodnoty relevančního faktoru na úspěšnost rozpoznávání v případě GMM-SVM systému*

již velmi blíží ML odhadu parametrů modelu[3] (i když pouze jedné iteraci, pro kterou je v případě všech modelů základem UBM model a díky tomu zůstává zachována určitá vazba mezi modely mluvčích). V případě systémů s 32 a 64 komponentami bylo dosaženo nejvyšší úspěšnosti rozpoznávání při použití relevančního faktoru 0,5, nižší počet komponent však znamená, že jednotlivým komponentám jsou přiřazeny vyšší hodnoty celkové okupační pravděpodobnosti. Na základě uvedených výsledků lze usuzovat, že diskriminativní povaze SVM vyhovuje, nejsou-li GMM supervektory příslušné mluvčím příliš koncentrovány blízko jednoho supervektoru (příslušného UBM modelu).

Obr. 6.5 zobrazuje výsledky systémů s aplikací metody NAP pro kompenzaci variability akustických podmínek. Z výsledků vyplývá, že metoda NAP nemá na rozdíl od UBM-GMM systému žádný vliv na optimální hodnotu relevančního faktoru. Ve shodě s výsledky získanými v případě UBM-GMM systému (viz obr. 6.3) se zdá, že variabilita akustických podmínek je převážně soustředěna v 20-rozměrném podprostoru. Aplikace metody NAP přináší jednoznačně zlepšení úspěšnosti rozpoznávání, nicméně ve srovnání s aplikací metody ECA u UBM-GMM systému je pozorované zlepšení v podstatně menším rozsahu. Důvodem je pravděpodobně odlišný princip metody ECA a metody NAP. Zatímco metoda ECA přizpůsobuje model mluvčího akustickým podmínkám testované nahrávky na základě kritéria maximální věrohodnosti, v případě metody NAP je prováděna pouze „slepá“ projekce supervektorů.

[3] Je-li hodnota relevančního faktoru 1,0, bude se nový odhad vektoru středních hodnot komponenty, pro kterou byla na základě předložených adaptačních dat stanovena celková hodnota okupační pravděpodobnosti rovna pouze 1,0, nacházet uprostřed mezi původním odhadem (UBM) a novým ML odhadem (stanoveným pouze na základě předložených dat). Jinými slovy, bude-li k dané komponentě exkluzivně přiřazen jediný příznakový vektor, bude ležet nový odhad střední hodnoty přímo uprostřed mezi původním odhadem a tímto vektorem.



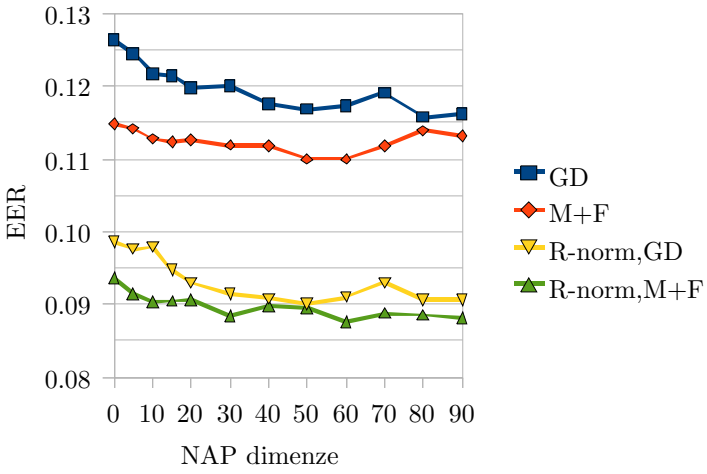
Obrázek 6.5: *Efekt aplikace metody NAP na výsledky GMM-SVM systému v závislosti na dimenzi NAP podprostoru*

6.2.5 MLLR-SVM systém

MLLR-SVM systém je založen na SVM klasifikaci na základě MLLR adaptačních koeficientů (viz podkapitola 5.2.3). Kvalita odhadu MLLR transformačních parametrů je velmi závislá na množství dat a pro nahrávky délky několika vteřin nelze očekávat robustní odhad těchto parametrů. Akustický model pro rozpoznávání řeči byl tvořen třístavovými modely 48 monofonů.

Nejprve byly porovnávány dvě verze MLLR-SVM systémů lišící se v počtu použitých regresních tříd. První systém používal jednu globální transformaci pro všechny Gaussovské komponenty modelu pro rozpoznávání řeči a druhý systém využíval tří regresních tříd. Gaussovské komponenty akustického modelu byly do těchto tříd rozděleny na základě shlukování akusticky podobných komponent. Ani v jednom případě nebyly při odhadu MLLR transformačních parametrů použity neřečové události. Pro SVM klasifikaci byla použita lineární jádrová funkce (ve formě skalárního součinu vektorů). První systém tedy používal SVM vektory tvořené MLLR transformačními parametry s 1 560 koeficienty (byly použity 39-dimenzionální akustické příznakové vektory) a druhý systém pracoval s 4 680 koeficienty. Úspěšnost rozpoznávání systému pracujícího se třemi regresními třídami byla výrazně nižší a lze tak soudit, že dostupné množství dat je nedostatečné pro stanovení tak velkého množství parametrů.

Použitý systém pro rozpoznávání řeči používá modely specifické pro pohlaví mluvčích (tzv. GD modely) a stanovené MLLR transformace jsou tak závislé na zvoleném modelu. MLLR-SVM systém pak ovšem selhává v případě, kdy je v důsledku chybné automatické detekce pohlaví použito různých modelů pro odvození MLLR transformačních parametrů v průběhu trénování SVM modelů a rozpoznávání. Přestože chybná identifikace pohlaví mluvčího pro testovanou nahrávku je spíše výjimečná, podílí se také na celkové míře neúspěšnosti rozpoznávání. Možným řešením, umožňujícím odstranit problém identifikace pohlaví, je využití MLLR transformací odvozených pro oba GD modely (Stolcke et al., 2006), tedy mužský (M) a ženský (F). GD modely jsou navíc trénovány



Obrázek 6.6: *Výsledky dosažené v případě MLLR-SVM systému*

zcela nezávisle a lze tak očekávat, že odvozené MLLR transformační parametry poskytnou vzájemně komplementární informaci. SVM příznakový vektor v tomto případě tvoří 3 120 koeficientů.

Obr. 6.6 poskytuje přehled dosažených výsledků. Výrazné zvýšení úspěšnosti rozpoznávání přináší aplikace metody R-norm (rank normalization) (Stolcke et al., 2006). Důvodem pro aplikaci R-norm je fakt, že SVM jádrové funkce jsou citlivé na rozsah hodnot jednotlivých příznaků. Aby nedocházelo k potlačení vlivu příznaků s malým rozsahem hodnot na úkor příznaků s velkým rozsahem, je nutné používat techniky pro úpravu rozsahu jejich hodnot. V případě GMM-SVM systému je normalizace součástí jádrové funkce. Metoda R-norm je založena na nahrazení hodnoty příznaku jeho pořadím (následně normalizovaným do intervalu 0 až 1) v referenční distribuci. Upozorněme, že aplikace metody R-norm vyžaduje provedení NAP projekce jak pro trénovací, tak i testovaná data.

Výsledky dále potvrzují přínos využití transformací odvozených pro oba GD modely (M+F). Aplikace metody NAP nepřinesla v případě MLLR-SVM systému významné zvýšení úspěšnosti rozpoznávání. Lze tak usuzovat, že příznaky tvořené MLLR transformačními parametry jsou robustní nejenom vůči různému fonetickému obsahu promluv ale i vůči variabilitě akustických podmínek. Výsledky tohoto systému jsou ale ve srovnání s výsledky předchozích porovnávaných systémů horší. Pravděpodobné příčiny jsou diskutovány v následující sekci.

6.3 Shrnutí výsledků

Tabulka 6.1 shrnuje sledované metriky a výpočetní čas[4] nutný pro zpracování testovacích dat (v násobku reálného času, výpočet akustických příznaků MFCC není zahrnut) pro nejúspěšnější konfigurace porovnávaných systémů. Obr. 6.7 pak poskytuje porovnání DET křivek vyhodnocených pro obě evaluační sady. Z provedených srovnání vyplývá, že UBM-GMM systém poskytuje nejvyšší

[4]Výpočetní čas je uváděn jako poměr k celkové délce všech testovaných nahrávek (real-time factor) a odpovídá činnosti jednoho jádra systému s procesorem Intel Core i7 920@2,66 GHz a 3 GB RAM (DDR3@1,6 GHz).

Tabulka 6.1: *Porovnání evaluačních metrik pro nejúspěšnější UBM-GMM, GMM-SVM a MLLR-SVM systémy*

	EER [%]			C_{llr}			
	sada 1	sada 2	⌀	sada 1	sada 2	⌀	× RT
UBM-GMM							
512 c, τ=8, ECA dim. 20	6,91	7,05	6,98	0,25	0,27	0,26	0,158
1024 c, τ=4, ECA dim. 30	7,11	6,58	6,85	0,25	0,26	0,25	0,546
GMM-SVM							
64 c, τ=0,5, NAP dim. 20	6,92	8,04	7,48	0,26	0,30	0,28	0,012
128 c, τ=0,1, NAP dim. 20	7,50	7,91	7,70	0,27	0,31	0,29	0,006
MLLR-SVM							
M+F, R-norm, bez NAP	8,93	9,80	9,37	0,31	0,34	0,33	-

úspěšnost. Výhodou GMM-SVM systému je především jeho rychlost. Důvodů, proč je GMM-SVM systém výrazně rychlejší ve srovnání s UBM-GMM systémem je několik – a) GMM-SVM systém používá modely s nižším počtem komponent, b) výpočet skóre pro SVM model spočívá ve výpočtu jednoho skalárního součinu ve srovnání s komplikovanějším výpočtem věrohodnosti pro GMM modely odvozené MAP adaptací UBM modelu (i v případě, že je uvažován pouze omezený počet nejlepších komponent pro každý příznakový vektor, viz 4.2.4), c) výpočetní nároky metody ECA se výrazně zvyšují s rostoucím počtem vlastních kanálů, protože kanálové faktory jsou stanoveny metodou maximální věrohodnosti, zatímco aplikace NAP nemá prakticky žádný dopad na výpočetní nároky GMM-SVM systému (bez ohledu na dimenzi NAP podprostoru).

Nejnižší úspěšnosti dosáhl MLLR-SVM systém. Pravděpodobným důvodem je velmi krátká délka nahrávek, která neumožňuje robustní odhad MLLR transformačních parametrů. Potvrzením tohoto vysvětlení jsou i výsledky systému využívajícího 3 regresní třídy. V tabulce 6.1 není uveden výpočetní čas pro MLLR-SVM systém, protože čas potřebný pro výpočet skóre pro modely a vektor reprezentující testovanou promluvu je zanedbatelný ($< 0{,}001 \times RT$). MLLR-SVM systém tak působí jako nejrychlejší z porovnávaných systémů. To ovšem platí pouze za předpokladu, že MLLR transformace jsou „vedlejším produktem“ systému pro přepis řeči využívajícího MLLR adaptaci akustických modelů. V případě příznakového vektoru tvořeného MLLR transformačními parametry odvozenými pro oba GD modely je nutné provést MLLR adaptaci druhého modelu (opačného pohlaví) a tento krok výrazně zvyšuje výpočetní náročnost.

6.4 Vzájemné využití informací mezi systémy rozpoznávání mluvčích a řeči

MLLR-SVM systém využívá výsledků systému pro rozpoznávání řeči pro odvození příznakové reprezentace nahrávek vhodné pro rozpoznávání mluvčích, která by v tomto případě neměla být výrazně

ovlivněna fonetickým obsahem nahrávky.

Naproti tomu se autor zabýval i možností využití fonetické závislosti příznakových vektorů. Silovsky et al. (2007) se zabývá vyhodnocením systémů využívajících modely specifické pro definované inventáře akustických jednotek. Od systému rozlišujícího pouze dvě obecné třídy zahrnující všechny znělé a neznělé hlásky až po systém, kde třídy odpovídají jednotlivým stavům modelů fonémů (systém pro rozpoznávání řeči využíval třístavové modely monofonů). Významné zvýšení úspěšnosti rozpoznávání bylo dosaženo při využití akustických modelů specifických pro jednotlivé monofony (celkem bylo uvažováno 40 monofonů, systém pro rozpoznávání řeči disponoval ještě osmi HMM modely neřečových událostí, které byly ignorovány pro účel rozpoznávání mluvčích), kdy došlo ke snížení hodnoty EER z 7,64 % na 5,26 %, tj. relativně o 31,1 % na dané evaluační databázi[5].

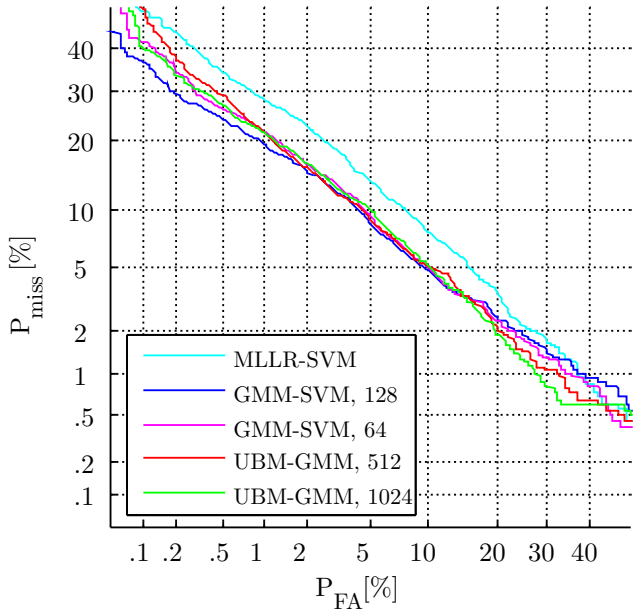
Předmětem zájmu byla také možnost využití informace o identitě mluvčího systémem pro rozpoznávání řeči. Konkrétně byla zkoumána možnost využití informace o mluvčím pro výběr akustického modelu jednoho ze zapsaných mluvčích (pro nejčastěji se vyskytující mluvčí byly natrénovány jejich vlastní akustické modely, v případě TVR záznamů se jedná především o moderátory a reportéry, v případě záznamů parlamentních jednání pak o členy parlamentu) nebo v rámci adaptace akustického modelu na mluvčího rozpoznávané promluvy (Silovsky – Cerva, 2006; Cerva et al., 2006).

V případě (Cerva et al., 2006) byl systém pro rozpoznávání mluvčích použit pro výběr kohorty mluvčích určité velikosti a akustické modely mluvčích v této kohortě byly zkombinovány s vahou nepřímo úměrnou jejich pořadí. Ve srovnání s referenčním systémem (využívajícím univerzální akustický model) bylo dosaženo pro kohortu o velikosti 25 mluvčích relativní snížení chyby rozpoznávání (byla použita míra WER) o 11,3 %. Menšího zlepšení ale bylo dosaženo ve srovnání s použitím modelů specifických pro pohlaví mluvčího (automaticky detekovaného), které odpovídalo relativnímu snížení WER o 4 %. Dále byly aplikovány dvě metody neřízené adaptace akustických modelů, které vyžadují znalost přepisu promluvy. Přesnější přepis umožňuje získání lepšího odhadu parametrů modelu v rámci neřízené adaptace, a proto je pro získání tohoto přepisu použit popsaný kombinovaný model. Aplikace metody provádějící (s využitím znalosti přepisů) maximálně věrohodný odhad vah pro kombinaci modelů z kohorty mluvčích (stanovené systémem pro rozpoznávání mluvčích) přinesla zlepšení míry WER o 22,3 % ve srovnání s referenčním systémem. Tato metoda tak v (Cerva et al., 2006) překonala i výsledky systému využívajího MLLR odhad parametrů.

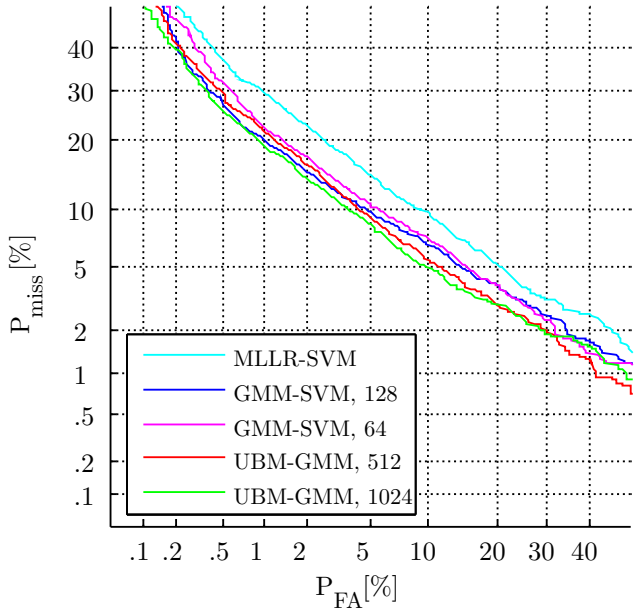
Popsaný způsob využívá kombinace modelů i v případě promluv, které pochází od jednoho z mluvčích, pro které je k dispozici jejich vlastní akustický model. V článku (Silovsky – Cerva, 2006) bylo proto navrženo řešení využívající verifikaci mluvčích. V případě, že si je systém pro rozpoznávání mluvčích dostatečně jistý identitou mluvčího, je použit jemu příslušný model, v opačném případě je aplikován model vytvořený na základě kombinace modelů z kohorty mluvčích. Zatímco

[5]Vyhodnocení bylo provedeno na datech ze stejného korpusu českých televizních a rozhlasových pořadů jaký byl použit pro vyhodnocení systémů popsané v této kapitole, definice evaluačních podmínek však byla odlišná od podmínek popsaných v sekci 6.1.

v případě aplikace kombinace modelů pro všechny promluvy bylo dosaženo snížení míry WER o 10,96 % ve srovnání se systémem využívajícím univerzální akustický model. Přímé použití modelů zapsaných mluvčích v případě, že jejich identita byla prohlášena jako ověřená, přineslo snížení WER o 12,67 %. V souvislosti s přímou aplikací modelu mluvčího navíc odpadá krok kombinace modelů a tím je dosaženo mírného snížení výpočetní náročnosti.



(a) První evaluační sada



(b) Druhá evaluační sada

Obrázek 6.7: *Porovnání DET křivek pro konfigurace UBM-GMM, GMM-SVM a MLLR-SVM systémů uvedené v tab. 6.1*

Kapitola 7

Systémy TUL pro NIST evaluaci rozpoznávání mluvčích 2010

V letech 2008 a 2010 se autor zúčastnil NIST evaluací rozpoznávání mluvčích (SRE) (Silovsky, 2008, 2010). První ročník NIST SRE byl uspořádán v roce 1996 a zatím poslední ročník pořádaný v roce 2010 byl v pořadí již třináctý. Účast autora v roce 2008 představovala první zapojení Laboratoře počítačového zpracování řeči TUL v sérii těchto evaluací. Tato kapitola se zabývá stručným popisem evaluace NIST SRE 2010 a autorem implementovaných řešení v rámci účasti v této evaluaci.

Každý ročník NIST SRE s sebou zpravidla přináší větší objem dat oproti předchozímu a případně také data nového typu. Původně byla evaluační data tvořena pouze záznamy telefonního signálu přenášeného pevnými linkami. Postupně došlo k rozšíření o data pocházející z mobilních sítí a následně o data nahrávaná na mikrofon. Záznam se přitom provádí současně až na čtrnácti různých mikrofonech (různého typu i umístění v záznamové místnosti) (Cieri et al., 2007; Brandschain et al., 2008). Takto rozsáhlý charakter evaluačních dat pak umožňuje vyhodnocení robustnosti systémů vůči rozdílným zdrojům trénovacích a testovacích dat. Zpravidla sledovanými aspekty jsou ale také například vliv pohlaví a jazyka. Je zřejmé, že výkon systémů rozpoznávání mluvčích je výrazně ovlivněn množstvím trénovacích a testovacích dat. V NIST evaluacích je proto definováno několik evaluačních zadání specifických množstvím a charakterem dat dostupných pro trénování modelů mluvčích a jejich rozpoznávání (NIST, 2008, 2010).

7.1 Specifikace NIST SRE 2010

Základní (pro všechny účastníky povinné) zadání NIST SRE 2010 zahrnuje tři druhy nahrávek, které jsou použity pro trénování modelů (model je přitom vždy trénován na základě jediné nahrávky) a jako testovací nahrávky. Jedná se o dvoukanálové nahrávky tvořené přibližně pětiminutovými výňatky záznamů telefonních hovorů, kdy každý z kanálů odpovídá jedné straně hovoru (dále jsou tato data označována „tel“). Dále jsou to dvoukanálové nahrávky telefonních hovorů, kdy je ovšem jedna strana hovoru zaznamenána na mikrofon (tato data jsou dále označovaná jako

„mtel“) umístěný v místnosti volajícího a druhý kanál je telefonní. Pro trénování modelu mluvčího nebo rozpoznávání přitom může být specifikován libovolný z kanálů. Poslední typ nahrávek je tvořen segmenty dvoukanálových záznamů interview (rozhovorů) dotazované osoby a tazatele v záznamové místnosti. Tyto segmenty mají délku od 3 do 15 minut. Poměrně velký rozsah délek nahrávek použitých v rámci jednoho evaluačního zadání (základního) je jedním z nových aspektů evaluace pořádané v roce 2010. Pro srovnání, v předchozím ročníku 2008, byly v případě základního zadání použity mikrofonní nahrávky tvořené tříminutovými segmenty. Ty vzhledem k zpravidla převažujícímu podílu řeči dotazované osoby odpovídají co do množství řeči zapisované či rozpoznávané osoby pětiminutovým nahrávkám telefonních hovorů, u kterých je většinou poměr řeči obou mluvčích vyrovnaný.

V případě mikrofonních záznamů interview je pro trénování modelu zapisovaného mluvčího nebo rozpoznávání vždy použit signál z prvního kanálu, který byl zaznamenán na jeden z 14 mikrofonů rozmístěných v místnosti dle specifikace (Brandschain et al., 2008) (dále jsou tato data označována „int“). Druhý kanál těchto nahrávek obsahuje signál zaznamenaný náhlavním mikrofonem tazatele s přidaným šumem za účelem překrytí případného záznamu hlasu dotazované osoby. Cílem je umožnit detekci intervalů, ve kterých hovoří tazatel. Tyto intervaly je pak nutné vynechat ze zpracování prvního kanálu, protože řeč tazatele je přirozeně v různé míře zaznamenána všemi mikrofony v místnosti. Zdůrazněme, že v rámci základního zadání je k trénování modelů i rozpoznávání použita jediná nahrávka.

Pro všechny trénovací i testovací nahrávky jsou k dispozici automatické přepisy, pro které je uváděna chybovost typicky v rozsahu 15 až 30 % (NIST, 2010). V případě promluv zaznamenaných souběžně několika mikrofony, byl rozpoznáván pouze jeden záznam, a přepis tak může být přesnější, než by odpovídalo provedení rozpoznávání záznamu skutečně použitého pro trénování modelu mluvčího nebo rozpoznávání. Použití těchto příliš optimistických přepisů je jedním z často kritizovaných aspektů evaluace. NIST ale hájí jejich použití úmyslem soustředit se na vyhodnocení systémů pro rozpoznávání mluvčích a ne řeči.

Na rozdíl od předchozích ročníků byly pro evaluaci vybrány pouze nahrávky v angličtině. Není ale požadováno, aby angličtina byla primárním jazykem mluvčích. Pro srovnání, evaluace pořádaná v roce 2008 zahrnovala nahrávky telefonních záznamů ve 26 jazycích (poskytnuté automatické přepisy nahrávek byly ale bez ohledu na jazyk dané promluvy anglické). V NIST SRE 2010 byly poprvé použity záznamy telefonních hovorů, při kterých bylo snahou přimět mluvčího ke sníženému nebo zvýšenému hlasovému úsilí (Brandschain et al., 2008) (pro data s nízkým hlasovým úsilím je dále používáno označení „nhú“ a pro data s vysokým úsilím označení „vhú“).

Protože jsou součástí základního zadání nahrávky různého typu (ve smyslu stylu konverzace, způsobu záznamu nebo evokovaného hlasového úsilí) a množství verifikačních soudů pro různé kombinace typu trénovacích a testovacích dat je značně rozdílné, není příliš vhodné provést jediné společné vyhodnocení přes všechny soudy. Z tohoto důvodu bylo v rámci základního zadání NIST SRE 2010 definováno celkem 9 evaluačních podmínek, které byly vyhodnoceny na příslušné pod-

množině verifikačních soudů v závislosti na použitém typu nahrávek použitých pro trénování modelů a testovacích nahrávek. Konkrétně byly jednotlivé evaluační podmínky tvořeny soudy zahrnujícími:

1. záznamy interview při použití dat ze stejného mikrofonu pro trénování modelu i rozpoznávání (na základě charakteru dat použitých pro trénování a testování lze podmínku označit int – int),
2. záznamy interview při použití dat z rozdílných mikrofonů pro trénování modelu a rozpoznávání (int – int),
3. záznamy interview pro trénování modelu a záznam telefonního hovoru s běžným hlasovým úsilím pro rozpoznávání (int – tel),
4. záznamy interview pro trénování modelu a mikrofonní záznam telefonního hovoru s běžným hlasovým úsilím pro rozpoznávání (int – mtel),
5. záznamy telefonních hovorů s běžným hlasovým úsilím pro trénování modelu i rozpoznávání (tel – tel),
6. záznamy telefonních hovorů s běžným hlasovým úsilím pro trénování modelu a zvýšeným hlasovým úsilím pro rozpoznávání (tel – tel,vhú),
7. mikrofonní záznamy telefonních hovorů s běžným hlasovým úsilím pro trénování modelu a mikrofonní záznamy telefonních hovorů se zvýšeným hlasovým úsilím pro rozpoznávání (mtel – mtel,vhú),
8. záznamy telefonních hovorů s běžným hlasovým úsilím pro trénování modelu a sníženým hlasovým úsilím pro rozpoznávání (tel – tel,nhú),
9. mikrofonní záznamy telefonních hovorů s běžným hlasovým úsilím pro trénování modelu a mikrofonní záznamy telefonních hovorů se sníženým hlasovým úsilím pro rozpoznávání (mtel – mtel,nhú).

Příslušnost verifikačních soudů k uvedeným evaluačním podmínkám není v době konání evaluace známa. Data pro verifikační soudy zahrnující telefonní záznamy pro trénování modelu i rozpoznávání jsou organizátory vybrána tak, že záznamy použité pro trénování a rozpoznávání odpovídají hovorům uskutečněným z různých tel. čísel (tím pádem je předpokládána i rozdílnost použitých tel. zařízení), a to i v případě oprávněných soudů.

Všechna evaluační data (včetně mikrofonních záznamů) jsou kódována v osmibitovém formátu μ-law s vzorkovací frekvencí 8 kHz.

Evaluační metriky

Primární evaluační metrikou je očekávaná pokuta detekčních chyb C_{det} (2.11), resp. její normalizovaná varianta $C_{norm} = \frac{C_{det}}{C_{det0}}$, kde hodnota C_{det0} je definována dle (2.12). Pro základní zadání

Tabulka 7.1: *Porovnání rozsahu evaluací pořádaných úřadem NIST v letech 2006, 2008 a 2010*

	ženy		muži		celkem	
počty	modelů	soudů	modelů	soudů	modelů	soudů
SRE06	462	30 674	354	23 292	816	53 966
SRE08	1 993	59 343	1 270	39 433	3 263	98 776
SRE10	3 026	336 961	2 434	273 787	5 460	610 748
SRE10–extended					11 983	6 924 149

NIST SRE 2010 byly pro výpočet C_{det} specifikovány aplikační parametry $C_{miss} = 1$, $C_{FA} = 1$ a $P_{tar} = 0{,}001$. Jak již bylo uvedeno v podkapitole 2.2.5, toto nastavení představuje významný posun pracovního bodu systémů ve srovnání s parametry $C_{miss} = 10$, $C_{FA} = 1$ a $P_{tar} = 0{,}01$ použitými ve všech předchozích ročnících.

Pro systémy jsou dále vyhodnoceny DET charakteristiky umožňující provést porovnání rozlišovacích schopností systému v celém možném rozsahu pracovních bodů a hodnoty C_{det}^{min} (2.13), které charakterizují rozlišovací schopnosti detektoru ve zvoleném pracovním bodě za předpokladu optimálně stanoveného rozhodovacího prahu.

Účastníci jsou také vyzýváni, aby uvedli, zda skóre určená jejich systémy pro verifikační soudy mohou být interpretována jako logaritmy poměru věrohodností. Evaluační specifikace (NIST, 2010) uvádí, že v tomto případě budou provedena aplikačně nezávislá vyhodnocení, zahrnující např. vyhodnocení C_{llr} (2.23). Autor práce si ale navzdory evaluační specifikaci není vědom, že by NIST po skončení evaluace (alespoň v případě ročníků 2008 a 2010) tato vyhodnocení prezentoval.

Rozsah evaluace

Žádný z verifikačních soudů v NIST SRE neprovádí vyhodnocení modelu a testovací nahrávky rozdílného pohlaví a evaluační soudy tak lze rozdělit na mužské a ženské. Tab. 7.1 shrnuje počet modelů a soudů pro základní zadání NIST SRE 2010 a zprostředkovává také porovnání se dvěma předchozími ročníky. Přestože ve srovnání s evaluací pořádanou v roce 2008 došlo k nárůstu počtu modelů a především vyhodnocovaných soudů, neumožňuje počet soudů, s ohledem na nově definovaný pracovní bod systémů, provedení vyhodnocení s požadovanou statistickou významností (viz podkapitola 2.2.1) pro všechny evaluační podmínky. Je totiž také nutné brát v úvahu, že vyhodnocení není prováděno přes celkový počet soudů uvedený v tab. 7.1, ale je prováděno pro podmnožiny, které odpovídají výše uvedeným evaluačním podmínkám. Po oficiálním skončení evaluace byly proto účastníci vyzváni k provedení (volitelného) vyhodnocení systémů pro rozšířené základní zadání (SRE10-extended).

Evaluační pravidla

Součástí evaluační specifikace je seznam pravidel upřesňujících jakým způsobem je možné s daty nakládat. Skóre určené systémem pro předložený verifikační soud nesmí být podle těchto pravidel

ovlivněno znalostí ostatních evaluačních dat a soudů. Není tak možné použít normalizaci skóre přes několik testovacích nahrávek nebo několik modelů, stejně jako není možné použít evaluační data pro reprezentaci modelu neoprávněných mluvčích. K tomuto účelu je možné použít data z předchozích ročníků.

Pravidla zakazují poslech evaluačních dat, tím pádem je vyloučeno použití ručních přepisů nebo jiných ručně anotovaných informací. Z uvedeného rozdělení soudů na mužské a ženské vyplývá, že informace o pohlaví je dostupná a je možné ji využít. Pro nahrávky je dále dostupná informace o tom, zda se jedná o záznam telefonního hovoru nebo interview mezi tazatelem a dotazovanou osobou v místnosti. V prvním případě je dále poskytnuta informace o použitém kanálu (telefonní nebo mikrofonní). V případě telefonního kanálu ale již není dostupná informace o způsobu přenosu signálu (pevná linka, přenosný telefon nebo mobilní telefon) a typu zařízení (telefon s běžným sluchátkem nebo hlasitým odposlechem, hands-free sada, sluchátka, aj.). U mikrofonních záznamů pak není dostupná informace o použitém mikrofonu.

Jako vývojová data jsou všem účastníkům evaluace dostupná data ze všech předešlých ročníků. Účastníci ale smějí použít libovolný další jim dostupný korpus. Protože v žádném z předešlých ročníků SRE 2010 nebyla použita data s jiným než běžným hlasovým úsilím, bylo účastníkům poskytnuto několik ukázkových nahrávek.

7.2 Systémy TUL

V rámci účasti TUL v NIST SRE 2010 byly implementovány tři systémy založené na odlišných principech:

- JFA systém - založený na úplném modelu sdružené faktorové analýzy
- UBM-GMM systém - využívající modely odvozené MAP adaptací UBM a metodu adaptace modelů mluvčích s využitím vlastních kanálů
- i-vektorový systém - využívající reprezentaci modelů a nahrávek pomocí i-vektorů a skóre založené na kosinové vzdálenosti páru těchto i-vektorů.

V rámci vývoje všech systémů byla použita pouze data z předchozích ročníků NIST SRE evaluací. Využití těchto dat shrnuje tab. 7.2. Objem dat použitých pro jednotlivé účely bude uveden v následujícím textu. Experimentální vyhodnocení systémů v rámci jejich vývoje bylo prováděno podle základního zadání evaluace pořádané v roce 2008.

7.2.1 Společné vlastnosti systémů

Navzdory odlišným principům všechny systémy vyžadují natrénování UBM modelu. Společných rysů je ale více. Všechny systémy byly vyvíjeny jako závislé na pohlaví a ve všech případech byly použity stejné akustické příznakové vektory.

Tabulka 7.2: *Přehled dat použitých při vývoji systémů TUL pro NIST SRE 2010 a jejich využití*

	UBM	JFA systém			UBM-GMM	i-vektorový systém			normalizace			kalibrace
		$\boldsymbol{V}$	$\boldsymbol{U}$	$\boldsymbol{D}$	$\boldsymbol{U}$	$\boldsymbol{T}$	LDA	WCCN	T-	Z-	S-norm	
SRE04	•	•	•	•	•	•	•	•	•	•	•	
SRE05	•	•	•	•	•	•	•	•	•	•	•	
SRE06	•	•	•	•	•	•	•	•	•	•	•	
SRE08	•											•

Detekce řečové aktivity a parametrizace řečového signálu

Pro detekci řečové aktivity byly použity poskytnuté automatické přepisy. Za účelem vytvoření souvislejších řečových segmentů byly ignorovány krátké neřečové intervaly mezi slovy. Pro jednotlivá slova byl navíc s cílem kompenzace možného vzájemného posunu signálu a přepisů jako řečový uvažován rozšířený interval. Rozsah tohoto rozšíření je přitom závislý na typu záznamu (tel. hovor nebo interview dvou osob v místnosti). V případě telefonních nahrávek byl následně použit energetický detektor pro určení tichých intervalů na základě výrazného poklesu energie signálu. V případě mikrofonních záznamů interview tazatele a dotazované osoby jsou ignorovány intervaly, pro které existují automatické přepisy odpovídající záznamu z náhlavního mikrofonu tazatele.

Signál byl parametrizován každých 10 ms s oknem délky 25 ms. Příznakové vektory byly tvořeny 40 příznaky typu MFCC (19 MFCC koeficientů + c_0 a jejich první derivace). Příznakové vektory byly normalizovány metodou feature warping (viz podkapitola 1.3.1), která provádí transformaci rozložení příznakových vektorů v rámci krátkého posuvného okna na normální rozložení. V našem případě bylo použito okno délky 3 s (300 příznakových vektorů). Tato normalizace ruší efekt případně předřazené CMS normalizace aplikované přes celou délku nahrávky a proto CMS normalizace nebyla aplikována. Důvodem pro zrušení efektu CMS je, že při transformaci rozložení nehraje roli absolutní hodnota příznaků, ale jejich vzájemné uspořádání. Odečtení libovolné hodnoty od všech příznaků tak toto pořadí nijak neovlivní.

UBM model

V souvislosti s návrhem systémů jako závislých na pohlaví byly natrénovány UBM modely specifické pro muže a ženy. Modely byly natrénovány jak na telefonních, tak na mikrofonních datech z předchozích evaluací (viz tab. 7.2). Pro trénování UBM modelu pro ženské mluvčí bylo použito 22 872 nahrávek o celkovém objemu 1 931 hod. a pro trénování UBM modelu pro mužské mluvčí 16 899 nahrávek v objemu 1 432 hod. UBM-GMM systém využívá UBM modely s 512 komponentami, zatímco ostatní systémy modely s 1024 komponentami.

UBM modely jsou nutné pro stanovení postačujících statistik nultého a prvního řádu. V případě žádného systému není pro vývojové, trénovací ani testovací nahrávky vyžadována znalost jiné informace, než je znalost postačujících statistik. Tyto statistiky tak lze chápat jako souhrnnou příznakovou reprezentaci nahrávek (je proto výhodné uchovávat pro každou nahrávku statistiky

uložené a neprovádět výpočet při každé práci s nahrávkou).

7.2.2 JFA systém

Tento systém využívá úplný model sdružené faktorové analýzy (viz podkapitola 4.3.2). Protože systém pracuje s UBM modely s 1024 komponentami, je rozměr příslušných supervektorů 40 960. Výpočet skóre pro předložené soudy nebyl proveden na základě vyhodnocení věrohodnosti JFA modelu dle vztahu (4.84), ale byl použit zjednodušený výpočet vycházející z aproximace logaritmu poměru věrohodností navržený autory (Glembek et al., 2009). Skóre pro model reprezentovaný supervektorem $\boldsymbol{s}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s$ (resp. tedy faktory $\boldsymbol{y}_s$ a $\boldsymbol{d}_s$) a testovanou nahrávku $\boldsymbol{O}$ je tak určeno jako

$$s_{linJFA}\left(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{w}\right) = (\boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{F} - \boldsymbol{N}\boldsymbol{\mu} - \boldsymbol{N}\boldsymbol{U}\boldsymbol{w}), \tag{7.1}$$

kde $\boldsymbol{w}$ je vektor bodového MAP odhadu faktorů specifických pro variabilitu akustických podmínek stanovený podle UBM modelu a $\boldsymbol{N}$ a $\boldsymbol{F}$ jsou postačující statistiky nultého a prvního řádu stanovené na základě UBM modelu pro nahrávku $\boldsymbol{O}$.

Odhad hyperparametrů modelu

GMM supervektor tvořený vektory středních hodnot komponent UBM modelu byl použit jako odhad supervektoru $\boldsymbol{\mu}$ nezávislého na mluvčím a konkrétní nahrávce (viz JFA model (4.45)). Obdobně byly použity diagonální kovarianční matice příslušné komponentám UBM pro odhad $\boldsymbol{\Sigma}$. Odhad $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ už nebyl dále aktualizován v průběhu odhadu ostatních hyperparametrů JFA modelu. Matice $\boldsymbol{V}$, $\boldsymbol{U}$ a $\boldsymbol{D}$ byly inicializovány náhodně a jejich odhad byl proveden odděleně (viz podkapitola 4.3.8).

Nejprve byl s využitím telefonních i mikrofonních dat proveden odhad matice $\boldsymbol{V}$ charakterizující variabilitu hlasů mluvčích. Pro oba typy kanálů byly vybrány nahrávky od mluvčích, pro které bylo k dispozici alespoň 8 nahrávek daného typu. Od jednotlivých mluvčích bylo použito nejvýše 32 nahrávek daného typu. Celkem bylo pro odhad $\boldsymbol{V}$ vybráno 9 010 nahrávek od 593 mužů a 11 989 nahrávek od 819 žen. Pro obě pohlaví byla zvolena hodnost matice $\boldsymbol{V}$ rovna 200.

Následně byl proveden odhad matice $\boldsymbol{U}$. Přitom byly zvlášť natrénovány dvě matice $\boldsymbol{U}_{tel}$ a $\boldsymbol{U}_{mic}$ a matice $\boldsymbol{U}$ byla následně sestavena z těchto dílčích matic jako $\boldsymbol{U} = \begin{pmatrix}\boldsymbol{U}_{tel} & \boldsymbol{U}_{mic}\end{pmatrix}$. Matice $\boldsymbol{U}_{tel}$ charakterizuje variabilitu supervektorů příslušných telefonním nahrávkám jednotlivých mluvčích a matice $\boldsymbol{U}_{mic}$ pak variabilitu mikrofonních nahrávek. Pro odhad těchto matic byly vybrány nahrávky mluvčích, pro které bylo k dispozici alespoň 8 nahrávek daného typu. V případě telefonních dat byl maximální počet nahrávek použitých od jednoho mluvčího omezen na 16 a v případě mikrofonních dat pak na 32. Matice $\boldsymbol{U}_{tel}$ byly stanoveny s využitím 6 218 nahrávek od 490 mužů a 8 691 nahrávek od 704 žen. Pro odhad $\boldsymbol{U}_{mic}$ bylo vybráno celkem 2 224 nahrávek[1] od 82 mužů a 2 688 nahrávek od 98 žen. Množství dat dostupné pro odhad $\boldsymbol{U}_{mic}$ je tak značně omezené. Větší množství mikrofonních

[1] Jedná se pouze o mikrofonní záznamy telefonních hovorů, protože mikrofonní záznamy interview tazatele a dotazované osoby vedené v záznamové místnosti byly poprvé přidány v NIST SRE 2008.

dat bylo začleněno do evaluací až od ročníku 2008, který byl ale v našem případě vyčleněn pro účely vyhodnocení vyvíjených systémů. Hodnost obou dílčích matic $\boldsymbol{U}_{tel}$ a $\boldsymbol{U}_{mic}$ byla zvolena rovna 100.

Nakonec byl proveden odhad diagonální matice $\boldsymbol{D}$. Trénovací sada byla v tomto případě tvořena nahrávkami od mluvčích, pro které bylo k dispozici alespoň 5 nahrávek, ale současně méně než 8 nahrávek. Vzhledem k požadavkům použitým při výběru dat pro trénování $\boldsymbol{V}$ a $\boldsymbol{U}$ to znamená, že trénovací sada v tomto případě zahrnovala data od jiných mluvčích, zatímco v případě trénovacích sad použitých pro trénování $\boldsymbol{V}$ a $\boldsymbol{U}$ byla většina mluvčích společná pro obě sady. Celkem bylo pro trénování matice $\boldsymbol{D}$ použito 392 nahrávek od 68 mužů a 528 nahrávek od 89 žen.

Normalizace skóre

Pro normalizaci skóre stanovených dle (7.1) byla použita kombinace metod Z-norm a T-norm (viz podkapitola 1.3.1), pro kterou se používá označení ZT-norm. Skóre byla normalizována v závislosti na pohlaví. Pro normalizaci skóre soudů mužských modelů a testovaných nahrávek bylo v rámci Z-norm použito 261 nahrávek (69 mikrofonních a 192 telefonních záznamů) a v rámci T-norm pak 154 modelů mluvčích (50 natrénovaných na mikrofonních datech a 104 na telefonních datech) modelujících rozložení možných neoprávněných mluvčích. V případě verifikačních soudů mezi ženskými modely a testovacími nahrávkami bylo použito 328 nahrávek (79 mikr. a 249 tel.) pro Z-norm a 205 T-norm modelů (64 mikr. a 141 tel.). Při provádění této normalizace nebyla brána v úvahu informace o typu záznamu použitého pro trénování modelu ověřovaného mluvčího a testované nahrávky.

7.2.3 UBM-GMM systém

UBM-GMM systém je založen na standardním přístupu využívajícím GMM modely mluvčích odvozené relevanční MAP adaptací UBM modelu. Vektory středních hodnot UBM modelu byly adaptovány s relevančním faktorem $\tau = 16$. Pro kompenzaci variability akustických podmínek byla použita metoda adaptace modelů s využitím vlastních kanálů (ECA). Tento systém tedy využívá stejný princip jaký byl popsán v podkapitole 6.2.3. Implementace ale byla v tomto případě odlišná a odpovídá linearizované variantě prezentované autory (Strasheim – Brummer, 2008). Výhodou tohoto přístupu je možnost provést efektivní vyhodnocení skóre pro předložený soud na základě skalárního součinu dvou supervektorů. ECA adaptace byla v tomto případě aplikována nejen při vyhodnocení skóre pro testovanou nahrávku, ale také při trénování modelů mluvčích. Tento systém jako jediný využíval UBM modely s 512 komponentami.

Stejně jako v případě JFA systému byla vytvořena matice $\boldsymbol{U}$ charakteristická pro variabilitu akustických podmínek z dílčích matic $\boldsymbol{U}_{tel}$ a $\boldsymbol{U}_{mic}$, které byly stanoveny na shodných datech. Na rozdíl od JFA systému byly ovšem tyto matice stanoveny metodou analýzy hlavních komponent (PCA). Pro oba typy záznamů bylo stanoveno 50 vlastních kanálů, hodnost matice $\boldsymbol{U}$ tedy byla rovna 100.

Pro normalizaci skóre byla opět použita ZT-norm normalizace závislá na pohlaví. Nahrávky a modely použité pro stanovení rozložení skóre pro neoprávněné soudy byly shodné jako v případě JFA systému.

7.2.4 I-vektorový systém

Tento systém je založen na reprezentaci nahrávek v nízkodimenzionálním prostoru celkové variability prostřednictvím i-vektorů. Skóre pro verifikační soudy je určeno jednoduše na základě kosinové vzdálenosti i-vektoru příslušného modelu (resp. trénovací nahrávce) $\boldsymbol{x}_{train}$ a i-vektoru příslušného testované nahrávce $\boldsymbol{x}_{test}$ podle vztahu (4.88), tedy

$$s_{CDS}(\boldsymbol{x}_{train}, \boldsymbol{x}_{test}) = \frac{\langle \boldsymbol{x}_{train}, \boldsymbol{x}_{test} \rangle}{\|\boldsymbol{x}_{train}\| \, \|\boldsymbol{x}_{test}\|}. \tag{7.2}$$

Systém pracuje s UBM modely s 1024 komponentami a tomu odpovídají 40 960-dimenzionální supervektory. Matice charakterizující prostor celkové variability $\boldsymbol{T}$ byla v případě obou na pohlaví závislých systémů natrénována s využitím telefonních i mikrofonních dat. Pro oba typy kanálů byly vybrány nahrávky od mluvčích, pro které byly k dispozici alespoň 4 nahrávky daného typu. Od jednotlivých mluvčích bylo použito nejvýše 8 mikrofonních nahrávek a 4 telefonní nahrávky. Celkem bylo tímto způsobem vybráno 2 800 nahrávek od 618 mužů a 3916 nahrávek od 881 žen. Prostor celkové variability byl zvolen jako 300-dimenzionální.

Pro potlačení vlivu variability akustických podmínek byla aplikována kombinace metody LDA (viz podkapitola 4.4.3) a WCCN (viz podkapitola 5.3.2). Transformační matice obou metod byly natrénovány na základě stejných dat, která byla použita pro stanovení prostoru celkové variability. V důsledku aplikace LDA byl snížen rozměr i-vektorů z 300 na 200.

Protože skóre vyhodnocená tímto systémem jsou symetrická, byla použita S-norm normalizace (Brummer – Strasheim, 2009). Sada nahrávek sloužící pro stanovení parametrů rozložení skóre určených systémem pro neoprávněné soudy byla shodná s nahrávkami použitými pro Z-norm normalizaci u ostatních systémů.

7.2.5 Kalibrace systémů

Výstupní skóre všech systémů (po provedení ZT-norm nebo S-norm normalizace) byla kalibrována, aby mohla být interpretována jako logaritmus poměru věrohodností. Pro kalibraci skóre byla použita lineární logistická regrese (viz sekce 3.1). Parametry transformační funkce byly stanoveny na základě evaluačních dat z NIST SRE 2008 (tedy stejných dat, které byly použity pro experimentální vyhodnocení systémů v rámci jejich vývoje).

Kalibrace byla provedena ve dvou fázích. Nejprve byla aplikována kalibrace podmíněná pohlavím a kombinaci typu (ve smyslu rozlišování mezi telefonním a mikrofonním záznamem) trénovací a testované nahrávky. Připomeňme, že v rámci základního zadání NIST SRE je použita k trénování modelů pouze jediná nahrávka. Ve druhé fázi pak byla aplikována kalibrace podmíněná pouze

pohlavím mluvčích. Hodnota rozhodovacího prahu byla v souladu se specifikovanými aplikačními parametry (NIST, 2010), kterým odpovídá hodnota efektivní apriorní pravděpodobnosti $\widetilde{P}_{tar} = 0{,}001$ (viz podkapitola 2.2.5), nastavena dle (2.22) na 6,9.

7.3 Vyhodnocení systémů na datech NIST SRE 2008

Jak již bylo zmíněno, pro experimentální vyhodnocení systémů v rámci jejich vývoje bylo použito základní zadání NIST SRE 2008. Jedním z nedostatků ve srovnání s evaluací pořádanou v roce 2010 je absence záznamů s jiným než běžným hlasovým úsilím. S ohledem na velmi omezené množství ukázkových dat poskytnutých organizátory tak nebyly vyvíjené systémy nijak přizpůsobeny práci se záznamy s evokovaným vyšším nebo nižším hlasovým úsilím. Podobně jako v případě SRE 2010 bylo i v případě SRE 2008 definováno s ohledem na charakter trénovacích a testovaných nahrávek několik evaluačních podmínek, pro které je prováděno vyhodnocení. Srovnatelné evaluační podmínky ročníků 2008 a 2010 jsou ty, které zahrnují:

1. záznamy interview při použití dat ze stejného mikrofonu pro trénování modelu i rozpoznávání (int – int),
2. záznamy interview při použití dat z rozdílných mikrofonů pro trénování modelu a rozpoznávání (int – int),
3. záznamy interview pro trénování modelu a záznam telefonního hovoru s běžným hlasovým úsilím pro rozpoznávání (int – tel),
4. záznamy anglických telefonních hovorů s běžným hlasovým úsilím pro trénování modelu i rozpoznávání (tel – tel).

V případě ročníku 2008 ale byly kromě anglických nahrávek použity nahrávky telefonních hovorů v dalších 25 jazycích. Pro anglické telefonní nahrávky bylo navíc zvlášť provedeno vyhodnocení pro podmnožinu soudů zahrnující pouze mluvčí, jejichž rodným jazykem je americká angličtina.

Tab. 7.3 shrnuje dosažené výsledky pro ženské verifikační soudy. Pro možnost posouzení vlivu jazyku jsou uvedeny i výsledky pro evaluační podmínky, které nemají v důsledku omezení na anglické nahrávky ekvivalentní podmínky v NIST SRE 2010. Hodnota C_{norm}^{min} byla stanovena s aplikačními parametry specifikovanými pro NIST SRE 2008 (NIST, 2008), tedy $C_{miss} = 10$, $C_{FA} = 1$ a $P_{tar} = 0{,}01$, kterým odpovídá hodnota $\widetilde{P}_{tar} \approx 0{,}092$. Při souhrnném posouzení lze konstatovat, že nejlepších výsledků dosáhl JFA systém.

Tab. 7.4 představuje výsledky vyhodnocené pro mužské verifikační soudy. Z porovnání s tab. 7.3 je patrné, že všechny tři systémy dosáhly, nezávisle na evaluační podmínce, lepších výsledků. Tento rozdíl v úspěšnosti vyhodnocení mužských a ženských verifikačních soudů byl zaznamenán také u systémů dalších účastníků NIST SRE 2008. Mužské verifikační soudy tak lze v případě této evaluace hodnotit jako jednodušší.

Tabulka 7.3: *Vyhodnocení systémů pro ženské verifikační soudy základního zadání NIST SRE 2008*

	int – int stejný mikr.	int – int různý mikr.	int – tel	tel – tel 26 jazyků	tel – tel angl.	tel – tel amer. angl.
JFA systém						
EER [%]	2,03	5,65	7,65	6,55	3,55	3,95
C_{norm}^{min}	0,120	0,364	0,398	0,340	0,177	0,211
UBM-GMM systém						
EER [%]	1,80	7,87	8,40	8,16	4,43	5,79
C_{norm}^{min}	0,066	0,445	0,358	0,425	0,198	0,253
I-vektorový systém						
EER [%]	2,33	7,53	8,99	8,70	5,58	6,34
C_{norm}^{min}	0,090	0,421	0,445	0,422	0,283	0,315

Tabulka 7.4: *Vyhodnocení systémů pro mužské verifikační soudy základního zadání NIST SRE 2008*

	int – int stejný mikr.	int – int různý mikr.	int – tel	tel – tel 26 jazyků	tel – tel angl.	tel – tel amer. angl.
JFA systém						
EER [%]	0,82	3,74	6,40	5,49	2,96	1,32
C_{norm}^{min}	0,024	0,228	0,301	0,249	0,151	0,104
UBM-GMM systém						
EER [%]	1,23	5,24	5,69	4,80	2,73	1,82
C_{norm}^{min}	0,036	0,299	0,221	0,287	0,185	0,166
I-vektorový systém						
EER [%]	0,82	6,43	5,29	7,55	6,38	4,42
C_{norm}^{min}	0,032	0,333	0,321	0,359	0,279	0,243

Souhrnné výsledky, bez rozlišování pohlaví mluvčích, shrnuje tab. 7.5. Z uvedených výsledků je mimo jiné dobře patrné, že nejhorší úspěšnosti dosahují systémy zpravidla v evaluační podmínce zahrnující rozdílný typ záznamu pro trénování modelů a pro rozpoznávání. V případě podmínek zahrnujících mikrofonní data pro trénování modelů i rozpoznávání dosahují systémy dle očekávání podstatně lepších výsledků pokud jsou použity shodné mikrofony. Naopak, poněkud v rozporu s očekáváním byl pozorován velký jazykový vliv. Vzhledem k povaze implementovaných systémů

Tabulka 7.5: *Vyhodnocení systémů pro všechny verifikační soudy základního zadání NIST SRE 2008*

	int – int stejný mikr.	int – int různý mikr.	int – tel	tel – tel 26 jazyků	tel – tel angl.	tel – tel amer. angl.
JFA systém						
EER [%]	1,73	4,87	7,14	6,19	3,18	3,29
C_{norm}^{min}	0,083	0,308	0,358	0,312	0,170	0,171
UBM-GMM systém						
EER [%]	1,56	6,74	7,24	6,98	3,91	4,42
C_{norm}^{min}	0,059	0,384	0,302	0,382	0,200	0,225
I-vektorový systém						
EER [%]	1,73	7,08	7,59	8,37	5,78	5,59
C_{norm}^{min}	0,069	0,385	0,397	0,403	0,288	0,291

byla očekávána spíše jejich jazyková nezávislost[2]. Možným vysvětlením může být absence vhodných dat při trénování metod pro kompenzaci variability akustických podmínek (příp. příslušného podprostoru) se zaměřením na variabilitu způsobenou rozdílem jazyků. Na druhé straně může být rozdíl způsoben i zahrnutím obtížněji rozhodnutelných verifikačních soudů.

7.4 Vyhodnocení systémů na datech NIST SRE 2010

Výsledky dosažené pro evaluační data NIST SRE 2010, která nebyla v rámci vývoje systémů viděná, shrnuje tab. 7.6. Hodnota C_{norm}^{min} byla v tomto případě stanovena s novými aplikačními parametry specifikovanými pro ročník 2010 (NIST, 2010), tedy $C_{miss} = 1$, $C_{FA} = 1$ a $P_{tar} = 0{,}001$, kterým odpovídá hodnota $\widetilde{P}_{tar} = 0{,}001$. Pro možnost porovnání s výsledky dosaženými v rámci vývojových experimentů je v tab. 7.6 uváděna i hodnota $C_{norm,old}^{min}$, která byla stanovena pro předešlé hodnoty aplikačních parametrů.

Z porovnání výsledků uvedených v tab. 7.5 a 7.6 je patrné, že v případě tří ze čtyř srovnatelných evaluačních podmínek došlo ke snížení úspěšnosti rozpoznávání. Jedinou výjimkou je podmínka „int – tel“ zahrnující mikrofonní záznamy interview pro trénování modelů a telefonní záznamy hovorů pro rozpoznávání. Tato evaluační podmínka je přitom v rámci NIST SRE 2010 jediná, ve které je použit odlišný typ záznamů pro trénování modelů a rozpoznávání. Ve srovnání s těmito výsledky jsou pak ve výrazném kontrastu výsledky dosažené pro podmínku „int – mtel“, která zahrnuje mikrofonní záznamy interview pro trénování modelů a mikrofonní záznamy telefonních hovorů pro rozpoznávání. Charakter konverzací trénovací a rozpoznávané nahrávky je tak sice

[2] Jazykovou závislost lze očekávat například u MLLR-SVM systému využívajícího automatický přepis pro odvození MLLR transformací, které jsou následně použity jako příznakové vektory charakterizující mluvčí.

Tabulka 7.6: *Vyhodnocení systémů pro všechny verifikační soudy základního zadání NIST SRE 2010*

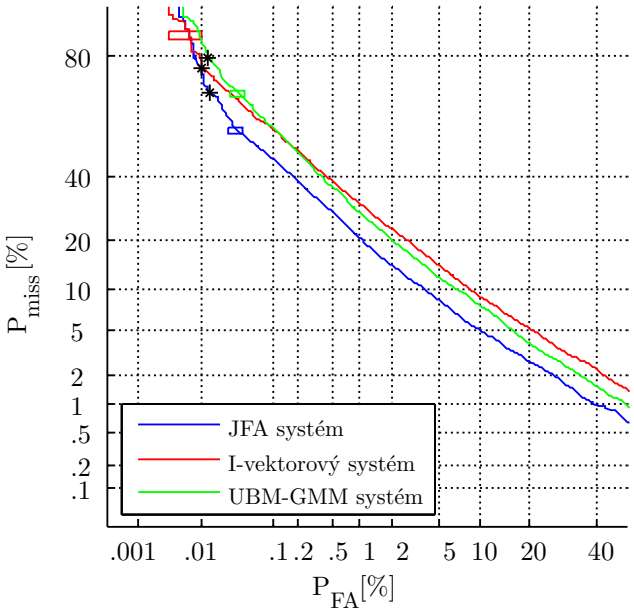
	int – int stejný mikr.	int – int různý mikr.	int – tel	int – mtel	tel – tel	tel – tel,vhú	mtel – mtel,vhú	tel – tel,nhú	mtel – mtel,nhú
JFA systém									
EER [%]	4,61	6,66	5,32	15,00	4,23	8,05	15,35	2,04	9,31
C_{norm}^{min}	0,746	0,823	0,693	0,851	0,852	0,781	0,877	0,418	0,855
$C_{norm,old}^{min}$	0,234	0,303	0,230	0,551	0,207	0,361	0,613	0,094	0,434
UBM-GMM systém									
EER [%]	6,00	8,64	5,32	17,96	5,11	10,80	18,64	2,29	11,72
C_{norm}^{min}	0,827	0,925	0,639	0,918	1,000	0,970	0,969	0,759	0,993
$C_{norm,old}^{min}$	0,282	0,375	0,231	0,635	0,341	0,470	0,691	0,141	0,545
I-vektorový systém									
EER [%]	4,69	9,33	7,17	16,96	8,19	10,58	18,13	4,36	12,08
C_{norm}^{min}	0,628	0,869	0,872	0,849	0,812	0,983	0,911	0,804	0,855
$C_{norm,old}^{min}$	0,213	0,404	0,334	0,585	0,345	0,568	0,697	0,237	0,509

rozdílný, ale v obou případech se jedná o mikrofonní záznam.

S ohledem na posouzení vlivu rozdílného hlasového úsilí lze konstatovat, že zatímco snížené hlasové úsilí nemá žádný negativní vliv na úspěšnost systémů, zvýšené hlasové úsilí vede k výraznému poklesu úspěšnosti rozpoznávání a to jak v případě telefonních „tel – tel,vhú“, tak mikrofonních „mtel – mtel,vhú“ záznamů.

Všechny systémy vykazují jednotně špatné výsledky především v podmínkách zahrnujících mikrofonní záznamy telefonních hovorů. To lze přisuzovat malému množství mikrofonních dat dostupnému pro vývoj systémů v důsledku omezení na data z předchozích ročníků evaluací a vyčlenění dat NIST SRE 2008 pro vyhodnocení vyvíjených systémů. Naproti tomu, vzhledem k uvedenému omezení, nebyly pro vývoj systémů k dispozici žádné mikrofonní záznamy interview vedených v místnosti, a přitom systémy dosahují pro tento typ záznamů lepší úspěšnosti rozpoznávání. Navíc tyto nahrávky vyžadují správnou detekci řeči dotazovaného, protože mikrofon umístěný v místnosti zaznamenává i hlas tazatele.

Společným problémem všech systémů je nízká úspěšnost rozpoznávání v oblasti nového pracovního bodu, jak dokládají hodnoty C_{norm}^{min}. Příčinou je vývoj systémů především s ohledem na dosažení nízké hodnoty EER. Pracovní bod odpovídající EER je přitom velmi vzdálený pracovnímu bodu, který odpovídá nově specifikovaným aplikačním parametrům. Obr. 7.1 zobrazuje srovnání DET charakteristik vyhodnocených pro implementované systémy pro evaluační podmínku „int-int“,



Obrázek 7.1: *Srovnání DET charakteristik implementovaných systémů při vyhodnocení pro evaluační podmínku „int – int“ NIST SRE 2010, ve které jsou pro trénování modelů a rozpoznávání použita data z různých mikrofonů*

ve které jsou pro trénování modelů a rozpoznávání použita data z různých mikrofonů. Z těchto charakteristik je zřejmé, že příčinou vysoké hodnoty C_{norm}^{min} je vysoká míra chybně zamítnutých soudů v pracovním bodě odpovídajícím nízké apriorní pravděpodobnosti výskytu cílových soudů.

7.4.1 Porovnání výpočetní náročnosti systémů

Tab. 7.7 shrnuje výpočetní náročnost systémů[3]. Uváděn je poměr doby nutné pro zpracování všech trénovacích či rozpoznávaných dat a jejich celkové délky (v případě testovacích nahrávek, pro které bylo vyhodnocováno více modelů, je jejich délka započítána pouze jednou). Všechny systémy vyžadují stanovení postačujících statistik jak pro trénovací, tak pro rozpoznávané nahrávky. Tento výpočet tak není nutné provádět zvlášť pro každý systém, ale stačí provést pouze jednou. Pro lepší představu o rozdílu výpočetní náročnosti systémů jsou v tab. 7.7 uvedeny také výpočetní náročnosti bez započtení doby nutné pro výpočet postačujících statistik a stanovení rozložení skóre možných neoprávněných mluvčích v rámci Z-norm a T-norm normalizace.

Z tab. 7.7 je zřejmé, že právě výpočet postačujících statistik se podílí na celkové výpočetní náročnosti v případě všech systémů podstatnou nebo dokonce převažující měrou. Nejnižší výpočetní náročnost má UBM-GMM systém. Celkovou náročnost tohoto systému v porovnání s ostatními snižuje mimo jiné práce s UBM modely s 512 komponentami. Naopak nejvyšší výpočetní náročnost

[3]Výpočetní čas odpovídá činnosti jednoho jádra systému s procesorem Intel Core i7 920@2,66 GHz a 3 GB RAM (DDR3@1,6 GHz).

Tabulka 7.7: *Porovnání výpočetní náročnosti implementovaných řešení*

	trénování modelů x RT	vyhodnocení soudů x RT
Výpočet postačujících statistik		
512 komponent	0,007	0,007
1024 komponent	0,012	0,012
JFA systém		
základní trénování / rozpoznávání	0,004	0,003
+ Z-norm / T-norm	0,012	0,008
+ výpočet statistik (celková náročnost)	0,024	0,019
UBM-GMM systém		
základní trénování / rozpoznávání	0,0008	0,0012
+ Z-norm / T-norm	0,001	0,002
+ výpočet statistik (celková náročnost)	0,008	0,009
I-vektorový systém		
základní trénování / rozpoznávání + S-norm	0,003	0,003
+ výpočet statistik (celková náročnost)	0,015	0,015

má JFA systém.

7.4.2 Post-evaluační experimenty

V rámci post-evaluačních experimentů byly provedeny experimenty s fúzí jednotlivých systémů. Fúze byla stejně jako kalibrace založena na lineární logistické regresi (viz podkapitola 3.3). Váhy příslušné jednotlivým systémům byly stanoveny na základě evaluačních dat NIST SRE 2008. Před aplikací fúze byla skóre jednotlivých systémů kalibrována způsobem popsaným v podkapitole 7.2.5. Fúze již na rozdíl od kalibrace nebyla nijak závislá na typu dat daného soudu.

Výsledky dosažené systémem založeným na fúzi systémů jsou uvedeny v tab. 7.8. Přínos této fúze ale není zcela jednoznačný. Zatímco pro některé podmínky došlo ke zlepšení úspěšnosti, pro jiné byl efekt ve srovnání s nejlepšími výsledky dosaženými jednotlivými systémy (především tedy JFA systémem) opačný.

Tabulka 7.8: *Vyhodnocení systému založeného na fúzi jednotlivých systémů pro všechny verifikační soudy základního zadání NIST SRE 2010*

	int – int stejný mikr.	int – int různý mikr.	int – tel	int – mtel	tel – tel	tel – tel,vhú	mtel – mtel,vhú	tel – tel,nhú	mtel – mtel,nhú
Fúze JFA, UBM-GMM a i-vektorového systému									
EER [%]	4,27	6,45	4,65	15,26	4,24	8,03	15,58	1,68	8,92
C_{norm}^{min}	0,655	0,799	0,658	0,835	0,990	0,884	0,894	0,520	0,899
$C_{norm,old}^{min}$	0,203	0,286	0,186	0,550	0,217	0,344	0,597	0,095	0,409

Kapitola 8

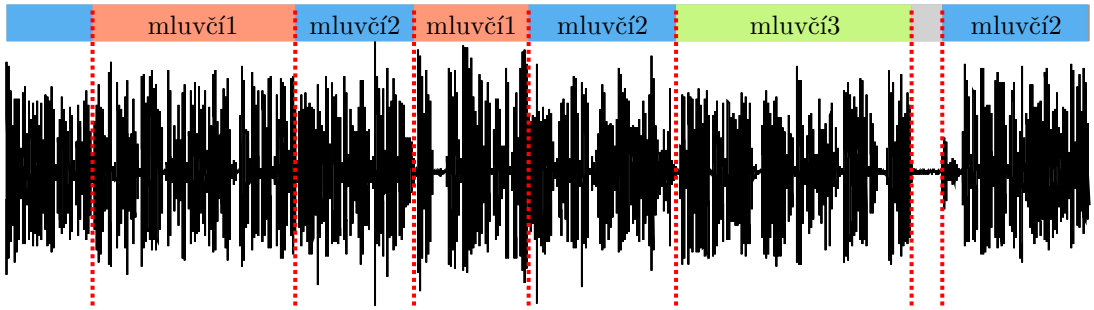
Metody založené na faktorové analýze v úloze diarizace mluvčích

Diariazace audiozáznamů je proces, jehož cílem je rozdělení signálu do homogenních segmentů vzhledem k určité charakteristice zdroje signálu. Tedy vytvoření záznamu (diáře) událostí nastávajících v audiozáznamu. Hovoříme-li o diarizici mluvčích, jsou jako zdroje signálu chápáni mluvčí a úlohu diarizace mluvčích tak lze výstižně formulovat jako otázku „kdy a kdo mluví?“. Informace o počtu a identitě osob hovořících ve zpracovávané nahrávce přitom není diarizačnímu systému známa. Diarizační systém nepracuje s žádnými modely mluvčích a úkolem systému není provést konkrétní identifikaci mluvčích. Obr. 8.1 ilustruje výstup systému pro diarizaci mluvčích.

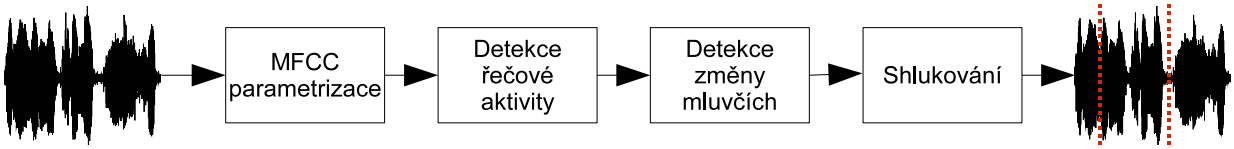
Diariazace mluvčích je užitečným krokem při předzpracování signálu v systémech pro rozpoznávání řeči i mluvčích. Je-li známo, že v daném segmentu hovoří pouze jedna osoba, je možné provést efektivní adaptaci akustických modelů pro rozpoznávání řeči. Je-li navíc známo, že daná osoba hovoří ve více segmentech, je výhodné použít pro adaptaci data všech těchto segmentů. S rostoucím množstvím dat je totiž možné stanovit lepší odhad adaptačních parametrů. Pro úlohu rozpoznávání mluvčích je přirozeně také důležité, aby byla nahrávka rozdělena do segmentů, ve kterých nehovoří více osob.

Diariazace mluvčích je také důležitou součástí systémů pro indexaci záznamů v archivech multimediálních pořadů. Diarizace usnadňuje čitelnost automatických přepisů a ve spolupráci se systémem pro rozpoznávání mluvčích umožňuje vyhledávat v přepisech nejen na základě toho co bylo řečeno, ale také kým. Případně je možné detekovat nové, pro identifikační systém neznámé mluvčí.

Autor práce se v rámci této kapitoly zabývá návrhem systému pro diarizaci mluvčích s využitím metod používaných pro rozpoznávání mluvčích, které jsou založeny na faktorové analýze. S ohledem na již zmíněné zaměření Laboratoře počítačového zpracování řeči TUL na zpracování záznamů televizních a rozhlasových pořadů bylo provedeno vyhodnocení na těchto datech.



Obrázek 8.1: *Ukázka výstupu systému pro diarizaci mluvčích*



Obrázek 8.2: *Schéma diarizačního systému*

8.1 Schéma diarizačního systému

Základní návrh vyvinutého diariazačního systému byl popsán v článku (Prazak – Silovsky, 2011). Schéma systému je standardní (Tranter – Reynolds, 2006) a sestává ze tří hlavních modulů, jak ilustruje obr. 8.2. Poté co je provedena extrakce příznakových vektorů, je aplikován detektor řečové aktivity. Následně je provedena segmentace signálu na základě detekce změny mluvčích. Nakonec je provedeno shlukování segmentů s cílem určit segmenty pocházející od shodného mluvčího. V ideálním případě je výsledkem jeden shluk pro každého mluvčího a všechny segmenty příslušné danému mluvčímu přiřazené v jemu příslušném shluku. Hlavní moduly jsou detailněji popsány v následujících sekcích.

8.1.1 Detekce řečové aktivity

Detektor řečové aktivity má dvě části – energetický detektor s adaptivním prahem a detektor využívající GMM modely. Zatímco cílem aplikace energetického detektoru je určit tiché intervaly v signálu, druhý detektor určuje intervaly odpovídající hlasitým neřečovým událostem, zejména hudbě a různým hlukům. Celkem byly natrénovány modely pro 3 řečové třídy (řeč, řeč s hudbou, řeč s ostatním hlukem) a 2 neřečové třídy (hudba, ostatní hluk). Význam rozšířených řečových tříd spočívá ve snaze potlačení možnosti vyhodnocení řeči s hudbou nebo hlukem na pozadí jako neřečového intervalu a tím pádem vyřazení příslušného segmentu z dalšího zpracování. Detektor používá modely s 64 Gaussovskými komponentami.

Energetický detektor provádí vyhodnocení řečové aktivity postupně po segmentech délky 10 ms, při rozhodnutí je přitom zvažován pravý a levý kontext délky 100 ms. Na závěr činnosti detektoru je provedeno vyhlazení výstupu zrušením neřečových intervalů kratších než 200 ms. Detektor založený na GMM provádí klasifikaci po segmentech délky 200 ms a při vyhodnocení je opět zvažován širší kontext daného segmentu, v tomto případě je to 150 ms v obou směrech. Sousední segmenty klasi-

fikované do jedné z řečových tříd jsou následně sloučeny do delších souvislých řečových segmentů (bez podrobnějšího rozlišování tříd) a postoupeny k dalšímu zpracování.

8.1.2 Detekce změny mluvčích

Modul pro segmentaci mluvčích postupně prochází signál s oknem proměnné délky, v rámci kterého je analyzována možná změna mluvčích. Pro každý příznakový vektor uvnitř okna je vypočtena míra, která hodnotí rozdíl mezi hypotézou, že všechna data v analyzovaném okně jsou reprezentována jedním rozložením, a hypotézou, že jsou data v levém a pravém okolí tohoto vektoru reprezentována rozdílnými rozloženími. Změna mluvčích je detekována v případě, že sledovaná míra převyšuje stanovený rozhodovací práh. Často je míra pro srovnání uvedených hypotéz odvozena na základě Bayesovského informačního kritéria (BIC). Testovací míra ΔBIC pro příznakový vektor v čase t je dle (Chen – Gopalakrishnan, 1998) definována následovně[1]

$$\Delta BIC(t) = (N_1 + N_2)log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| - N_2 log\,|\mathbf{\Sigma}_2| - \alpha P, \tag{8.1}$$

kde N je počet příznakových vektorů a Σ je úplná kovarianční matice. Není-li použit žádný dolní index, jsou uvažována všechna data v analyzovaném okně v intervalu $(a; b)$, index 1 značí příslušnost k intervalu v rozmezí od počátku analyzovaného okna po t, tj. $(a; t)$, a index 2 k intervalu od t do konce analyzovaného okna, tj. $(t + 1, b)$. P je penalizační člen, který je pro d-dimenzionální příznakové vektory roven

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) log(N_1 + N_2) \tag{8.2}$$

a konečně α v (8.1) je penalizační váha (v následujících experimentech byla pro segmentaci vždy použita váha $\alpha = 1$).

Počáteční délka analyzovaného okna je nastavena na w_{min} příznakových vektorů (bylo použito $w_{min} = 100$) a délka je navyšována (dochází k posunu konce analyzovaného okna b) o 50 vektorů vždy, když není nalezena žádná změna mluvčího. V případě, že délka okna dosáhne maximální délky w_{max} (v našem případě 2000 vektorů, tj. 20 s) a stále není v okně detekována žádná změna mluvčích, je umístěna fiktivní změna mluvčích do okamžiku odpovídajícího aktuální hodnotě konce okna b a následně je nově inicializováno okno s počátkem v okamžiku $b + 1$ a délkou w_{min}. Cílem tohoto postupu je snížení výpočetní náročnosti, protože analýza dlouhých oken je výpočetně náročná. Ve většině případů jsou fiktivní změny mluvčích eliminovány ve fázi shlukování.

Pro snížení výpočetní náročnosti nebylo prováděno okamžité vyhodnocení ΔBIC pro každý příznakový vektor analyzovaného okna, ale nejprve byla pro vektory stanovena Hottelingova T^2 statistika

$$T^2(t) = \frac{N_1 N_2}{N_1 + N_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \tag{8.3}$$

[1] Autoři nejsou zcela jednotní v uvádění této míry a někdy, viz např. (Tranter – Reynolds, 2006; Zhou – Hansen, 2000), je uváděna ve tvaru $\Delta BIC(t) = \frac{1}{2}\left((N_1 + N_2)log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| - N_2 log\,|\mathbf{\Sigma}_2|\right) - \alpha P$. Rozdíl obou kritérií je ovšem eliminován případnou změnou váhy α. Tvar (8.1) uvádí kromě (Chen – Gopalakrishnan, 1998) např. (Ben et al., 2004; Zhu et al., 2008).

kde $\boldsymbol{\mu}$ představuje střední hodnotu dat. Kovarianční matice $\boldsymbol{\Sigma}$ v (8.3) je shodná pro všechna t v daném okně a výpočet je ve srovnání s (8.1) méně náročný. V rámci analyzovaného okna je statistika T^2 použita pro vytipování domnělého okamžiku změny mluvčího. Pouze v okolí její maximální hodnoty je pro ověření případné změny mluvčích vyhodnocena ΔBIC testovací statistika. V případě, že je v rámci analyzovaného okna nalezena změna mluvčích v čase t^*, je provedena inicializace nového okna s počátkem v čase $t^* + 1$ a délkou w_{min}.

Rozhodovací práh pro posuzování kritéria ΔBIC byl stanoven na vývojových datech.

8.1.3 Modul pro shlukování mluvčích

Modul pro shlukování mluvčích využívá metodu hierarchického aglomerativního shlukování. Na počátku činnosti algoritmu je každý segment určený segmentačním modulem samostatným shlukem a postupně jsou shluky sdružovány dokud není splněna ukončovací podmínka. Činnost algoritmu lze popsat v následujících krocích:

1. inicializace shluků a vypočtení vzájemné vzdálenosti mezi všemi shluky
2. kontrola ukončovací podmínky a v případě jejího splnění ukončení činnosti
3. sloučení dvou nejbližších shluků do jednoho
4. aktualizace vzdálenosti mezi nově vytvořeným shlukem a všemi ostatními
5. návrat do kroku 2

Klíčovou otázkou je volba vhodné míry pro vyhodnocení vzájemné vzdálenosti mezi shluky. V našem případě je cílem, aby tato vzdálenost reflektovala podobnost mezi mluvčími shluků.

Jak uvádí Tranter – Reynolds (2006), zpravidla jsou shluky reprezentovány jedním vícerozměrným Gaussovým rozložením s plnou kovarianční maticí. V literatuře lze ovšem nalézt i další přístupy, které jsou založeny například na reprezentaci shluků pomocí vícemodálních rozložení ve formě GMM modelů. Možným způsobem odhadu parametrů těchto GMM modelů je odvození modelů MAP adaptací UBM modelu, případně je možné (s ohledem na zvýšení robustnosti) namísto UBM modelu využít GMM model natrénovaný na všech datech zpracovávané nahrávky. Poslední zmíněný přístup byl vyhodnocen i v rámci námi navrženého systému pro diarizaci TVR dat. Vzdálenost mezi shluky byla stanovena na základě míry navržené v (Ben et al., 2004), která je založena na porovnání parametrů GMM modelů příslušných shlukům. Dosažené výsledky shrnuje (Prazak – Silovsky, 2011).

Za základní přístup pro stanovení vzdálenosti shluků lze považovat přístup založený na BIC kritériu. Tento přístup je zde použit v rámci *referenčního systému*.

8.2 Metody pro shlukování mluvčích

Byly navrženy tři přístupy pro shlukování mluvčích založené na metodách běžně používaných v úloze rozpoznávání mluvčích. Všechny prezentované metody využívají reprezentaci řečových segmentů pomocí i-vektorů. První metoda provádí vyhodnocení podobnosti segmentů na základě kosinové vzdálenosti i-vektorů, zatímco zbylé dvě metody jsou založeny na pravděpodobnostní lineární diskriminační analýze (PLDA) (Silovsky et al., 2011). Stejně jako pro systémy rozpoznávání mluvčích je i pro diarizační systémy pracující se záznamy televizních a rozhlasových pořadů důležitá robustnost vůči velké variabilitě nahrávacích podmínek (často proměnlivých) a přenosových kanálů. Výhodou aplikovaných metod založených na i-vektorech je, že umožňují použití technik pro kompenzaci variability akustických podmínek. Dále je prezentováno dvoufázové schéma shlukování, které využívá v první fázi metodu odvozenou od BIC kritéria a ve druhé fázi jednu z metod založených na aplikaci i-vektorů.

8.2.1 Shlukování založené na Bayesovském informačním kritériu

Shluky jsou reprezentovány vícerozměrným Gaussovým rozložením s plnou kovarianční maticí a při vyhodnocení vzájemné vzdálenosti dvou shluků g_1 a g_2 je porovnáváno BIC kritérium pro model, ve kterém jsou oba shluky reprezentovány svými rozloženími, a model, ve kterém jsou shluky sloučeny a tím pádem reprezentovány společným rozložením (pro reprezentaci dat tak stačí nižší počet parametrů). Na základě rozdílu v hodnotě BIC kritéria je definována míra (Chen – Gopalakrishnan, 1998):

$$\Delta BIC(g_1, g_2) = (N_1 + N_2) log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| - N_2 log\,|\mathbf{\Sigma}_2| - \alpha P, \tag{8.4}$$

kde N_i je počet příznakových vektorů přiřazených do shluku g_i a $\mathbf{\Sigma}_i$ je příslušná plná kovarianční matice, $i = 1, 2$. Kovarianční matice $\mathbf{\Sigma}$ je vyhodnocena přes data obou shluků, α je penalizační váha a P je penalizační člen, který je v případě d-dimenzionálních příznakových vektorů roven

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) log(N). \tag{8.5}$$

Dle (Chen – Gopalakrishnan, 1998) je N rovno počtu příznakových vektorů v celé zpracovávané nahrávce (řečových segmentech nahrávky). Tranter – Reynolds (2006) označují tuto variantu jako *globální* a uvádí, že lepších výsledků je dosahováno aplikací *lokální* varianty, pro kterou je $N = N_1 + N_2$. Penalizační faktor pak reflektuje změnu BIC kritéria pouze s ohledem na dvojici porovnávaných shluků. V našem případě byla použita lokální varianta.

Je-li možné popsat dvojici shluků dobře pomocí jednoho Gaussova rozložení, bude hodnota ΔBIC nízká, zatímco pokud je vhodné použít dvě samostatná rozložení, bude hodnota ΔBIC vysoká. V každém cyklu procesu shlukování je sloučen pár shluků s nejnižší hodnotou ΔBIC. Ukončovací podmínka algoritmu je splněna v okamžiku, kdy je nejnižší hodnota ΔBIC vyšší než stanovený práh λ (typicky 0). Výhodou tohoto přístupu je malé množství parametrů, které je nutné

stanovit. Nejsou použity žádné modely jejichž parametry by bylo nutné stanovit a penalizační váha α a práh λ představují jediné parametry.

8.2.2 Shlukování na základě kosinové vzdálenosti i-vektorů

Tento přístup je založen na reprezentaci shluků prostřednictvím i-vektorů a vyhodnocení vzdálenosti mezi shluky na základě kosinové vzdálenosti i-vektorů, dále je označován jako CDS (Cosine Distance Scoring). Pro každý segment je nejprve nutné určit jemu příslušný i-vektor. Pro shluk i jsou tak nejprve na základě UBM modelu vypočteny postačující statistiky nultého a prvního řádu $\boldsymbol{N}_{i,c}$ a $\tilde{\boldsymbol{F}}_{i,c}(\boldsymbol{\mu}_c)$ příslušné jednotlivým komponentám c, kde $c = 1, \ldots, C$ a $\boldsymbol{\mu}_c$ je vektor středních hodnot komponenty c. Zřetězením těchto statistik je získána příslušná reprezentace v prostoru supervektorů $\boldsymbol{N}_i$ a $\tilde{\boldsymbol{F}}_i(\boldsymbol{\mu})$ (viz podkapitola 4.3.4). Na jejich základě je pak stanoven i-vektor $\boldsymbol{x}_i$ jako MAP odhad faktorů v podprostoru celkové variability v souladu s (4.61), tedy platí

$$\boldsymbol{x}_i = \left(\boldsymbol{I} + \boldsymbol{T}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_i\boldsymbol{T}\right)^{-1}\boldsymbol{T}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}_i(\boldsymbol{\mu}), \tag{8.6}$$

kde matice $\boldsymbol{T}$ definuje podprostor celkové variability (viz podkapitola 4.4). Tímto způsobem je zajištěna projekce sekvence libovolné délky na vektory konstantního rozměru.

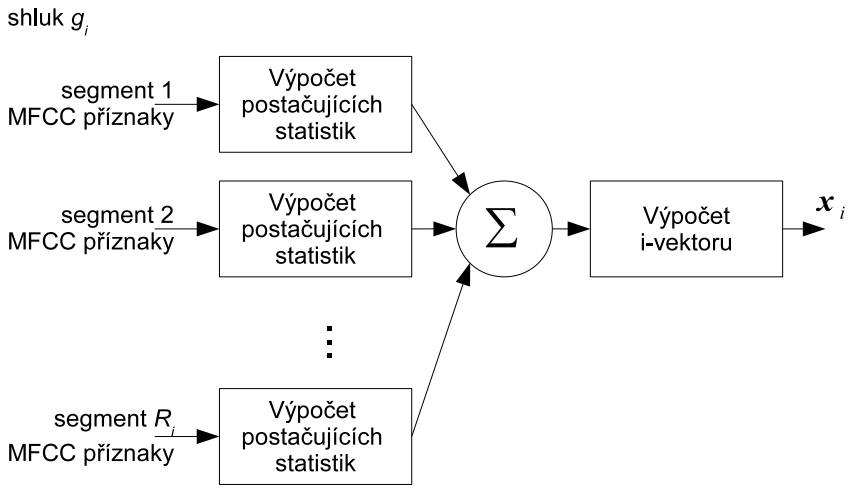
Vzájemná vzdálenost shluků g_1 a g_2 reprezentovaných i-vektory $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ je vyhodnocena na základě kosinové vzdálenosti přesně podle vztahu (4.88), tedy

$$s_{CDS}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \, \|\boldsymbol{x}_2\|}. \tag{8.7}$$

Přístup založený na aplikaci i-vektorů pro reprezentaci řečových segmentů a vyhodnocení podobnosti shluků na základě kosinové vzdálenosti byl aplikován také v (Franco-Pedroso et al., 2010), výsledky porovnání systému navrženého autory s dalšími systémy pak shrnuje (Zelenak et al., 2010). Námi navržený přístup se ale liší od přístupu navrženého v (Franco-Pedroso et al., 2010) způsobem jakým jsou reprezentovány shluky. V našem přístupu je při sloučení dvou shluků provedeno sečtení postačujících statistik příslušných všem řečovým segmentům přiřazeným v dosavadním běhu algoritmu do těchto shluků a i-vektor nového shluku stanoven na základě těchto sečtených statistik podle (8.6) (viz obr. 8.3), zatímco přístup navržený autory (Franco-Pedroso et al., 2010) reprezentuje nový shluk průměrným i-vektorem stanoveným na základě všech i-vektorů příslušných řečovým segmentům.

V průběhu shlukování je v každém cyklu algoritmu sloučen pár shluků s nejvyšší hodnotou kosinové vzdálenosti s_{CDS}. Ukončovací podmínka je splněna pokud je v daném cyklu maximální hodnota s_{CDS} nižší než práh λ stanovený na vývojových datech.

Pro potlačení variability akustických podmínek je využívána lineární diskriminační analýza (viz podkapitola 4.4.3). Cílem LDA je nalézt transformaci příznakových vektorů vedoucí k maximálnímu rozptylu mezi třídami a minimálnímu rozptylu uvnitř tříd. V našem případě tak byly při trénování transformační matice jednotlivé třídy tvořeny všemi segmenty mluvčího v rámci daného audiozáznamu.



Obrázek 8.3: *Popis způsobu odvození reprezentace shluků v případě CDS systému*

8.2.3 Shlukování s využitím PLDA modelu i-vektorů

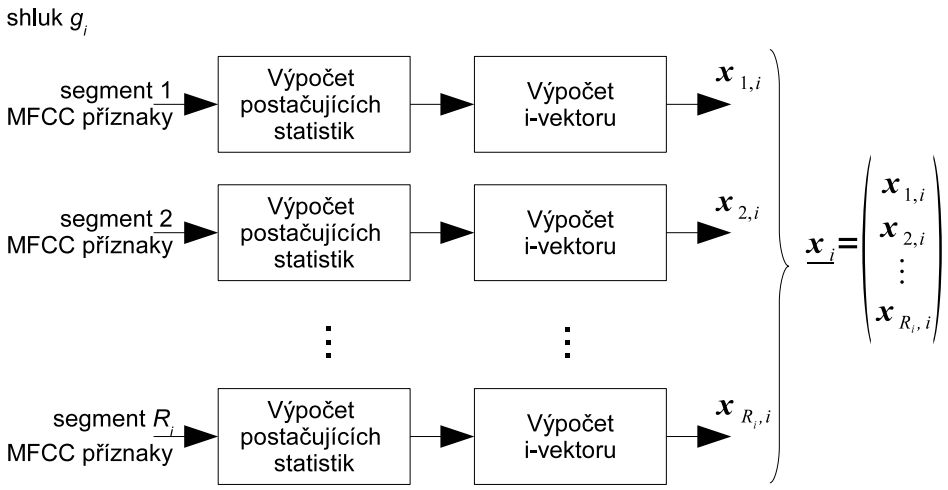
V tomto případě jsou shluky opět reprezentovány i-vektory, pro které je ovšem uvažován PLDA generativní model (4.96). Pro každý segment byly opět nejprve vyhodnoceny postačující statistiky nultého a prvního řádu pro daný UBM model a na jejich základě určen příslušný i-vektor. V případě metod založených na PLDA modelu byly i-vektory normalizovány na jednotkovou délku (Garcia-Romero – Espy-Wilson, 2011).

Míra hodnotící vzájemnou vzdálenost dvou shluků g_1 a g_2 je formulována jako logaritmus poměru věrohodností modelu $\mathcal{M}_1$ a modelu $\mathcal{M}_2$. Model $\mathcal{M}_1$ odpovídá situaci, kdy je mluvčí příslušný oběma shlukům (a tím pádem všech do nich přiřazených segmentů) totožný. Znamená to pak, že sdílejí stejné faktory specifické pro mluvčího $\boldsymbol{y}$. Zatímco model $\mathcal{M}_2$ odpovídá situaci, kdy jsou mluvčí příslušní oběma shlukům rozdílní. Mluvčí shluku g_1 (tedy všech jemu přidružených řečových segmentů) je reprezentován faktory $\boldsymbol{y}_1$ a mluvčí shluku g_2 faktory $\boldsymbol{y}_2$. Jedná se tedy o stejné modely, které jsou porovnávány v úloze verifikace mluvčích (viz podkapitola 4.5.2). V průběhu shlukování jsou v každém cyklu sloučeny shluky s nejvyšším hodnotou logaritmu poměru věrohodností stanovenou na základě (4.116). Činnost algoritmu je ukončena v případě, že je v daném cyklu nejvyšší hodnota nižší než stanovený práh λ.

Vyhodnocení věrohodnosti pro uvedené modely umožňuje, na rozdíl od vyhodnocení kosinové vzdálenosti, brát v úvahu více i-vektorů reprezentujících jednotlivé shluky. Jsou tak uvažovány dva přístupy lišící se právě v reprezentaci shluků, první je označován jako *jednosložkový PLDA* přístup a druhý jako *vícesložkový PLDA* přístup.

Jednosložkový PLDA přístup

V případě jednosložkového PLDA přístupu jsou shluky reprezentovány jediným i-vektorem. Ten je odvozen na základě součtu statistik příslušných jednotlivým řečovým segmentům přiřazeným do



Obrázek 8.4: *Popis způsobu odvození reprezentace shluků v případě systému založeného na vícesložkovém PLDA přístupu*

daného shluku, tedy stejným způsobem jako v případě CDS přístupu (viz obr. 8.3). Pro shluky g_1 a g_2 reprezentované i-vektory $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$ je logaritmus poměru věrohodností modelů $\mathcal{M}_1$ a $\mathcal{M}_2$ určen na základě (4.116) jako

$$s_{oPLDA} = \log P(\boldsymbol{x}_1, \boldsymbol{x}_2|\mathcal{M}_1) - \log P(\boldsymbol{x}_1|\mathcal{M}_2) - \log P(\boldsymbol{x}_2|\mathcal{M}_2), \tag{8.8}$$

kde první člen na pravé straně představuje logaritmus věrohodnosti kombinovaného PLDA modelu s jedním mluvčím (4.97). Hodnota tohoto členu je stanovena na základě (4.117), kde $\underline{\boldsymbol{x}}$ je vytvořeno zřetězením $\boldsymbol{x}_1$ a $\boldsymbol{x}_2$. Ostatní členy na pravé straně (8.8) odpovídají věrohodnosti jednoduchého PLDA modelu, které je také provedeno podle (4.117), v tomto případě je však $\underline{\boldsymbol{x}}$ tvořeno jediným vektorem.

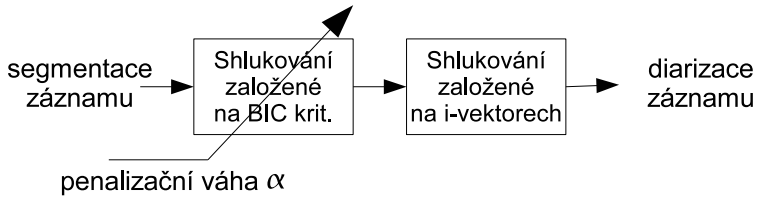
Vícesložkový PLDA přístup

Vícesložkový PLDA přístup je založen na reprezentaci shluku na základě sady všech i-vektorů příslušných řečovým segmentům přiřazeným do daného shluku. Shluk g_i je tedy reprezentován množinou R_i i-vektorů $\{\boldsymbol{x}_{1,i}, \ldots, \boldsymbol{x}_{R_i,i}\}$ jejichž zřetězením je vytvořen $R_i D_{ivec}$-dimenzionální vektor $\underline{\boldsymbol{x}}_i$ (viz obr. 8.4). Logaritmus poměru věrohodností je v tomto případě roven

$$s_{mPLDA} = \log P(\underline{\boldsymbol{x}}_1, \underline{\boldsymbol{x}}_2|\mathcal{M}_1) - \log P(\underline{\boldsymbol{x}}_1|\mathcal{M}_2) - \log P(\underline{\boldsymbol{x}}_2|\mathcal{M}_2), \tag{8.9}$$

kde pro každý z uvedených členů platí, že odpovídá logaritmu věrohodnosti kombinovaného PLDA modelu s jedním mluvčím (4.97) a vyhodnocení je tedy provedeno na základě (4.117).

Ve srovnání s jednosložkovým PLDA přístupem je výhodou, že při sloučení shluků není nutné stanovit i-vektor reprezentující nový shluk. Dochází pouze ke sloučení množin i-vektorů příslušných všem segmentům zařazeným do slučovaných shluků. Na druhé straně náročnost výpočtu (8.9) na rozdíl od náročnosti výpočtu (8.8) narůstá společně s počtem segmentů zařazených do posuzovaných shluků.



Obrázek 8.5: *Proces dvoufázového shlukování*

8.2.4 Dvoufázové shlukování

Dvoufázové shlukování již bylo úspěšně použito v systémech pro diarizaci mluvčích v záznamech jednání, přednášek nebo TVR dat, využívajících kombinaci různých technik pro shlukování například autory (Zhu et al., 2005, 2008).

Motivace pro aplikaci dvoufázového shlukování v našem případě vychází z předpokladu, že bodové odhady i-vektorů (faktorů v prostoru celkové variability) nemohou být s ohledem na velmi krátkou délku řečových segmentů stanoveny příliš robustně a to může (především v počátečním stádiu) narušit shlukovací proces. Pro zmírnění tohoto problému je shlukování rozděleno do dvou fází.

V první fázi je provedeno shlukování založené na BIC kritériu. Je však použito takové nastavení, aby došlo k přerušení činnosti dříve než by odpovídalo kompletnímu provedení shlukovacího procesu. Konkrétně je použita nulová hodnota rozhodovacího prahu λ a penalizační váha α vedoucí k nižší úrovni shlukování segmentů (zachování vyššího počtu shluků). Čím nižší je hodnota α, tím více jsou penalizovány dlouhé segmenty. V důsledku toho je první fáze dříve přesušena a na vstupu druhé fáze je větší počet shluků. Hlavním cílem první fáze je zredukovat počet shluků, kterým je přiřazeno malé množství dat. Ve druhé fázi je aplikováno shlukování využívající jednu z metod založených na i-vektorech. Schéma dvoufázového shlukování ilustruje obr. 8.5.

8.3 Specifikace evaluační databáze a vyhodnocení systémů

Experimentální vyhodnocení bylo provedeno s využitím databáze COST278 (Vandecatseye et al., 2004). Tato databáze zahrnuje záznamy televizních zpravodajských pořadů v 9 evropských jazycích. Autoři databáze rozdělili data pro každý jazyk do dvou sad – trénovací, obsahující kolem 2 hod. dat, a testovací, obsahující přibližně 1 hod. dat.

Pro účely našeho vyhodnocení byly definovány tři různé sady dat. První sada obsahovala všechna chorvatská, česká, maďarská, portugalská a slovenská data označená v COST278 databázi jako trénovací. Celkové délka těchto dat je 11,5 hod. Tato sada byla použita pro trénování UBM modelu, stanovení prostoru celkové variability a odhad hyperparametrů PLDA modelu.

Druhá sada obsahovala záznamy 13 pořadů různé délky (v rozsahu od 8,5 do 53,8 min.), vybrané také z trénovacích dat databáze COST278. Celkový objem dat v této sadě je 5,9 hod. a tato sada byla použita jako vývojová. Na základě této sady tak byla hledána optimální nastavení systémů, např. práh pro segmentaci mluvčích nebo práh ukončovací podmínky shlukování.

Konečně třetí sada byla použita pro vyhodnocení systémů. Tato sada obsahovala záznamy 15 pořadů různé délky (v rozsahu od 4,1 do 53,2 min.), vybrané z testovacích dat databáze COST278. Celková délka těchto záznamů je 6,3 hod. Výběr dat pro vývojovou a testovací sadu byl omezen na holandská, česká, maďarská, slovinská a slovenská data.

Záznamy některých pořadů v databázi COST278 obsahují také reklamní bloky, které ale nejsou v referenčních přepisech anotovány. Příslušné intervaly byly proto z nahrávek odstraněny.

8.3.1 Evaluační metriky

Systémy pro diarizaci mluvčích jsou obvykle porovnávány na základě míry DER (Diarization Error Rate), která byla definována úřadem NIST (NIST, 2009). Míru DER je možné vyjádřit na základě součtu třech složek

$$DER = SPKE + FA + MISS, \tag{8.10}$$

kde SPKE (speaker error rate) reprezentuje míru chybného přiřazení dat mezi mluvčími při optimálním spárování referenčních mluvčích a mluvčích ve výstupu diarizačního systému. Míra FA odpovídá objemu neřečových segmentů vyhodnocených jako řečových a míra MISS odpovídá objemu řečových segmentů vyhodnocených jako neřečových. Tyto míry jsou v našem případě závislé pouze na úspěšnosti detektoru řečové aktivity, protože dále již neprobíhá žádná změna klasifikace příznakových vektorů jako řečových nebo neřečových.

Protože všechny zde vyhodnocované systémy sdílejí shodný modul pro detekci řečové aktivity a segmentaci mluvčích, je jako primární metrika použita míra SPKE. V souladu s (NIST, 2009) byl pro každou hranici použit toleranční interval délky 0,25 s (v obou směrech).

8.4 Experimentální vyhodnocení systémů

8.4.1 Využití trénovacích dat

Průměrná délka segmentů v trénovacích datech, tj. ručně anotovaných, je 17,8 s. UBM model byl natrénován na datech 1 007 mluvčích, zahrnujících 2 530 segmentů o celkové délce 11,5 hod. Byly natrénovány modely různé velikosti až po model s 1024 komponentami. Matice charakterizující prostor celkové variability $\boldsymbol{T}$ byla stanovena na základě podmnožiny trénovacích dat, která vznikla výběrem segmentů o minimální délce 3 s a s omezením použití maximálně 8 segmentů od jednoho mluvčího. Výsledná sada obsahovala 2 050 segmentů (10,2 hod.) od 909 mluvčích. Hyperparametry PLDA modelu, tj. matice $\boldsymbol{V}$ a $\boldsymbol{U}$, byly stanoveny (společným odhadem) na základě dat od mluvčích, pro které byly k dispozici alespoň 3 segmenty o minimální délce 3 s. Celkem bylo použito 1 528 segmentů (7,5 hod.) od 280 mluvčích. Tato sada byla použita i pro stanovení LDA transformačních matic.

Při trénování všech modelů založených na faktorové analýze byl použit EM algoritmus aplikující v každé iteraci kombinovaný ML-MD odhad parametrů popsaný v podkapitole 4.5.1.

8.4.2 Parametrizace řečového signálu

Všechny moduly využívají MFCC akustické příznaky. Signál byl parametrizován každých 10 ms oknem délky 25 ms. Detektor řečové aktivity využívající GMM modely používá 39-rozměrné příznakové vektory, ty jsou tvořeny 12 MFCC statickými příznaky a koeficientem c_0, rozšířenými o dynamické příznaky prvního a druhého řádu. Příznakové vektory byly normalizovány metodou odečítání kepstrálního průměru (CMS). Normalizace byla aplikována *lokálně* pro prostřední vektor posuvného okna délky 400 vektorů. Naše zkušenost je, že lokální aplikace CMS přináší lepší výsledky než odečítání globální střední hodnoty. Pokud je tedy dále v této kapitole uváděno použití CMS normalizace, jedná se vždy o její lokální aplikaci.

Moduly pro segmentaci mluvčích a shlukování využívají menší příznakové vektory tvořené pouze 12 MFCC statickými příznaky. V rámci segmentace nebyla aplikována CMS normalizace a efekt její aplikace v rámci shlukování bude diskutován zvlášť pro jednotlivé přístupy.

8.4.3 Vyhodnocení činnosti detektorů řeči a změny mluvčích

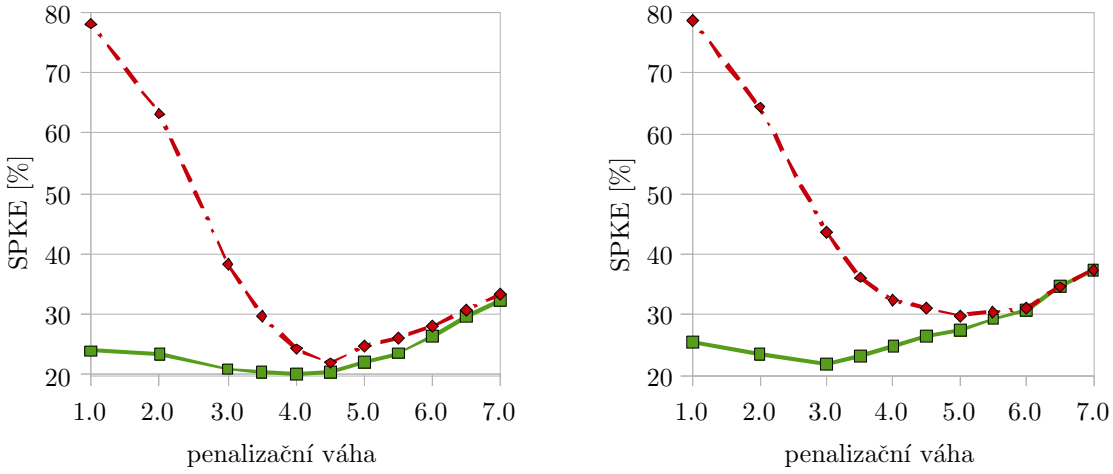
Jak již bylo uvedeno, modul detekce řečové aktivity a segmentace mluvčích je společný pro všechny systémy. Na testovací sadě byly vyhodnoceny míry FA 0,8 % a MISS 3,2 %. Vyšší míra MISS je způsobena nepřesností referenčních anotací, kdy pro některé záznamy nebyly anotovány dlouhé pauzy v řeči jako neřečové události. Kontrolou výstupu detektoru řečové aktivity bylo zjištěno, že detektor byl schopen spolehlivě určovat hranice řečových segmentů. Modul pro segmentaci mluvčích byl nastaven tak, že připouští vyšší počet falešných detekcí změny mluvčího. Důvodem je, že zatímco chyba způsobená falešnou detekcí může být (a zpravidla je) napravena v průběhu shlukování, chyba způsobená opominutou detekcí již nemůže být v žádném dalším kroku napravena (Prazak – Silovsky, 2011). Výsledkem jsou však poměrně krátké segmenty. Průměrná délka segmentů je 3,6 s (zatímco u referenčních anotací je to 17,8 s).

8.4.4 Referenční systém založený na BIC kritériu

Referenční systém využívá shlukování založené na BIC kritériu. V souvislosti s tím byl nejprve analyzován vliv penalizační váhy α, která je volitelným parametrem kritéria ΔBIC (8.4). Na základě provedených experimentů bylo zjištěno, že systémy, pro které byla použita nenulová hodnota ukončovacího prahu λ, stanovená na základě vývojových dat, dosahují lepších výsledků než systémy s nulovou hodnotou prahu. Rozdíl je patrný především v případě systémů pracujících s nízkou hodnotou penalizační váhy α jak ilustruje obr. 8.6.

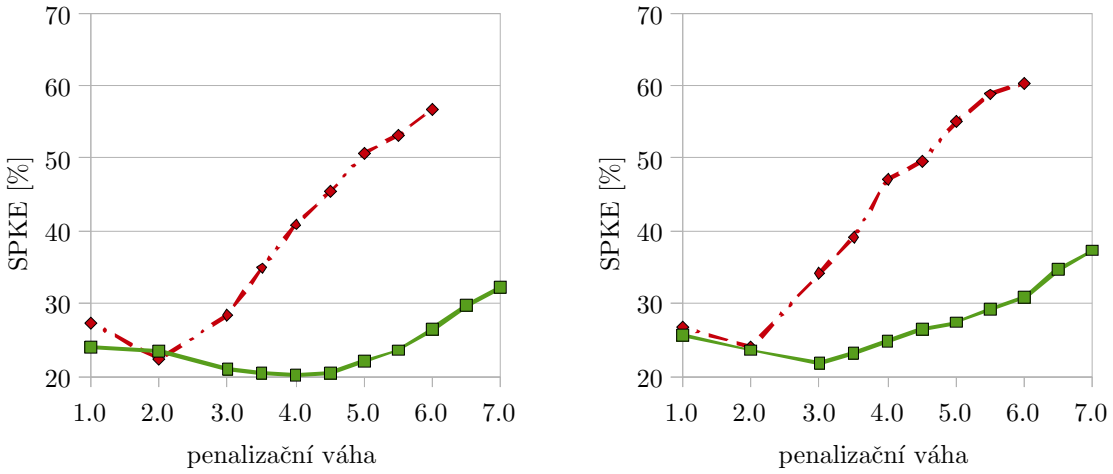
Nejlepší výsledky na vývojové sadě byly dosaženy systémem s penalizační vahou $\alpha = 4{,}0$. Výsledky dosažené pro toto nastavení na testovací sadě jsou tak považovány za referenční při vyhodnocení ostatních systémů. Systém dosáhl na testovací sadě míry SPKE 24,8 %, které odpovídá DER 28,8 %.

Dále byl zkoumán vliv aplikace CMS normalizace. Bylo přitom zjištěno, že aplikace CMS nor-



(a) vývojová sada (b) testovací sada

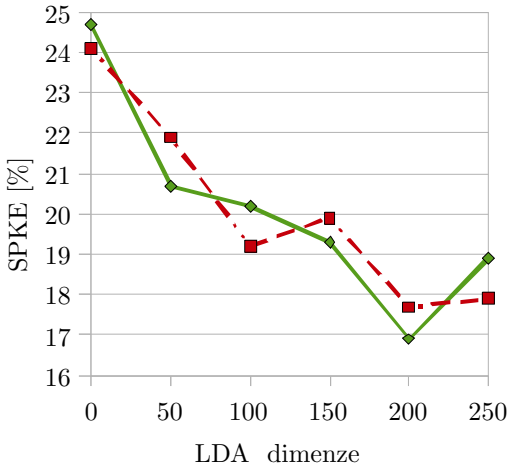
Obrázek 8.6: *Vliv hodnoty penalizační váhy α v případě referenčního systému pracujícího s nulovou (červená čára) a optimalizovanou (zelená čára) hodnotou ukončovacího prahu λ při shlukování vyhodnocený na (a) vývojové a (b) testovací sadě*



(a) vývojová sada (b) testovací sada

Obrázek 8.7: *Porovnání výsledků s (červená čára) a bez (zelená čára) aplikace CMS normalizace pro referenční systém využívající shlukování založené na BIC kritériu. Systémy využívají optimalizovanou hodnotu ukončovacího prahu. Vyhodnocení provedeno na (a) vývojové a (b) testovací sadě*

malizace způsobuje výrazné zhoršení míry SPKE, zejména pro vyšší hodnoty penalizační váhy α. Obr. 8.7 dokládá efekt aplikace CMS normalizace v případě systému využívajícího optimalizovanou hodnotu ukončovacího prahu λ.



Obrázek 8.8: *Efekt dimenze LDA transformace v případě CDS systémů pracujících s 300- (červená čára) a 400-dimenzionálními (zelená čára) i-vektory*

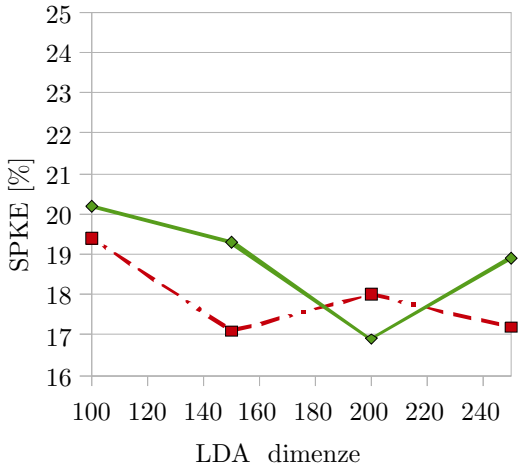
8.4.5 Výsledky systému založeného na CDS přístupu

V případě systému založeného na CDS přístupu byla zkoumána různá nastavení lišící se ve velikosti použitého UBM modelu, rozměru prostoru celkové variability (tj. rozměru i-vektorů) a rozměru LDA transformační matice. Nejlepší výsledky byly dosaženy pro systémy využívající UBM model s 256 komponentami a tedy 3 072-dimenzionální supervektory. Obr. 8.8 ilustruje vliv dimenze LDA transformace pro systémy pracující s 300 a 400-dimenzionálními i-vektory. Nulová dimenze LDA v obr. 8.8 odpovídá systému bez aplikace LDA transformace. S ohledem na omezené množství mluvčích, pro které je k dispozici dostatečný počet řečových segmentů, v trénovací databázi je možné stanovit nanejvýš 279-dimenzionální LDA transformaci.

Bez aplikace LDA poskytuje systém založený na CDS přístupu výsledky srovnatelné s referenčním systémem. Míra SPKE 24,1 % dosažená pro systém pracující s 300-dimenzionálními i-vektory představuje ve srovnání s hodnotou 24,8 %, dosaženou referenčním systémem, relativní zlepšení o 2,8 %. Odstranění nežádoucí variability prostřednictvím LDA však přináší významné zlepšení úspěšnosti. Nejnižší hodnota SPKE 16,9 % byla dosažena systémem pracujícím s 400-dimenzionálními i-vektory a 200-dimenzionální LDA transformací.

Obr. 8.9 zobrazuje efekt aplikace CMS normalizace pro systém pracující s 400-dimenzionálními i-vektory. Upozorněme, že aplikace CMS v případě systémů založených na CDS přístupu vyžaduje přetrénování UBM modelu, prostoru celkové variability a LDA transformace na normalizovaných datech. Na rozdíl od systému založeného na BIC shlukování nebyl pozorován výrazný efekt aplikace CMS. Pro většinu vyhodnocených nastavení (včetně těch, které nejsou zahrnuty v obr. 8.9) přinesla aplikace CMS normalizace mírné snížení SPKE.

Dalšího snížení míry SPKE bylo dosaženo použitím dvoufázového shlukování, které navíc vede ke snížení celkové výpočetní náročnosti shlukovacího procesu. Přístup založený na BIC kritériu je

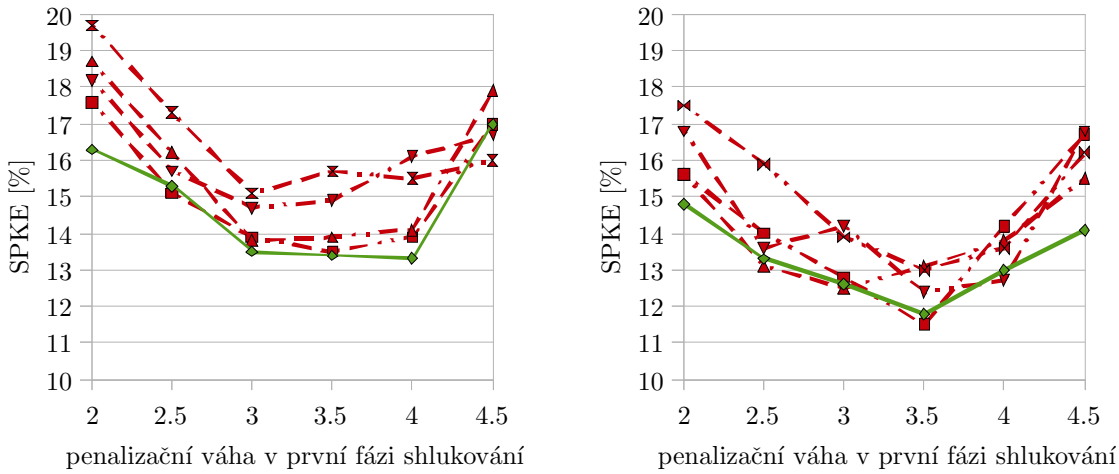


Obrázek 8.9: *Efekt aplikace CMS normalizace (červená čára) pro CDS systém pracující s 400-dimenzionálními i-vektory*

totiž ve srovnání s přístupy založenými na i-vektorech méně výpočetně náročný a navíc náročnost přístupů založených na i-vektorech narůstá v závislosti na počtu vstupních segmentů (shluků) rychleji. V důsledku snížení počtu shluků provedenému v první rychlé fázi (založené na BIC kritériu), je u dvoufázového shlukování dosaženo podstatné redukce výpočetní náročnosti. V případě první fáze byla pro shlukování založené na BIC kritériu použita nulová hodnota ukončovacího prahu λ. Počet shluků, které vstupují do druhé fáze, tak ovlivňuje hodnota penalizační váhy α. Čím nižší je tato hodnota, tím dříve je první fáze přerušena a tím více shluků vstupuje do druhé fáze. Obr. 8.10 dokládá vliv hodnoty penalizační váhy α použité v první fázi pro různé konfigurace CDS přístupu aplikované v druhé fázi dvoufázového shlukování. Nejlepší konfigurace CDS přístupu je zvýrazněna plnou čarou.

V případě systémů, které nepoužívají CMS normalizaci ani ve druhé fázi shlukování (viz obr. 8.10(a)), bylo dosaženo nejlepších výsledků pro penalizační váhy v intervalu od 3,0 do 4,0. V případě systémů, které aplikují CMS normalizaci ve druhé fázi shlukování (viz obr. 8.10(b)), bylo u většiny systémů dosaženo nejlepších výsledků pro penalizační váhu 3,5. Zdá se tak, že u systémů aplikujících CMS normalizaci je optimální hodnota penalizační váhy α snáze určitelná. Aplikace CMS normalizace ve druhé fázi shlukování ale především přináší snížení míry SPKE. S ohledem na výsledky dosažené pro různé hodnoty penalizační váhy α, lze jako nejlepší hodnotit systém pracující s 400-dimenzionálními i-vektory a 200-dimenzionální LDA transformací. Ten při hodnotě $\alpha = 3{,}5$ dosáhl SPKE 11,8 % (představující 52,4% relativní zlepšení oproti referenčním výsledkům).

Naše vysvětlení pro rozdílný přínos CMS normalizace v případě shlukování založeného pouze na CDS přístupu a dvoufázového shlukování je následující. První fáze využívající shlukování založené na BIC kritériu je méně náchylná k provedení chyby v důsledku malého množství dat (krátké délky segmentů). Navíc není v první fázi aplikována CMS normalizace a tak lze předpokládat, že v rámci této fáze budou slučovány přednostně sousední segmenty, tím by v ideálním případě mělo dojít k eliminaci falešných detekcí změny mluvčího. Tímto je pro druhou fázi zredukováno množství



(a) bez aplikace CMS

(b) s aplikací CMS

Obrázek 8.10: *Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování pro různá nastavení aplikovaná ve druhé fázi založené na CDS přístupu. Zelená čára vyznačuje výsledky dosažené pro nejlepší konfiguraci CDS přístupu. Obr. (a) ilustruje výsledky dosažené bez aplikace CMS normalizace ve druhé fázi, obr. (b) pak s aplikací CMS normalizace*

shluků, pro které je dostupné pouze malé množství dat. Pro shluky, pro které je k dispozici více dat, je pak možné získat přesnější MAP odhady i-vektorů. Aplikace CMS normalizace ve druhé fázi navíc usnadňuje shlukování vzdálených segmentů (resp. jim příslušných shluků) pocházejících od stejného mluvčího, ale zaznamenaných za různých podmínek.

8.4.6 Výsledky systému založeného na jednosložkovém PLDA přístupu

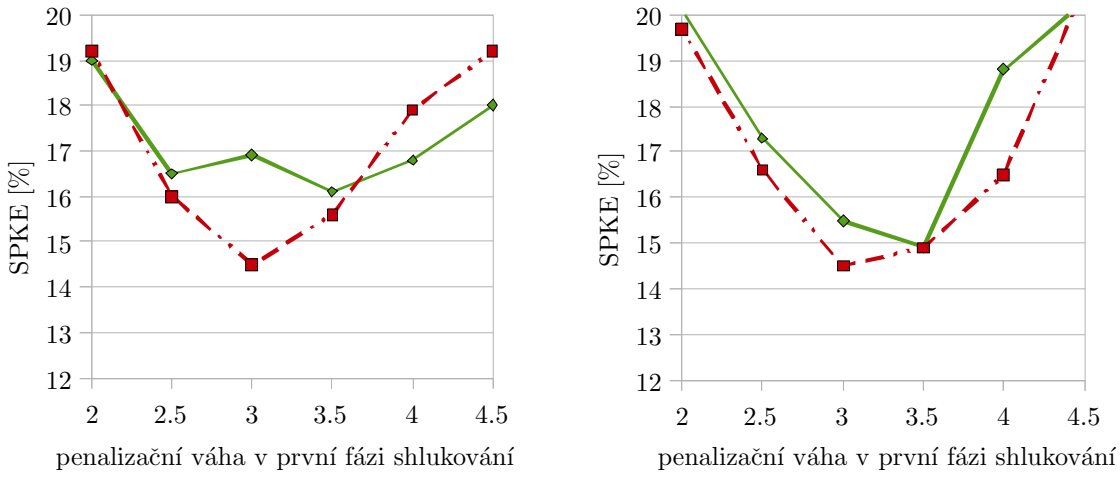
Stejně jako v případě přístupu založeného na použití kosinové vzdálenosti byl nejprve zkoumán vliv počtu Gaussovských komponent UBM modelu, rozměru podprostoru celkové variability a hodností matic $\boldsymbol{V}$ a $\boldsymbol{U}$ PLDA modelu. Tab. 8.1 shrnuje vybrané výsledky dosažené v rámci testovaných nastavení. Ve všech případech byl použit UBM model s 256 komponentami. Přestože byla testována i nastavení s nesouhlasnými hodnostmi matic $\boldsymbol{V}$ a $\boldsymbol{U}$, v případě souhlasné hodnosti bylo vždy dosaženo lepších výsledků. Systém pracující s 300-dimenzionálními i-vektory a hodností $\boldsymbol{V}$ a $\boldsymbol{U}$ rovnou 150 dosáhl míry SPKE 21,8 % (odpovídající 12,1% rel. zlepšení). Aplikace CMS normalizace přinesla v případě tohoto systému mírné zlepšení míry SPKE na 20,9 %. Přínos aplikace CMS normalizace v případě jednosložkového PLDA přístupu však není zcela jednoznačný, jak je patrné z výsledků uvedených v tab. 8.1.

Výrazného snížení SPKE bylo dosaženo v případě dvoufázového shlukování, jak dokládá obr. 8.11. Nejnižší míry SPKE 14,5 % (41,5% rel. zlepšení) dosáhl systém s 300-dimenzionálními i-vektory (rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150) v případě použití penalizační váhy ΔBIC kritéria $\alpha = 3{,}0$

Tabulka 8.1: *Výsledky systémů využívajících jednosložkové PLDA shlukování*

			bez aplikace CMS		s aplikací CMS	
rk($\boldsymbol{T}$)[a]	rk($\boldsymbol{V}$)	rk($\boldsymbol{U}$)	SPKE [%]	rel. změna [%]	SPKE [%]	rel. změna [%]
300	100	100	22,5	9,3	22,7	8,5
300	150	150	21,8	12,1	20,9	15,7
400	200	200	23,0	7,3	25,5	-2,8

[a] rk($\boldsymbol{T}$) = D_{ivec}



(a) bez aplikace CMS

(b) s aplikací CMS

Obrázek 8.11: *(a) Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování pro dvě konfigurace jednosložkového PLDA přístupu aplikovaného ve druhé fázi. Červená čára odpovídá systému s 300-dimenzionálními i-vektory (*rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150*) a zelená čára systému s 400-dimenzionálními i-vektory (*rk($\boldsymbol{V}$) = 200,rk($\boldsymbol{U}$) = 200*). (b) Vyhodnocení pro shodná nastavení s aplikací CMS normalizace ve druhé fázi shlukování*

v první fázi shlukování. Z porovnání obr. 8.11(a) a 8.11(b) je zřejmé, že aplikace CMS normalizace ve druhé fázi shlukování založené na jednosložkovém PLDA přístupu přináší ve většině případů pouze nepatrné snížení míry SPKE a celkový vliv lze hodnotit spíše jako zanedbatelný. Možným vysvětlením je, že variabilita akustických podmínek je v tomto případě plně kompenzována vlastním PLDA modelem a efekt CMS normalizace je tak potlačen. Úspěšnost nejlepšího uvedeného systému zůstala při aplikaci CMS normalizace nezměněna.

Tabulka 8.2: *Výsledky systémů využívajících vícesložkové PLDA shlukování*

			bez aplikace CMS		s aplikací CMS	
rk($\boldsymbol{T}$)	rk($\boldsymbol{V}$)	rk($\boldsymbol{U}$)	SPKE [%]	rel. změna [%]	SPKE [%]	rel. změna [%]
300	150	150	17,2	30,6	14,3	42,3
400	200	200	15,9	35,9	14,8	40,3

8.4.7 Výsledky systému založeného na vícesložkovém PLDA přístupu

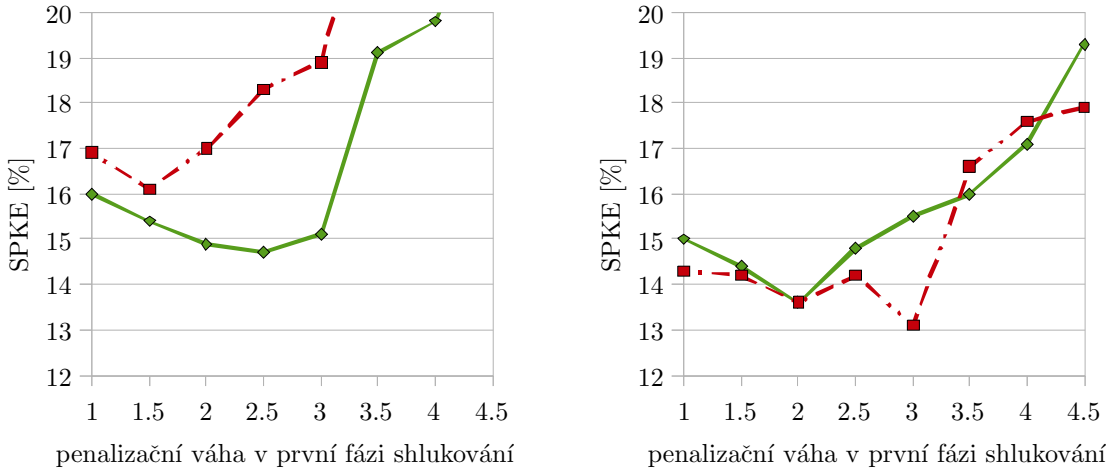
Také v případě systému využívajícího shlukování založené na vícesložkovém PLDA přístupu byl nejprve analyzován vliv počtu komponent UBM, rozměru podprostoru celkové variability a hodností matic $\boldsymbol{V}$ a $\boldsymbol{U}$ PLDA modelu. Nejlepších výsledků v tomto vyhodnocení dosáhly opět systémy používající UBM s 256 komponentami a souhlasnou hodnost matic $\boldsymbol{V}$ a $\boldsymbol{U}$. Výsledky dosažené pro nejlepší konfigurace systémů pracujících s dvěma rozdílnými rozměry i-vektorů jsou uvedeny v tab. 8.2. Systém pracující s 400-dimenzionálními i-vektory (rk($\boldsymbol{V}$) = 200,rk($\boldsymbol{U}$) = 200) dosáhl míry SPKE 15,9 % (35,9% rel. zlepšení). Po aplikaci CMS normalizace došlo k dalšímu snížení míry SPKE na 14,8 %. Ještě výraznější přínos CMS normalizace však byl zaznamenán v případě systému pracujícího s 300-dimenzionálními i-vektory (rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150). V tomto případě došlo ke snížení míry SPKE na 14,3 % (42,3% rel. zlepšení). To představuje nejlepší dosažený výsledek při použití jednofázového shlukování.

Výsledky dosažené aplikací dvoufázového shlukování ilustruje obr. 8.12. Ve srovnání s předešlými přístupy je v tomto případě dosaženo pouze mírného zlepšení míry SPKE, zejména pokud není ve druhé fázi provedena CMS normalizace (viz 8.12(a)). Výhodou dvoufázového shlukování je ovšem nižší výpočetní náročnost. Z obr. 8.12(a) je patrné, že optimální hodnoty penalizační váhy ΔBIC kritéria aplikovaného v první fázi dvoufázového shlukování jsou při použití vícesložkového PLDA shlukování ve druhé fázi nižší než v případě aplikace ostatních přístupů (k přerušení první fáze tak dochází dříve). Pro systém pracující s 300-dimenzionálními i-vektory (rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150) dochází k rychlému zvýšení SPKE společně s růstem penalizační váhy již nad hodnotu 1,5. Nejnižší míry SPKE 14,7 % (40,7% rel. zlepšení) dosáhl systém s 400-dimenzionálními i-vektory (rk($\boldsymbol{V}$) = 200,rk($\boldsymbol{U}$) = 200) a penalizační vahou ΔBIC kritéria $\alpha = 2{,}5$.

Aplikace CMS normalizace ve druhé fázi shlukování přináší patrné zlepšení úspěšnosti, viz obr. 8.12(b). Zejména pro systém s 300-dimenzionálními i-vektory, který pro penalizační váhu $\alpha =$ 3,0 dosáhl míry SPKE 13,1 % (47,2% rel. zlepšení).

8.5 Shrnutí výsledků

Tab. 8.3 zprostředkovává přehled nejlepších výsledků dosažených pro prezentované přístupy při aplikaci jednofázového shlukování včetně výpočetní náročnosti uváděné jako násobek reálné délky



(a) bez aplikace CMS

(b) s aplikací CMS

Obrázek 8.12: *(a) Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování pro dvě konfigurace vícesložkového PLDA přístupu aplikovaného ve druhé fázi. Červená čára odpovídá systému s 300-dimenzionálními i-vektory* ($\mathrm{rk}(\boldsymbol{V}) = 150, \mathrm{rk}(\boldsymbol{U}) = 150$) *a zelená čára systému s 400-dimenzionálními i-vektory* ($\mathrm{rk}(\boldsymbol{V}) = 200, \mathrm{rk}(\boldsymbol{U}) = 200$). *(b) Vyhodnocení pro stejné systémy s aplikací CMS normalizace ve druhé fázi shlukování*

zpracovávaných záznamů[2]. Výpočetní náročnost se pro jednotlivé nahrávky v testovací sadě liší. Je ovlivněna především počtem segmentů, které jsou výstupem segmentačního modulu. Tento počet přirozeně roste s počtem mluvčích v nahrávce a s ohledem na charakter zpracovávaných pořadů obecně roste s délkou nahrávky. Uváděná hodnota představuje poměr celkové délky zpracování[3] všech nahrávek a součtu jejich délky.

Pro všechny přístupy založené na reprezentaci segmentů pomocí i-vektorů byly dosaženy nejlepší výsledky při použití UBM modelu s 256 komponentami. Z hlediska úspěšnosti dosáhl nejlepších výsledků jednoznačně systém založený na vícesložkovém PLDA přístupu. Oproti referenčnímu systému založenému na BIC kritériu došlo k 42,3% relativnímu snížení míry SPKE. Nevýhodou tohoto přístupu je však nejvyšší výpočetní náročnost, která je téměř 6krát vyšší ve srovnání s referenčním systémem. S ohledem na shodnou reprezentaci shluků pomocí jediného i-vektoru (odvozeného na základě statistik sečtených přes všechny segmenty zařazené do daného shluku) u systémů založených na CDS a jednosložkovém PLDA přístupu, je zajímavý poměrně velký rozdíl nejnižší dosažené míry SPKE u těchto systémů. Systém založený na CDS přístupu má navíc vedle vyšší úspěšnosti

[2]Výpočetní čas odpovídá činnosti jednoho jádra systému s procesorem Intel Core i7 920@2,66 GHz a 3 GB RAM (DDR3@1,6 GHz).

[3]Je brána v úvahu pouze doba příslušná shlukování, činnost ostatních modulů (detekce řeči a změny mluvčího) není zahrnuta.

Tabulka 8.3: *Shrnutí výsledků dosažených při využití prezentovaných přístupů založených na i-vektorech v rámci jednofázového shlukování*

systém	konfigurace	CMS	SPKE [%]	Δ SPKE [%]	x RT
BIC	$\alpha = 4{,}0$ (referenční systém)[a]	ne	24,8	-	0,04
CDS	rk($\boldsymbol{T}$) = 400, LDA dimenze 200	ne	16,9	31,9	0,07
jednoslož. PLDA	rk($\boldsymbol{T}$) = 300,rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150	ano	20,9	15,7	0,13
víceslož. PLDA	rk($\boldsymbol{T}$) = 300,rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150	ano	14,3	42,3	0,23

[a] Byla použita nenulová hodnota prahu ukončovací podmínky optimalizovaná na vývojových datech

také nižší výpočetní náročnost.

Velmi rozdílný byl, v závislosti na použitém přístupu, efekt aplikace CMS normalizace. Zatímco v případě systému založeného na BIC kritériu dochází v souvislosti s aplikací CMS pro hodnoty penalizační váhy ΔBIC kritéria vyšší než 2,0 k prudkému zhoršení úspěšnosti. V případě systému využívajícího vícesložkový PLDA přístup přinesla aplikace CMS podstatné zvýšení úspěšnosti. U ostatních přístupů nedošlo k významnému ovlivnění úspěšnosti v důsledku aplikace CMS normalizace.

Jednoznačně pozitivní přínos měla aplikace dvoufázového shlukování, kdy je v první fázi provedeno „předshlukování“ založené na BIC kritériu. Vybrané výsledky shrnuje tab. 8.4. Ukončovací podmínka první fáze je splněna v okamžiku, kdy nejnižší hodnota kritéria ΔBIC pro libovolný pár shluků přesáhne nulový práh λ. Jediným parametrem ovlivňujícím okamžik přerušení první fáze je tak hodnota penalizační váhy α. Čím je tato hodnota nižší, tím více shluků je na vstupu do druhé fáze a s tím roste i celková výpočetní náročnost shlukování, protože přístup založený na BIC kritériu je výpočetně nejméně náročný. Pro lepší ilustraci vlivu penalizační váhy ΔBIC kritéria na celkovou výpočetní náročnost v rámci dvoufázového shlukování jsou v tab. 8.4 uvedeny výsledky systému založeného na vícesložkovém PLDA přístupu pro různé hodnoty α. Volbu α však nelze provádět s ohledem na požadovanou výpočetní náročnost ale vzhledem k vlivu na úspěšnost diarizace. Nevhodná volba α použitá v rámci dvoufázového shlukování (zejména příliš vysoká hodnota vedoucí k pozdnímu přerušení první fáze) může vést ke zhoršení výsledků ve srovnání s výsledky dosaženými v rámci jednofázového shlukování. V případě systémů založených na CDS a jednosložkovém PLDA přístupu je vhodnou volbou hodnota α v intervalu od 3,0 do 3,5. V případě vícesložkového PLDA přístupu je přínos dosažený aplikací dvoufázového shlukování nejnižší a tomu odpovídá i preference nižších hodnot α (viz obr. 8.12), která vede k přenechání více činnosti na druhou fázi.

V případě dvoufázového shlukování vedla aplikace CMS normalizace ve druhé fázi shlukování pro všechny prezentované přístupy ke snížení míry SPKE. Systém využívající CDS přístup dosáhl míry SPKE 11,8 % (52,4% rel. zlepšení), což představuje vůbec nejlepší dosažený výsledek. Výhodou tohoto systému je také nízká výpočetní náročnost. Doba potřebná pro zpracování nahrávek odpovídá pouze 0,06 násobku jejich délky.

Tabulka 8.4: *Shrnutí výsledků dosažených v rámci dvoufázového shlukování*

systém	konfigurace[a]	BIC α[b]	SPKE [%]	Δ SPKE [%]	x RT
CDS	rk($\boldsymbol{T}$) = 400, LDA dimenze 200	3,5	11,8	52,4	0,06
jednoslož. PLDA	rk($\boldsymbol{T}$) = 300,rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150	3,0	14,5	41,5	0,07
víceslož. PLDA	rk($\boldsymbol{T}$) = 300,rk($\boldsymbol{V}$) = 150,rk($\boldsymbol{U}$) = 150	1,0	14,3	42,3	0,17
		2,0	13,6	45,2	0,14
		3,0	13,1	47,2	0,11

[a] Ve všech případech byla ve druhé fázi shlukování aplikována CMS normalizace

[b] Penalizační váha ΔBIC kritéria α aplikovaná v první fázi shlukování

Kapitola 9

Závěr

Výzkum v oblasti zabývající se otázkami rozpoznávání mluvčího prochází v posledních letech nebývalým rozvojem. Důvodů je hned několik: V prvé řadě roste počet a důležitost aplikací, kde lze výsledky bezprostředně uplatnit, počínaje bezpečnostními úlohami, přes nasazení v různých oblastech využívající biometrii, až třeba po projekty zaměřené na zpracování archivů mluvené řeči. Dalším činitelem rychlého rozvoje je existence pravidelných mezinárodních evaluací, které stimulují vývoj nových metod a umožňují jejich bezprostřední a objektivní vyhodnocování. Významnou roli hraje i to, že se jedná o jeden z mála podoborů v oblasti zpracování řeči, který je nezávislý na jazyku, a výzkum zde tedy není tříštěn specifickými jazykovými a národními aspekty. Jedním z hlavních cílů této práce bylo proto podchytit a podrobně zmapovat současný stav oboru a jednotným způsobem podat souhrnný výklad všech moderních metod a přístupů k textově nezávislému rozpoznávání mluvčích, včetně navazující úlohy, kterou je diarizace mluvčích.

Významným, často však opomíjeným, tématem souvisejícím s vývojem a aplikací systémů pro rozpoznávání mluvčích je vyhodnocování těchto systémů. Tradičně bývá vyhodnocení prováděno na základě DET charakteristiky nebo hodnoty EER. Část práce věnovaná problematice vyhodnocování systémů se kromě těchto základních metod, které provádí vyhodnocení pouze rozlišovacích schopností systémů (ve smyslu schopnosti rozlišovat oprávněné a neoprávněné soudy), zabývá popisem řady pokročilých metod umožňujících mimo jiné provedení tzv. aplikačně nezávislého vyhodnocení, nebo vyhodnocení vhodného z pohledu interpretovatelnosti v případě forenzních aplikací. V souvislosti s rostoucím zájmem o použití technologie automatického rozpoznávání mluvčích bezpečnostními složkami a v oblasti justice je potřeba uceleného výkladu těchto metod velmi aktuální a jeho provedení v rámci této práce přínosné. Uvedené metody vyhodnocení jsou navíc aplikovatelné i v dalších úlohách, které jsou formulovány jako úloha dichotomie, případně v úlohách, které mohou být reformulovány jako úloha klasifikace do dvou tříd obdobným způsobem, jako je možné převést úlohu identifikace mluvčích na verifikační úlohu.

Navzdory velké oblibě metod založených na faktorové analýze je v současné době obtížné nalézt v odborné literatuře souhrnný přehled umožňující rychlou orientaci v těchto metodách i zájemcům, kteří se problematice rozpoznávání mluvčích dlouhodobě nevěnují. Kapitola věnovaná generativním

klasifikátorům se proto zabývá poměrně detailním výkladem těchto metod, zahrnujícím popis JFA modelu a významu jeho složek, odvození i-vektorové reprezentace nebo popis metody PLDA. Popis jednotlivých metod přitom zahrnuje veškeré vztahy nutné pro jejich implementaci.

Vývoj systémů pro rozpoznávání mluvčích je v posledních letech ovlivňován především pravidelnými evaluacemi pořádanými americkým Úřadem pro standardy a technologii. Autor se zúčastnil dvou posledních ročníků NIST SRE evaluací pořádaných v letech 2008 a 2010. Účast v roce 2008 přitom představovala první zapojení Laboratoře počítačového zpracování řeči TUL v sérii těchto evaluací. Standardní evaluační data jsou sice ideálním prostředkem pro vyhodnocení úspěšnosti nově navržených metod, na druhé straně existují reálné aplikace, které pracují s daty charakterem vzdálenými od dat zahrnutých v evaluacích pořádaných NIST.

Jedním z příkladů takové aplikace je rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů, které je vyžadováno řadou řešení vyvíjených Laboratoří počítačového zpracování řeči TUL. V rámci této práce byly shrnuty výsledky experimentálního vyhodnocení několika metod v této úloze. V případě metod založených na stanovení modelu mluvčího relevančním MAP odhadem parametrů byl, v rozporu s často uváděným tvrzením, pozorován výrazný vliv hodnoty relevančního faktoru. To může být jedním z důsledků rozdílného množství dat dostupného pro trénování modelů jednotlivých mluvčích a rozpoznávání, které je přirozené pro reálné úlohy, ne však pro standardní evaluační databáze. Výsledky vyhodnocení provedených na těchto databázích tak nelze zcela automaticky přejímat při nasazení systémů v praktických aplikacích.

V rámci řešení úlohy diarizace mluvčích byly navrženy tři alternativní přístupy pro shlukování mluvčích, které využívají reprezentaci řečových segmentů pomocí i-vektorů. Jejich aplikací se podařilo ve srovnání s referenčním systémem založeným na Bayesovském informačním kritériu dosáhnout snížení chybové míry SPKE relativně v rozsahu o 15 až 42 %. Provedené experimentální vyhodnocení je mimo jiné specifické omezeným množstvím dat dostupných pro trénování modelů založených na faktorové analýze, kdy jejich celkový objem nepřevyšuje 10 hod. Z praktického hlediska je tak podstatným závěrem, vzhledem k dosaženým výsledkům, že i v případě takto omezeného množství dat je možné provést odhady hyperparametrů modelů založených na faktorové analýze.

Dále bylo navrženo schéma dvoufázového shlukování, které kombinuje přístup založený na Bayesovském informačním kritériu a přístup založený na i-vektorech. Motivací byla především snaha o zredukování počtu velmi krátkých segmentů, pro které je obtížné získat spolehlivý odhad koeficientů i-vektorů. Výhodou dvoufázového shlukování je navíc redukce celkové výpočetní náročnosti procesu shlukování. Důraz byl mimo jiné kladen na vhodnou aplikaci normalizace příznakových vektorů, která může v případně nevhodné aplikace zásadním způsobem narušit výsledky shlukování. Aplikací dvoufázového shlukování se podařilo dosáhnout relativního snížení SPKE až o 52,4 % oproti referenčnímu systému. Doba nutná pro shlukování přitom odpovídá v průměru 0,06 násobku délky zpracovávané nahrávky ve srovnání s 0,04 násobkem v případě referenčního systému.

9.1 Shrnutí přínosů k rozvoji vědního oboru

V práci je

- podán jednotný výklad základních i pokročilých metod pro textově nezávislé rozpoznávání mluvčích, založených jak na generativním, tak diskriminativním přístupu a umožňujících provádět kompenzaci variability akustických podmínek za účelem zvýšení robustnosti systémů;
- podán ucelený přehled základních a v komunitě zabývající se problematikou rozpoznávání mluvčích obecně přijímaných metod pro vyhodnocení systémů, které umožňují získat množství informací o chování systému za různých aplikačních podmínek;
- vytvořena evaluační databáze záznamů televizních a rozhlasových pořadů umožňující vyhodnocení systémů pro rozpoznávání mluvčích;
- provedeno experimentální porovnání systémů založených na generativním i diskriminativním přístupu na této databázi;
- provedeno experimentální vyhodnocení autorem implementovaných systémů založených na faktorové analýze s využitím standardních evaluačních databázích NIST SRE 2008 a 2010;
- formou spoluautorství vytvořena na základě dat korpusu COST278 evaluační databáze pořadů pro účely vyhodnocení systémů pro diarizaci mluvčích;
- proveden návrh tří přístupů založených na faktorové analýze pro shlukování segmentů v úloze diarizace mluvčích;
- navržena metoda dvoufázového shlukování využívající shlukování založeného na Bayesovském informačním kritériu v první fázi a jednu z metod založených na faktorové analýze ve druhé;
- experimentálně ověřen přínos prezentovaných metod v úloze diarizace mluvčích s využitím vytvořené databáze televizních a rozhlasových pořadů odvozené z korpusu COST 278;

9.2 Shrnutí přínosů pro praxi

Všechny v práci popsané metody byly autorem implementovány a vybraná řešení jsou součástí reálných systémů vyvíjených Laboratoří počítačového zpracování řeči TUL. Moduly pro diarizaci a rozpoznávání mluvčích jsou nedílnou součástí systému pro automatické zpracování rozsáhlých archivů mluvených dokumentů (jedná se např. o záznamy TVR pořadů nebo záznamy různých jednání).

Systém pro rozpoznávání mluvčích v záznamech telefonních hovorů vytvořený autorem práce byl testován bezpečnostními složkami v rámci projektu „Překlenutí jazykové bariéry, komplikující vyšetřování financování terorismu a závažné finanční kriminality.“

Citovaná literatura

ADAMI, A. G. et al. Modeling Prosodic Dynamics for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2003*, Hong Kong, 2003.

ALEXANDER, A. *Forensic Automatic Speaker Recognition Using Bayesian Interpretation and Statistical Compensation for Mismatched Conditions*. PhD thesis, Ecole Polytechnique Federale de Lausanne, 2005.

AUCKENTHALER, R. – CAREY, M. – LLOYD-THOMAS, H. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*. 2000, 10, 1–3, s. 42–54.

AYER, M. et al. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*. 1955, 26, 4, s. 641–647.

BEN, M. et al. Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs. In *in Intl. Conf. on Speech and Language Processing*, s. 2329–2332, Jeju Island, Korea, 2004.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. 2006.

BOTTI, F. – ALEXANDER, A. – DRYGAJLO, A. An Interpretation Framework for the Evaluation of Evidence in Forensic Automatic Speaker Recognition with Limited Suspect Data. In *Proceedings of 2004: A Speaker Odyssey*, s. 63–68, Toledo, Spain, 2004.

BRANDSCHAIN, L. et al. Speaker Recognition: Building the Mixer 4 and 5 Corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA).

BRICKER, P. et al. Statistical techniques for talker identification. *The Bell System Technical Journal*. 1971, 50, s. 1427–1454.

BRUMMER, N. *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, University of Stellenbosch, October 2010a.

BRUMMER, N. The EM algorithm and Minimum Divergence. Technical report, 2010b.

BRUMMER, N. – PREEZ, J. Application-independent evaluation of speaker detection. *Computer Speech and Language*. 2006, 20, 2, s. 230–275.

BRUMMER, N. – STRASHEIM, A. AGNITIO's Speaker Recognition System for EVALITA 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, December 2009.

BRUMMER, N. et al. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007, 15, 7, s. 2072–2084.

BRUMMER, N. et al. ABC System description for NIST SRE 2010. In *Proc. NIST 2010 Speaker Recognition Evaluation*, s. 1–20. Brno University of Technology, 2010.

BURGES, C. J. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*. 1998, 2, 2, s. 121–167.

BURGET, L. et al. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007, 15, 7, s. 1979–1986.

BURGET, L. et al. BUT system description: NIST SRE 2008. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, s. 1–4. National Institute of Standards and Technology, 2008.

BURGET, L. et al. Investigation into variants of Joint Factor Analysis for speaker recognition. In *10th Annual Conference of the International Speech Communication Association - Interspeech 2009*, Brighton, GB, 2009.

BURGET, L. et al. Discriminatively Trained Probabilistic Linear Discriminant Analysis for Speaker Verification. In *Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011*, s. 4832–4835. IEEE Signal Processing Society, 2011.

CAMPBELL, W. M. et al. Support vector machines for speaker and language recognition. *Computer Speech and Language*. 2006a, 20, s. 210–229.

CAMPBELL, W. M. Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002*, Orlando, FL, USA, 2002.

CAMPBELL, W. M. et al. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2006*, Toulouse, France, 2006b.

CAMPBELL, W. et al. Estimating and Evaluating Confidence for Forensic Speaker Recognition. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1, s. 717 – 720, 18-23, 2005.

CASTALDO, F. et al. Loquendo–Politecnico di Torino's 2010 NIST Speaker Recognition Evaluation System. In *Proc. IEEE ICASSP*, s. 5464–5467, Prague, May 2011.

ČERVA, P. *Řízená a neřízená adaptace na mluvčího v systémech rozpoznávání řeči.* PhD thesis, Technická univerzita v Liberci, Prosinec 2007.

CERVA, P. – NOUZA, J. – SILOVSKY, J. Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination. In *International Conference on Spoken Language Processing Interspeech - ICSLP 2006*, Pittsburgh, USA, 2006.

CHANG, C.-C. – LIN, C.-J. LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`, 2005.

CHEN, S. – GOPALAKRISHNAN, P. Speaker, environment and channel change detection and clustering via the Bayesian information criterion. In *Proceedings 1998 DARPA Broadcast News Transcription and Understanding Workshop*, s. 127–132, 1998.

CHEN, Z. – TANG, H. Sparse Bayesian approach to classification. In *Networking, Sensing and Control, 2005. Proceedings. 2005 IEEE*, s. 914–917, march 2005.

CIERI, C. et al. Resources for new research directions in speaker recognition: the Mixer 3, 4 and 5 corpora. In *INTERSPEECH*, s. 950–953. ISCA, 2007.

DEHAK, N. et al. Front-End Factor Analysis for Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on*. May 2011, 19, 4, s. 788 –798. ISSN 1558-7916.

DEHAK, N. et al. Support vector machines and joint factor analysis for speaker verification. In *Proc. ICASSP 2009*, s. 4. IEEE Signal Processing Society, 2009.

DEMPSTER, A. P. – LAIRD, N. M. – RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1977, 39, 1, s. 1–38. ISSN 00359246.

DO, M. N. Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*. 2003, 10, 4, s. 115–118.

DODDINGTON, G. Speaker Recognition based on Idiolectal Differences between Speakers. In *7th European Conference on Speech Communication and Technology - Eurospeech 2001*, Aalborg, Denmark, 2001.

DODDINGTON, G. R. et al. The NIST speaker recognition evaluation - Overview, methodology, systems, results, perspective. *Speech Communication*. 2000, 31, 2-3, s. 225 – 254. ISSN 0167-6393.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* June 2006, 27, s. 861–874. ISSN 0167-8655.

FAWCETT, T. – NICULESCU-MIZIL, A. PAV and the ROC convex hull. *Machine Learning*. 2007, 68, s. 97–106. ISSN 0885-6125. 10.1007/s10994-007-5011-0.

FERRAS, M. et al. Constrained MLLR for Speaker Recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 4, s. IV–53–IV–56, April 2007.

FERRER, L. et al. System combination using auxiliary information for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2008*, Las Vegas, NV, USA, 2008.

FINE, S. – NAVRATIL, J. – GOPINATH, R. A. A Hybrid GMM/SVM Approach to Speaker Identification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2001*, Salt Lake City, UT, USA, 2001.

FRANCO-PEDROSO, J. et al. ATVS-UAM System Description for the Audio Segmentation and Speaker Diarization Albayzin 2010 Evaluation. In *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop (FALA2010)*, s. 415–417, November 2010.

FURUI, S. 40 years of progress in automatic speaker recognition. *Advances in Biometrics*. 2009, 5558/2009, s. 1050–1059.

FURUI, S. 50 years of progress in speech and speaker recognition. In *SPECOM 2005*, Patras, Greece, 2005.

GALES, M. J. F. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*. 1998, 12, 2, s. 75–98.

GARCIA-ROMERO, D. – ESPY-WILSON, C. Y. Analysis of I-vector Length Normalization in Speaker Recognition Systems. In *Interspeech'11*, s. 249–252, Florence, Italy, August 2011.

GAUVAIN, J. L. – LEE, C.-H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *Speech and Audio Processing, IEEE Transactions on*. 1994, 2, 2, s. 291–298.

GLEMBEK, O. et al. Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis. In *Proc. ICASSP 2009*, s. 4. IEEE Signal Processing Society, 2009.

GONZALEZ-RODRIGUEZ, J. et al. Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems. *Forensic Science International*. December 2005, 155, 2-3, s. 126–140.

HATCH, A. O. – STOLCKE, A. Generalized Linear Kernels for One-Versus-All Classification: Application to Speaker Recognition. In *Proc. IEEE ICASSP*, 5, s. 585–588, Toulouse, May 2006.

HATCH, A. O. – KAJAREKAR, S. – STOLCKE, A. Within-Class Covariance Normalization for SVM-based Speaker Recognition. In *Proc. ICSLP*, s. 1471–1474, Pittsburgh, PA, September 2006.

HEIJDEN, F. *Classification, parameter estimation, and state estimation: an engineering approach using MATLAB*. Wiley, 2004. ISBN 9780470090138.

HERMANSKY, H. – MORGAN, N. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*. 1994, 2, 4, s. 578–589.

HUANG, X. – ACERO, A. – HON, H.-W. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Upper Saddle River, NJ, USA : Prentice Hall PTR, 1st edition, 2001. ISBN 0130226165.

JAIN, A. K. – ROSS, A. – PRABHAKAR, S. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. 2004, 14, s. 4–20.

KARAM, Z. N. – CAMPBELL, W. M. A new kernel for SVM MLLR based speaker recognition. In *Interspeech*, s. 290–293, 2007.

KATZ, M. *Discriminative Classifiers for Speaker Recognition*. PhD thesis, Otto-von-Guericke-Universitat Magdeburg, March 2008.

KENNY, P. et al. Factor Analysis Simplified. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, 1, s. 637 – 640, 18-23, 2005.

KENNY, P. et al. Speaker and Session Variability in GMM-Based Speaker Verification. *Audio, Speech, and Language Processing, IEEE Transactions on*. may 2007a, 15, 4, s. 1448 –1460. ISSN 1558-7916.

KENNY, P. Joint factor analysis of speaker and session variability : Theory and algorithms. Technical Report CRIM-06/08-13, Centre de recherche informatique de Montreal - CRIM, Montreal, Canada, 2005.

KENNY, P. – DUMOUCHEL, P. Disentangling speaker and channel effects in speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2004*, s. 37–40, Montreal, Canada, 2004a.

KENNY, P. – DUMOUCHEL, P. Experiments in Speaker Verification using Factor Analysis Likelihood Ratios. In *Proceedings of 2004: A Speaker Odyssey*, s. 219–226, Toledo, Spain, 2004b.

KENNY, P. et al. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*. 2007b, 15, 4, s. 1435–1447.

KENNY, P. et al. Development of the primary CRIM system for the NIST 2008 speaker recognition evaluation. In *INTERSPEECH*, s. 1401–1404, 2008a.

KENNY, P. et al. A Study of Interspeaker Variability in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*. 2008b, 16, 5, s. 980–988.

KINNUNEN, T. – LI, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*. January 2010, 52, s. 12–40. ISSN 0167-6393.

KINNUNEN, T. – HAUTAMAKI, V. – FRANTI, P. On the Use of Long-Term Average Spectrum in Automatic Speaker Recognition. In *Int. Symp. on Chinese Spoken Language Processing (ISCSLP06)*, s. 559–567, December 2006.

LEGGETTER, C. J. – WOODLAND, P. C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech & Language*. April 1995, 9, 2, s. 171–185. ISSN 08852308.

LONGWORTH, C. *Kernel Methods for Text-Independent Speaker Verification*. PhD thesis, University of Cambridge, February 2010.

LUQUE, J. – HERNANDO, J. Robust Speaker Identification for Meetings: UPC CLEAR'07 Meeting Room Evaluation System. In *Multimodal Technologies for Perception of Humans, International Evaluation Workshops CLEAR 2007 and RT 2007*, s. 266–275, 2007.

MARIETHOZ, J. – BENGIO, S. A Unified Framework for Score Normalization Techniques Applied to Text-Independent Speaker Verification. *Signal Processing Letters, IEEE*. July 2005, 12, 7, s. 532–535.

MARTIN, A. et al. The DET Curve In Assessment Of Detection Task Performance. In *Proc. Eurospeech '97*, s. 1895–1898, Rhodes, Greece, 1997.

MASON, M. et al. Data-Driven Clustering for Blind Feature Mapping in Speaker Verification. In *9th European Conference on Speech Communication and Technology - Interspeech 2005*, Lisboa, Portugal, 2005.

MEUWLY, D. – DRYGAJLO, A. Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM). In *Proceedings of The Workshop on Speaker Recognition "2001: A Speaker Odyssey"*, Proc. of IEEE, s. pp 145–150. SPIE, 2001.

MINKA, T. P. A comparison of numerical optimizers for logistic regression. Technical report, 2003.

NIST. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan, 2009.

NIST. The NIST Year 2008 Speaker Recognition Evaluation Plan. Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf`, 2008.

NIST. The NIST Year 2010 Speaker Recognition Evaluation Plan. Available at `http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf`, 2010.

NOUZA, J. et al. Continual On-Line Monitoring of Czech Spoken Broadcast Programs. In *International Conference on Spoken Language Processing Interspeech - ICSLP 2006*, Pittsburgh, USA, 2006.

OSUNA, E. – FREUND, R. – GIROSI, F. Support Vector Machines: Training and Applications. Technical report, Cambridge, MA, USA, 1997.

PARK, A. – HAZEN, T. J. ASR Dependent Techniques for Speaker Identification. In *International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

PELECANOS, J. – SRIDHARAN, S. Feature Warping for Robust Speaker Verification. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001.

PESKIN, B. et al. Using Prosodic and Conversational Features for High-Performance Speaker Recognition: Report from JHU WS'02. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2003*, Hong Kong, 2003.

PETERSEN, K. B. – PEDERSEN, M. S. The Matrix Cookbook, oct 2008. Version 20081110.

PIGEON, S. – DRUYTS, P. – VERLINDE, P. Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions. *Digital Signal Processing*. 2000, 10, 1-3, s. 237–248. ISSN 1051-2004.

PLATT, J. C. *Fast training of support vector machines using sequential minimal optimization*, s. 185–208. MIT Press, Cambridge, MA, USA, 1999.

PRAZAK, J. – SILOVSKY, J. Comparison of Segmentation and Clustering Methods for Speaker Diarization of Broadcast Stream Audio. In *Analysis of Verbal and Nonverbal Communication and Enactment*, 6800 / *Lecture Notes in Computer Science*, s. 214–222. Springer, 2011.

PRESS, W. et al. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, October 1992. ISBN 0521431085.

PRINCE, S. – ELDER, J. Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings ICCV 2007*, Rio de Janeiro, Brazil, October 2007.

PURDY, P. M. et al. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *In Advances in Neural Information Processing Systems 16*, , 2003. MIT Press.

RAMOS, D. *Forensic Evaluation of the Evidence Using Automatic Speaker Recognition Systems*. PhD thesis, Universidad Autonoma de Madrid, November 2007.

RAMOS, D. – GONZALEZ-RODRIGUEZ, J. Cross-Entropy Analysis of the Information in Forensic Speaker Recognition. In *Proc. of Odyssey*, January 2008.

RAMOS-CASTRO, D. et al. Between-sources modelling for likelihood ratio computation in forensic biometric recognition. In *Proc. 5th IAPR Intl. Conf. on Audio- and Video-based Biometric Person Authentication, AVBPA*, 3546 / *LNCS*, s. 1080–1089. Springer, July 2005.

REYNOLDS, D. A. SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2003*, Hong Kong, 2003.

REYNOLDS, D. A. Comparison of Background Normalization Methods for Text-Independent Speaker Verification. In *5th European Conference on Speech Communication and Technology - Eurospeech '97*, Rhodes, Greece, 1997.

REYNOLDS, D. A. – QUATIERI, T. F. – DUNN, R. B. Speaker verification using Adapted Gaussian mixture models. *Digital Signal Processing*. 2000, 1–3, s. 19–41.

SCHEFFER, N. et al. The SRI NIST 2010 Speaker Recognition evaluation system. In *Proc. IEEE ICASSP*, s. 5292–5295, Prague, May 2011.

SCHMIDT, M. – GISH, H. Speaker Identification via Support Vector Classifiers. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 1996*, Atlanta, Ga, USA, 1996.

SHRIBERG, E. et al. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*. 2005, 46, 3–4, s. 455–472.

SILOVSKY, J. TUL System for the NIST 2008 Speaker Recognition Evaluation. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

SILOVSKY, J. TUL NIST 2010 SRE System Description. In *Proc. 2010 NIST Speaker Recognition Evaluation Workshop*, 2010.

SILOVSKY, J. – CERVA, P. Study on Speaker Recognition aided Broadcast Stream Transcription. In *16th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2006.

SILOVSKY, J. – NOUZA, J. Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream. *Radioengineering*. 2006, 15, 3, s. 42–48.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. Text-Independent Speaker Verification Supported by ASR. In *17th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2007.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. Comparison of Generative and Discriminative Approaches for Speaker Recognition with Limited Data. *Radioengineering*. 2009, 18, 3, s. 307–316.

SILOVSKY, J. et al. PLDA-based Clustering for Speaker Diarization of Broadcast Streams. In *INTERSPEECH'11*, s. 2909–2912. ISCA, August 2011.

SMITH, N. – GALES, M. Speech Recognition using SVMs. In *Advances in Neural Information Processing Systems*, 14, s. 1197–1204. MIT Press, 2002.

SOLOMONOFF, A. – QUILLEN, C. – CAMPBELL, W. M. Channel Compensation for SVM Speaker Recognition. In *Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

SOLOMONOFF, A. – CAMPBELL, W. M. – BOARDMAN, I. Advances in channel compensation for SVM speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2005*, Philadelphia, PA, USA, 2005.

SPEARMAN, C. General Intelligence, objectively determined and measured. *American J. Psychology*. 1904, 15, s. 201–293.

STOLCKE, A. et al. MLLR transforms as features in speaker recognition. In *9th European Conference on Speech Communication and Technology - Interspeech 2005*, Lisboa, Portugal, 2005.

STOLCKE, A. – FERRER, L. – KAJAREKAR, S. S. Improvements in MLLR-transform-based speaker recognition. In *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.

STRASHEIM, A. – BRUMMER, N. SUNSDV System Description: NIST SRE 2008. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

STURIM, D. et al. The MIT LL 2010 Speaker Recognition Evaluation System: Scalable Language-Independent Speaker Recognition. In *Proc. IEEE ICASSP*, s. 5272–5275, Prague, May 2011.

STURIM, D. E. et al. Speaker Verification using Text-Constrained Gaussian Mixture models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002*, Orlando, FL, USA, 2002.

TEUNEN, R. – SHAHSHAHANI, B. – HECK, L. A Model-Based Transformational Approach to Robust Speaker Recognition. In *6th International Conference on Spoken Language Processing - ICSLP 2000*, Beijing, China, 2000.

TRANTER, S. – REYNOLDS, D. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*. sept. 2006, 14, 5, s. 1557 –1565. ISSN 1558-7916.

LEEUWEN, D. A. – BRUMMER, N. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. *Lecture Notes in Computer Science*. 2007, 4343/2007, s. 330–353.

LEEUWEN, D. A. et al. NIST and NFI-TNO evaluations of automatic speaker recognition. *Computer Speech & Language*. 2006, 20, 2-3, s. 128 – 158. ISSN 0885-2308. Odyssey 2004: The speaker and Language Recognition Workshop - Odyssey-04.

VAN VUUREN, S. *Speaker verification in a time-feature space*. PhD thesis, 1999. AAI9920388.

VANDECATSEYE, A. et al. The COST278 pan-European Broadcast News Database. s. 873–876, 2004.

VAPNIK, V. N. *Statistical Learning Theory*. Wiley-Interscience, September 1998. ISBN 0471030031.

VOGT, R. – SRIDHARAN, S. Explicit Modelling of Session Variability for Speaker Verification. *Computer Speech and Language*. 2008, 22, 1, s. 17–38.

VOGT, R. – BAKER, B. – SRIDHARAN, S. Modelling Session Variability in Text-Independent Speaker Verification. In *9th European Conference on Speech Communication and Technology - Interspeech 2005*, Lisboa, Portugal, 2005.

WAN, V. – RENALS, S. Speaker verification using sequence discriminant support vector machines. *Speech and Audio Processing, IEEE Transactions on*. 2005, 13, 2, s. 203–210.

WAN, V. – CAMPBELL, W. M. Support Vector Machines for Speaker Verification and Identification. In *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, Sydney, Australia, 2000.

WILBUR, W. – YEGANOVA, L. – KIM, W. The Synergy Between PAV and AdaBoost. *Machine Learning*. 2005, 61, s. 71–103. ISSN 0885-6125. 10.1007/s10994-005-1123-6.

XIANG, B. et al. Short-time Gaussianization for robust speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002*, Orlando, FL, USA, 2002.

ZDANSKY, J. – CHALOUPKA, J. – NOUZA, J. Joint Audio-Visual Processing, Representation and Indexing of TV News Programmes. In *IEEE 10th Workshop on Multimedia Signal Processing - MMSP 2008*, Cairns, Australia, 2008.

ZELENAK, M. – SCHULZ, H. – HERNANDO, J. Albayzin 2010 Evaluation Campaign: Speaker Diarization. In *VI Jornadas en Tecnologia del Habla and II Iberian SLTech Workshop (FALA2010)*, s. 301–304, November 2010.

ZHOU, B. – HANSEN, J. Unsupervised Audio Stream Segmentation And Clustering Via The Bayesian Information Criterion. In *in Proc. ISCLP 2000*, s. 714–717, 2000.

ZHU, X. et al. Combining speaker identification and BIC for speaker diarization. In *Interspeech'05, ISCA*, Lisbon, September 2005.

ZHU, X. et al. Multimodal Technologies for Perception of Humans. Berlin, Heidelberg: Springer-Verlag, 2008. Multi-stage Speaker Diarization for Conference and Lecture Meetings, s. 533–542. ISBN 978-3-540-68584-5.

Seznam vlastních publikací

CERVA, P. – NOUZA, J. – SILOVSKY, J. Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination. In *International Conference on Spoken Language Processing Interspeech - ICSLP 2006*, Pittsburgh, USA, 2006.

CERVA, P. – ZDANSKY, J. – SILOVSKY, J. – NOUZA, J. Study on Speaker Adaptation Methods in the Broadcast News Transcription Task. In *Proceedings of the 11th international conference on Text, Speech and Dialogue*, TSD '08, s. 277–284, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87390-7.

CERVA, P. – PALECEK, K. – SILOVSKY, J. – NOUZA, J. An Investigation into VTLN for Improved Transcription of Czech Broadcast Programs. In *53rd International IEEE Symposium ELMAR-2011*, s. 201–204, September 2011a. ISBN 978-953-7044-12-1.

CERVA, P. – PALECEK, K. – SILOVSKY, J. – NOUZA, J. Using Unsupervised Feature-Based Speaker Adaptation for Improved Transcription of Spoken Archives. In *INTERSPEECH*, s. 2565–2568. ISCA, August 2011b.

CHALOUPKA, J. – NOUZA, J. – ZDANSKY, J. – CERVA, P. – SILOVSKY, J. – KROUL, M. Voice Technology Applied for Building a Prototype Smart Room. In *Multimodal Signals: Cognitive and Algorithmic Issues*. Berlin, Heidelberg: Springer-Verlag, 2009. s. 104–111. ISBN 978-3-642-00524-4.

LOPEZ-COZAR, R. – CALLEJAS, Z. – KROUL, M. – NOUZA, J. – SILOVSKY, J. Two-Level Fusion to Improve Emotion Classification in Spoken Dialogue Systems. In *Proceedings of the 11th international conference on Text, Speech and Dialogue*, TSD '08, s. 617–624, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-87390-7.

LOPEZ-COZAR, R. – SILOVSKY, J. – GRIOL, D. Enhancement of spoken dialogue systems by Means of User Emotion Recognition. *Spanish Journal On Natural Language Processing (SEPLN)*. September 2010a, 45, s. 191–198. ISSN 1558-7916. ve španělšině.

LOPEZ-COZAR, R. – SILOVSKY, J. – GRIOL, D. F2 – new technique for recognition of user emotional states in spoken dialogue systems. In *Proceedings of the 11th Annual Meeting of the*

Special Interest Group on Discourse and Dialogue, SIGDIAL '10, s. 281–288. Association for Computational Linguistics, 2010b. ISBN 978-1-932432-85-5.

LOPEZ-COZAR, R. – SILOVSKY, J. – KROUL, M. Enhancement of emotion detection in spoken dialogue systems by combining several information sources. *Speech Communication*. November 2011, 53, s. 1210–1228. ISSN 0167-6393.

NOUZA, J. – SILOVSKY, J. Fast Keyword Spotting in Telephone Speech. *Radioengineering*. 2009, 18, 4, s. 665–670.

NOUZA, J. – SILOVSKY, J. Adapting lexical and language models for transcription of highly spontaneous spoken Czech. In *Proceedings of the 13th international conference on Text, speech and dialogue*, TSD '10, s. 377–384, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15759-9, 978-3-642-15759-2.

NOUZA, J. – CHALOUPKA, J. – ZDANSKY, J. – SILOVSKY, J. – KROUL, M. – MADER, Z. Voice Controlled Center for Homes of Motor-Handicapped Persons. In *Speech and Computer International Conference - Specom 2007*, s. 714–719, Moscow, Russia, 2007.

NOUZA, J. – SILOVSKY, J. – ZDANSKY, J. – CERVA, P. – KROUL, M. – CHALOUPKA, J. Czech-to-Slovak Adapted Broadcast News Transcription System. In *9th Annual Conference of the International Speech Communication Association - Interspeech 2008*, s. 2683–2686, Brisbane, Australia, 2008.

NOUZA, J. – ZDANSKY, J. – CERVA, P. – SILOVSKY, J. Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak). In ESPOSITO, A. – CAMPBELL, N. – VOGEL, C. – HUSSAIN, A. – NIJHOLT, A. (Ed.) *COST 2102 Training School*, 5967 / *Lecture Notes in Computer Science*, s. 225–241. Springer, 2009. ISBN 978-3-642-12396-2.

NOUZA, J. – BLAVKA, K. – BOHAC, M. – CERVA, P. – ZDANSKY, J. – SILOVSKY, J. – PRAZAK, J. Voice technology to enable sophisticated access to historical Czech Radio audio archive. In *MM4CH 2011*, 247 / *Communications in Computer and Information Science*, s. 27–38. Springer, 2011.

PRAZAK, J. – SILOVSKY, J. Comparison of Segmentation and Clustering Methods for Speaker Diarization of Broadcast Stream Audio. In *Analysis of Verbal and Nonverbal Communication and Enactment*, 6800 / *Lecture Notes in Computer Science*, s. 214–222. Springer, 2011a.

PRAZAK, J. – SILOVSKY, J. Speaker Diarization Using PLDA-based Speaker Clustering. In *The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, s. 347–350, September 2011b.

SILOVSKY, J. TUL System for the NIST 2008 Speaker Recognition Evaluation. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

SILOVSKY, J. TUL NIST 2010 SRE System Description. In *Proc. 2010 NIST Speaker Recognition Evaluation Workshop*, 2010.

SILOVSKY, J. – CERVA, P. Study on Speaker Recognition aided Broadcast Stream Transcription. In *16th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2006.

SILOVSKY, J. – CERVA, P. Analysis of Eigenchannel Adaptation in Broadcast News Speaker Recognition System. In *9th International workshop on Electronics, Control, Modelling, Measurement and Signals 2009*, Mondragon, Spain, 2009. ISBN 978-84-608-0941-8.

SILOVSKY, J. – NOUZA, J. Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream. *Radioengineering*. 2006, 15, 3, s. 42–48.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. Text-Independent Speaker Verification Supported by ASR. In *17th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2007.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. Comparison of Generative and Discriminative Approaches for Speaker Recognition with Limited Data. *Radioengineering*. 2009a, 18, 3, s. 307–316.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. MLLR Transforms Based Speaker Recognition in Broadcast Streams. In ESPOSITO, A. – VICH, R. (Ed.) *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, 5641 / *Lecture Notes in Computer Science*. Prague,Czech Republic: Springer Berlin / Heidelberg, 2009b. s. 423–431. ISBN 978-3-642-03319-3.

SILOVSKY, J. – CERVA, P. – ZDANSKY, J. Assessment of Speaker Recognition on Lossy Codecs Used for Transmission of Speech. In *53rd International IEEE Symposium ELMAR-2011*, s. 205–208, September 2011a. ISBN 978-953-7044-12-1.

SILOVSKY, J. – PRAZAK, J. – CERVA, P. – ZDANSKY, J. – NOUZA, J. PLDA-based Clustering for Speaker Diarization of Broadcast Streams. In *INTERSPEECH'11*, s. 2909–2912. ISCA, August 2011b.

Příloha A

Přehled výpočtů pro oddělený odhad hyperparametrů JFA modelu

Tato příloha obsahuje přehled všech důležitých výpočtů nutných pro oddělený odhad hyperparametrů JFA modelu $\boldsymbol{\Lambda} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{D}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$ principiálně popsaný v podkapitole 4.3.8. Předpokládejme, že je k dispozici trénovací databáze obsahující nahrávky od S mluvčích, přitom pro každého mluvčího s je k dispozici R_s nahrávek, kde $R_s > 1$. Pro každou nahrávku byly s využitím UBM vyhodnoceny postačující statistiky nultého, prvního a druhého řádu podle (4.23), tj. $N_{r,s,c}$, $\boldsymbol{F}_{r,s,c}$ a $\boldsymbol{S}_{r,s,c}$.

Nejprve je proveden odhad zátěžové matice $\boldsymbol{V}$ společně s odhadem $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$. Celý proces trénování shrnuje Tab. A.1. Počáteční model je tvořen pětic hyperparametrů $\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \boldsymbol{0}, \boldsymbol{0}, \hat{\boldsymbol{\Sigma}}\}$. Inicializace $\hat{\boldsymbol{\mu}}$ a $\hat{\boldsymbol{\Sigma}}$ je provedena na základě parametrů UBM a matice $\hat{\boldsymbol{V}}$ je inicializována náhodně.

Pro úplnost uveďme, že platí

$$\mathbb{E}[\acute{\boldsymbol{y}}_s] = \begin{bmatrix} \mathbb{E}[\boldsymbol{y}_s] \\ 1 \end{bmatrix} \tag{A.1}$$

$$\mathbb{E}[\acute{\boldsymbol{y}}_s \acute{\boldsymbol{y}}_s^T] = \begin{pmatrix} \mathbb{E}[\boldsymbol{y}_s \boldsymbol{y}_s^T] & \mathbb{E}[\boldsymbol{y}_s] \\ \mathbb{E}[\boldsymbol{y}_s^T] & 1 \end{pmatrix} \tag{A.2}$$

Tab. A.2 shrnuje následný odhad matice $\boldsymbol{U}$ reprezentující podprostor charakteristický pro variabilitu akustických podmínek. Počáteční hodnoty hyperparametrů $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{V}}$ a $\hat{\boldsymbol{\Sigma}}$ odpovídají odhadům provedeným v předchozím kroku a inicializace $\hat{\boldsymbol{U}}$ je provedena náhodně. Cílem je stanovit na základě trénovacích dat matici $\boldsymbol{U}$ vyhovující modelu

$$\boldsymbol{m}_{r,s} = \mathbb{E}[\boldsymbol{s}_s] + \boldsymbol{U}\boldsymbol{w}_{r,s}. \tag{A.3}$$

Pro každou nahrávku z trénovací databáze jsou postačující statistiky vycentrovány kolem supervektoru $\boldsymbol{s}_s$ charakteristického pro mluvčího dané nahrávky. Odhad tohoto supervektoru $\mathbb{E}[\boldsymbol{s}_s]$ je získán na základě posteriorního rozložení $\boldsymbol{y}_s$, stanoveného s využitím statistik sečtených přes všechny nahrávky mluvčího s (viz sekce 4.3.8), jako

$$\mathbb{E}[\boldsymbol{s}_s] = \hat{\boldsymbol{\mu}} + \mathbb{E}[\boldsymbol{y}_s]. \tag{A.4}$$

Tabulka A.1: *Trénování hyperparametrů JFA modelu - odhad matice* $\boldsymbol{V}$

Vstupní model	$\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \boldsymbol{0}, \boldsymbol{0}, \hat{\boldsymbol{\Sigma}}\}$
Vstupní postačující statistiky	$\boldsymbol{N}_s = \sum_r \boldsymbol{N}_{r,s}$ $\boldsymbol{F}_s = \sum_r \boldsymbol{F}_{r,s}$ $\boldsymbol{S}_s = \sum_r \boldsymbol{S}_{r,s}$
Posteriorní rozložení skrytých proměnných	$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{V}$ $\mathbb{E}[\boldsymbol{y}_s] = \boldsymbol{L}^{-1}\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}_s(\hat{\boldsymbol{\mu}})$ $\mathrm{cov}(\boldsymbol{y}_s, \boldsymbol{y}_s) = \boldsymbol{L}^{-1}$
Pomocné statistiky a proměnné	$\boldsymbol{\mathfrak{A}}_c = \sum_s N_{s,c}\mathbb{E}[\acute{\boldsymbol{y}}_s\acute{\boldsymbol{y}}_s^T]$ $\boldsymbol{\mathfrak{B}} = \sum_s \boldsymbol{F}_s\mathbb{E}[\acute{\boldsymbol{y}}_s^T]$ $\boldsymbol{N} = \sum_s \boldsymbol{N}_s$ $\boldsymbol{\mu}_{\boldsymbol{y}} = \frac{1}{S}\sum_s \mathbb{E}[\boldsymbol{y}_s]$ $\boldsymbol{K}_{\boldsymbol{yy}}^{1/2}(\boldsymbol{K}_{\boldsymbol{yy}}^{1/2})^T = \frac{1}{S}\sum_s \mathbb{E}[\boldsymbol{y}_s\boldsymbol{y}_s^T] - \boldsymbol{\mu}_{\boldsymbol{y}}\boldsymbol{\mu}_{\boldsymbol{y}}^T$
ML odhad	$\acute{\boldsymbol{V}}_c^{ML} = \boldsymbol{\mathfrak{B}}_c\boldsymbol{\mathfrak{A}}_c^{-1}$ $\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\left(\sum_s \boldsymbol{S}_s - \mathrm{diag}\left(\boldsymbol{\mathfrak{B}}(\acute{\boldsymbol{V}}^{ML})^T\right)\right)$
MD odhad	$\boldsymbol{V} = \boldsymbol{V}^{ML}\boldsymbol{K}_{\boldsymbol{yy}}^{1/2}$ $\boldsymbol{\mu} = \boldsymbol{\mu}^{ML} + \boldsymbol{V}^{ML}\boldsymbol{\mu}_{\boldsymbol{y}}$
Výstupní model	$\boldsymbol{\Lambda} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{0}, \boldsymbol{0}, \boldsymbol{\Sigma}\}$

Nakonec je proveden odhad matice $\boldsymbol{D}$. V tomto případě jsou statistiky vycentrovány kolem supervektoru $\boldsymbol{h}_{r,s}$. Jeho odhad je stanoven na základě posteriorního rozložení faktorů $\boldsymbol{y}_s$ a $\boldsymbol{w}_{r,s}$ jako

$$\mathbb{E}[\boldsymbol{h}_{r,s}] = \mathbb{E}[\boldsymbol{s}_s] + \boldsymbol{U}\mathbb{E}[\boldsymbol{w}_{r,s}], \tag{A.5}$$

kde odhad $\mathbb{E}[\boldsymbol{s}_s]$ je proveden stejně jako v předchozím kroku odhadu $\boldsymbol{U}$ (tedy s využitím součtů postačujících statistik vyhodnocených pro jednotlivé nahrávky mluvčího). Následně jsou statistiky odpovídající jednotlivým nahrávkám vycentrovány kolem supervektoru $\mathbb{E}[\boldsymbol{s}_s]$ a je stanoveno posteriorní rozložení faktorů $\boldsymbol{w}_{r,s}$. Pak již nic nebrání vyjádření (A.5). Vyjma náhodné inicializace matice $\hat{\boldsymbol{D}}$ odpovídají hodnoty ostatních hyperparametrů odhadům provedeným v předchozích krocích. Postup trénování $\boldsymbol{D}$ shrnuje Tab. A.3.

Tabulka A.2: *Trénování hyperparametrů JFA modelu - odhad matice* $\boldsymbol{U}$

Vstupní model	$\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{U}}, \boldsymbol{0}, \hat{\boldsymbol{\Sigma}}\}$
Vstupní postačující statistiky	$\boldsymbol{N}_{r,s}$ $\boldsymbol{F}_{r,s}$ $\boldsymbol{S}_{r,s}$
Posteriorní rozložení skrytých proměnných	$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_{r,s}\boldsymbol{U}$ $\mathbb{E}[\boldsymbol{w}_{r,s}] = \boldsymbol{L}^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}_{r,s}(\mathbb{E}[\boldsymbol{s}_s])$ $\mathrm{cov}(\boldsymbol{w}_{r,s}, \boldsymbol{w}_{r,s})\ \boldsymbol{L}^{-1}$
Pomocné statistiky a proměnné	$\boldsymbol{\mathfrak{A}}_c = \sum_s \sum_r N_{r,s,c}\mathbb{E}[\boldsymbol{w}_{r,s}\boldsymbol{w}_{r,s}^T]$ $\boldsymbol{\mathfrak{B}} = \sum_s \sum_r \tilde{\boldsymbol{F}}_{r,s}(\mathbb{E}[\boldsymbol{s}_s])\mathbb{E}[\boldsymbol{w}_{r,s}^T]$ $\boldsymbol{K}_{\boldsymbol{ww}}^{1/2}(\boldsymbol{K}_{\boldsymbol{ww}}^{1/2})^T = \frac{1}{\sum_s R_s}\sum_s \sum_r \mathbb{E}[\boldsymbol{w}_{r,s}\boldsymbol{w}_{r,s}^T]$
ML odhad	$\boldsymbol{U}_c^{ML} = \boldsymbol{\mathfrak{B}}_c\boldsymbol{\mathfrak{A}}_c^{-1}$
MD odhad	$\boldsymbol{U} = \boldsymbol{U}^{ML}\boldsymbol{K}_{\boldsymbol{ww}}^{1/2}$
Výstupní model	$\boldsymbol{\Lambda} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \boldsymbol{U}, \boldsymbol{0}, \hat{\boldsymbol{\Sigma}}\}$

Tabulka A.3: *Trénování hyperparametrů JFA modelu - odhad diagonální matice* $\boldsymbol{D}$

Vstupní model	$\hat{\boldsymbol{\Lambda}} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{U}}, \hat{\boldsymbol{D}}, \hat{\boldsymbol{\Sigma}}\}$
Vstupní postačující statistiky	$\boldsymbol{N}_s = \sum_r \boldsymbol{N}_{r,s}$ $\tilde{\boldsymbol{F}}_s(\mathbb{E}[\boldsymbol{h}_{.,s}]) = \sum_r \tilde{\boldsymbol{F}}_{r,s}(\mathbb{E}[\boldsymbol{h}_{r,s}])$ $\tilde{\boldsymbol{S}}_s(\mathbb{E}[\boldsymbol{h}_{.,s}]) = \sum_r \tilde{\boldsymbol{S}}_{r,s}(\mathbb{E}[\boldsymbol{h}_{r,s}])$
Posteriorní rozložení skrytých proměnných	$\boldsymbol{L} = \boldsymbol{I} + \boldsymbol{D}^2\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s$ $\mathbb{E}[\boldsymbol{z}_s] = \boldsymbol{L}^{-1}\boldsymbol{D}\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{F}}_s(\mathbb{E}[\boldsymbol{h}_{.,s}])$ $\mathrm{cov}(\boldsymbol{z}_s, \boldsymbol{z}_s) = \boldsymbol{L}^{-1}$
Pomocné statistiky a proměnné	$\boldsymbol{\mathfrak{A}} = \sum_s \mathrm{diag}\left(\boldsymbol{N}_s\mathbb{E}[\boldsymbol{z}_s\boldsymbol{z}_s^T]\right)$ $\boldsymbol{\mathfrak{B}} = \sum_s \mathrm{diag}\left(\tilde{\boldsymbol{F}}_s(\mathbb{E}[\boldsymbol{h}_{.,s}])\mathbb{E}[\boldsymbol{z}_s^T]\right)$ $\boldsymbol{N} = \sum_s \boldsymbol{N}_s$ $\boldsymbol{K}_{\boldsymbol{zz}}^{1/2}\boldsymbol{K}_{\boldsymbol{zz}}^{1/2} = \frac{1}{S}\sum_s \mathrm{diag}\left(\mathbb{E}[\boldsymbol{z}_s\boldsymbol{z}_s^T]\right)$
ML odhad	$\boldsymbol{D}^{ML} = \boldsymbol{\mathfrak{B}}\boldsymbol{\mathfrak{A}}^{-1}$ $\boldsymbol{\Sigma} = \boldsymbol{N}^{-1}\left(\sum_s \tilde{\boldsymbol{S}}_s(\mathbb{E}[\boldsymbol{z}_s\boldsymbol{z}_s^T]) - \mathrm{diag}\left(\boldsymbol{\mathfrak{B}}(\boldsymbol{D}^{ML})^T\right)\right)$
MD odhad	$\boldsymbol{D} = \boldsymbol{D}^{ML}\boldsymbol{K}_{\boldsymbol{zz}}^{1/2}$
Výstupní model	$\boldsymbol{\Lambda} = \{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{V}}, \hat{\boldsymbol{U}}, \boldsymbol{D}, \boldsymbol{\Sigma}\}$

Příloha B

Výpočet parametrů sdruženého posteriorní rozložení faktorů PLDA modelu

Stejně jako v případě trénování hyperparametrů JFA modelu je i pro trénování hyperparametrů PLDA modelu klíčový výpočet parametrů posteriorního rozložení skrytých proměnných při daných pozorovatelných (trénovacích) datech (viz sekce 4.3.4),

Uvažujme kombinovaný model (4.97), parametry posteriorního rozložení $\underline{\boldsymbol{\kappa}}$ jsou pak v souladu s (4.101) a (4.102) stanoveny následovně (pro zjednodušení zápisu je zde vynechána reference na mluvčího s):

$$\mathbb{E}[\underline{\boldsymbol{\kappa}}] = \underline{\boldsymbol{L}}^{-1}\underline{\boldsymbol{W}}\underline{\boldsymbol{\Sigma}}^{-1}\underline{\tilde{\boldsymbol{x}}} \tag{B.1}$$

$$\mathrm{cov}(\underline{\boldsymbol{\kappa}}, \underline{\boldsymbol{\kappa}}) = \underline{\boldsymbol{L}}^{-1} \tag{B.2}$$

kde vektor $\underline{\tilde{\boldsymbol{x}}}$ je stanoven podle (4.100) a

$$\underline{\boldsymbol{L}} = \underline{\boldsymbol{I}} + \underline{\boldsymbol{W}}^T\underline{\boldsymbol{\Sigma}}^{-1}\underline{\boldsymbol{W}}. \tag{B.3}$$

Na základě znalosti struktury matice $\underline{\boldsymbol{W}}$ je možné provést vyjádření

$$\underline{\boldsymbol{L}} = \begin{pmatrix} \mathfrak{A} & & & \mathfrak{B} \\ & \ddots & & \vdots \\ & & \mathfrak{A} & \mathfrak{B} \\ \mathfrak{B}^T & \dots & \mathfrak{B}^T & \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{I} \end{pmatrix} = \begin{pmatrix} \underline{\mathfrak{A}} & \underline{\mathfrak{B}} \\ \underline{\mathfrak{B}}^T & \mathfrak{D} \end{pmatrix} \tag{B.4}$$

kde

$$\mathfrak{A} = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U} + \boldsymbol{I} \tag{B.5}$$

$$\mathfrak{B} = \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} \tag{B.6}$$

Inverzi $\underline{\boldsymbol{L}}$ pak lze vyjádřit s využitím uvedené dekompozice blokově takto[1] (Press et al., 1992)

$$\underline{\boldsymbol{L}}^{-1} = \begin{pmatrix} (\underline{\mathfrak{A}} - \underline{\mathfrak{B}}\mathfrak{D}^{-1}\underline{\mathfrak{B}}^T)^{-1} & -\underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}}(\mathfrak{D} - \underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}})^{-1} \\ -(\mathfrak{D} - \underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}})^{-1}\underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1} & (\mathfrak{D} - \underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}})^{-1} \end{pmatrix} = \begin{pmatrix} \underline{\mathfrak{K}} & \underline{\mathfrak{L}} \\ \underline{\mathfrak{L}}^T & \mathfrak{N} \end{pmatrix} \tag{B.7}$$

Protože matice $\underline{\mathfrak{D}}$ je blokově diagonální, platí $\underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}} = R\mathfrak{B}^T\mathfrak{A}^{-1}\mathfrak{B}$ a

$$\mathfrak{N} = \left(R(\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \mathfrak{B}^T\mathfrak{A}^{-1}\mathfrak{B}) + \boldsymbol{I}\right)^{-1} \tag{B.8}$$

$$\underline{\mathfrak{L}}^T = -\mathfrak{N}\underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1} = -\begin{pmatrix} \mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} & \dots & \mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} \end{pmatrix} \tag{B.9}$$

Aplikací Woodburyho maticové identity[2] (Petersen – Pedersen, 2008) získáme vyjádření

$$\begin{aligned} \underline{\mathfrak{K}} &= \underline{\mathfrak{A}}^{-1} + \underline{\mathfrak{A}}^{-1}\underline{\mathfrak{B}}\mathfrak{N}\underline{\mathfrak{B}}^T\underline{\mathfrak{A}}^{-1} \\ &= \begin{pmatrix} \mathfrak{A}^{-1} & & \\ & \ddots & \\ & & \mathfrak{A}^{-1} \end{pmatrix} + \begin{pmatrix} \mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} & \dots & \mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} \\ \vdots & \vdots\ \vdots\ \vdots & \vdots \\ \mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} & \dots & \mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} \end{pmatrix} \end{aligned} \tag{B.10}$$

Nyní lze vyjádřit $\underline{\boldsymbol{L}}^{-1}$ následovně

$$\underline{\boldsymbol{L}}^{-1} = \begin{pmatrix} \mathfrak{K}_d & \mathfrak{K}_o & \dots & \mathfrak{K}_o & \mathfrak{L} \\ \mathfrak{K}_o & \mathfrak{K}_d & \ddots & \mathfrak{K}_o & \mathfrak{L} \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \mathfrak{K}_o & \mathfrak{K}_o & \dots & \mathfrak{K}_d & \mathfrak{L} \\ \mathfrak{L}^T & \mathfrak{L}^T & \dots & \mathfrak{L}^T & \mathfrak{N} \end{pmatrix} \tag{B.11}$$

kde

$$\mathfrak{K}_o = \mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N}\mathfrak{B}^T\mathfrak{A}^{-1} \tag{B.12}$$

$$\mathfrak{K}_d = \mathfrak{A}^{-1} + \mathfrak{K}_o \tag{B.13}$$

$$\mathfrak{L} = -\mathfrak{A}^{-1}\mathfrak{B}\mathfrak{N} \tag{B.14}$$

Pro účely odhadu hyperparametrů nás nezájímá přímo kovarianční matice odpovídající posteriornímu rozložení vektoru skrytých proměnných $\underline{\boldsymbol{\kappa}}$, tj. $\mathrm{cov}(\underline{\boldsymbol{\kappa}}, \underline{\boldsymbol{\kappa}})$, ale kovarianční matice pro vektor $\boldsymbol{\kappa}_r$, definovaný podle (4.103), ta má podobu

$$\mathrm{cov}(\boldsymbol{\kappa}_r, \boldsymbol{\kappa}_r) = \begin{pmatrix} \mathfrak{K}_d & \mathfrak{L} \\ \mathfrak{L}^T & \mathfrak{N} \end{pmatrix} \tag{B.15}$$

a je shodná pro všechny nahrávky $r = 1, \dots, R$.

---

[1] $\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{pmatrix}^{-1} = \begin{pmatrix} (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1} & -\boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \\ -(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & (\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{pmatrix}$

[2] $(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$

Nyní provedeme vyjádření $\mathbb{E}[\boldsymbol{\kappa}_r]$. Na základě (B.1) lze získat

$$\begin{aligned}\mathbb{E}[\underline{\boldsymbol{\kappa}}] &= \underline{\boldsymbol{L}}^{-1} \begin{pmatrix} \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_1 \\ \vdots \\ \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_R \\ \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\sum_r \tilde{\boldsymbol{x}}_r \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\mathfrak{A}}^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_1 + \boldsymbol{\mathfrak{K}}_o\boldsymbol{\mathfrak{S}}_U + \boldsymbol{\mathfrak{L}}\boldsymbol{\mathfrak{S}}_V \\ \vdots \\ \boldsymbol{\mathfrak{A}}^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_R + \boldsymbol{\mathfrak{K}}_o\boldsymbol{\mathfrak{S}}_U + \boldsymbol{\mathfrak{L}}\boldsymbol{\mathfrak{S}}_V \\ \boldsymbol{\mathfrak{L}}^T\boldsymbol{\mathfrak{S}}_U + \boldsymbol{\mathfrak{N}}\boldsymbol{\mathfrak{S}}_V \end{pmatrix}\end{aligned} \tag{B.16}$$

kde

$$\boldsymbol{\mathfrak{S}}_U = \sum_r \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_r \tag{B.17}$$

$$\boldsymbol{\mathfrak{S}}_V = \sum_r \boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_r \tag{B.18}$$

Očekávána hodnota vektoru skrytých proměnných $\boldsymbol{\kappa}_r$ je pak rovna

$$\mathbb{E}[\boldsymbol{\kappa}_r] = \mathbb{E}\begin{bmatrix} \boldsymbol{w}_r \\ \boldsymbol{y} \end{bmatrix} = \begin{pmatrix} \boldsymbol{\mathfrak{A}}^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{x}}_r + \boldsymbol{\mathfrak{K}}_o\boldsymbol{\mathfrak{S}}_U + \boldsymbol{\mathfrak{L}}\boldsymbol{\mathfrak{S}}_V \\ \boldsymbol{\mathfrak{L}}^T\boldsymbol{\mathfrak{S}}_U + \boldsymbol{\mathfrak{N}}\boldsymbol{\mathfrak{S}}_V \end{pmatrix} \tag{B.19}$$

Poznamejme, že $\mathbb{E}[\boldsymbol{y}]$ je shodný pro všechny nahrávky, což odpovídá původnímu předpokladu.

Pro úplnost ještě uveďme výpočet hodnoty determinantu matice $\underline{\boldsymbol{L}}$. Determinant matice lze na základě blokového rozkladu (B.4) vyjádřit takto[3]:

$$|\underline{\boldsymbol{L}}| = |\underline{\boldsymbol{\mathfrak{A}}}| \left|\underline{\boldsymbol{\mathfrak{D}}} - \underline{\boldsymbol{\mathfrak{B}}}^T\underline{\boldsymbol{\mathfrak{A}}}^{-1}\underline{\boldsymbol{\mathfrak{B}}}\right|. \tag{B.20}$$

Protože je matice $\underline{\boldsymbol{\mathfrak{A}}}$ blokově diagonální platí[4]

$$\log|\underline{\boldsymbol{\mathfrak{A}}}| = R\log|\boldsymbol{\mathfrak{A}}| \tag{B.21}$$

a dále platí

$$\log|\underline{\boldsymbol{\mathfrak{D}}} - \underline{\boldsymbol{\mathfrak{B}}}^T\underline{\boldsymbol{\mathfrak{A}}}^{-1}\underline{\boldsymbol{\mathfrak{B}}}| = \log|R(\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\mathfrak{B}}^T\boldsymbol{\mathfrak{A}}^{-1}\boldsymbol{\mathfrak{B}}) + \boldsymbol{I}|. \tag{B.22}$$

Konečně dostáváme

$$\log|\underline{\boldsymbol{L}}| = R\log|\boldsymbol{\mathfrak{A}}| + \log|\underline{\boldsymbol{\mathfrak{D}}} - \underline{\boldsymbol{\mathfrak{B}}}^T\underline{\boldsymbol{\mathfrak{A}}}^{-1}\underline{\boldsymbol{\mathfrak{B}}}| = \log|R(\boldsymbol{V}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{V} + \boldsymbol{\mathfrak{B}}^T\boldsymbol{\mathfrak{A}}^{-1}\boldsymbol{\mathfrak{B}}) + \boldsymbol{I}|. \tag{B.23}$$

---

[3] $\begin{vmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{vmatrix} = |\boldsymbol{A}|\left|\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B}\right|$ (Press et al., 1992)

[4] $\begin{vmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B} \end{vmatrix} = |\boldsymbol{A}||\boldsymbol{B}|.$

Příloha C

Výpočet věrohodnosti PLDA modelu

Předpokládejme, že máme k dispozici R nahrávek reprezentovaných D-dimenzionálními i-vektory $\boldsymbol{x}_r$, a chceme vyhodnotit věrohodnost PLDA modelu s hyperparametry $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$, tj. $P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta})$, kde podoba $\underline{\boldsymbol{x}}$ vyplývá z rovnosti (4.97) a (4.98) (pro zjednodušení zápisu je vynechána reference na mluvčího s). Rozložení $P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta})$ odpovídá normálnímu rozložení (4.117) a platí tedy

$$\log P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta}) = -\frac{1}{2}RD\log(2\pi) - \frac{1}{2}\log\left|\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}\right| - \frac{1}{2}\tilde{\boldsymbol{x}}^T(\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}})^{-1}\tilde{\boldsymbol{x}}, \tag{C.1}$$

kde $\tilde{\boldsymbol{x}}$ odpovídá (4.100) a $\underline{\boldsymbol{W}}$ vyplývá z rovnosti (4.97) a (4.98). Výpočet (C.1) komplikuje závislost rozměru vektoru $\tilde{\boldsymbol{x}}$ a matice $\underline{\boldsymbol{W}}$ na počtu nahrávek R. Cílem následujících kroků je proto vyjádření vztahu pro výpočet věrohodnosti na základě členů, jejichž rozměr se s počtem nahrávek nemění.

Nejprve proveďme vyjádření $\left|\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}\right|$. Jednoduše lze ukázat, že kovarianční matici $\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}$ je možné vyjádřit následovně

$$\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}} = \underline{\boldsymbol{V}}\underline{\boldsymbol{V}}^T + \underline{\boldsymbol{\Upsilon}}, \tag{C.2}$$

kde

$$\underline{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{V} \\ \vdots \\ \boldsymbol{V} \end{pmatrix} \quad \text{a} \quad \underline{\boldsymbol{\Upsilon}} = \begin{pmatrix} \boldsymbol{\Upsilon} & & \\ & \ddots & \\ & & \boldsymbol{\Upsilon} \end{pmatrix}, \tag{C.3}$$

kde $\boldsymbol{\Upsilon} = \boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{\Sigma}$. Aplikací Sylvesterova teorému pro determinant získáme vyjádření

$$\log|\underline{\boldsymbol{V}}\underline{\boldsymbol{V}}^T + \underline{\boldsymbol{\Upsilon}}| = \log|\underline{\boldsymbol{\Upsilon}}| + \log|\underline{\boldsymbol{I}} + \underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}}|, \tag{C.4}$$

protože matice $\underline{\boldsymbol{\Upsilon}}$ je blokově diagonální, platí

$$\log|\underline{\boldsymbol{\Upsilon}}| = R\log|\boldsymbol{\Upsilon}| \tag{C.5}$$

a

$$\underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}} = R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}. \tag{C.6}$$

Zkombinováním (C.2), (C.4), (C.5) a (C.6) získáme

$$\log|\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}}| = R\log|\boldsymbol{\Upsilon}| + \log|\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}|. \tag{C.7}$$

Nyní se zaměřme na výpočet členu $\underline{\tilde{\boldsymbol{x}}}^T(\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}})^{-1}\underline{\tilde{\boldsymbol{x}}}$ v (C.1). Výpočet inverze kovarianční matice představuje hlavní výpočetní zátěž při vyhodnocení věrohodnosti. Aplikací Woodburyho maticové identity dostáváme

$$\left(\underline{\boldsymbol{V}}\underline{\boldsymbol{V}}^T + \underline{\boldsymbol{\Upsilon}}\right)^{-1} = \underline{\boldsymbol{\Upsilon}}^{-1} - \underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}}(\underline{\boldsymbol{I}} + \underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}})^{-1}\underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}. \tag{C.8}$$

Opět využijeme toho, že matice $\underline{\boldsymbol{\Upsilon}}$ je blokově diagonální, pak je možné zapsat

$$\underline{\tilde{\boldsymbol{x}}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\tilde{\boldsymbol{x}}} = \sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r \tag{C.9}$$

a

$$\underline{\tilde{\boldsymbol{x}}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}}\left(\underline{\boldsymbol{I}} + \underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}}\right)^{-1}\underline{\boldsymbol{V}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\tilde{\boldsymbol{x}}} = \left(\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)\left(\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)^{-1}\left(\sum_{r=1}^{R}\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r\right), \tag{C.10}$$

kde bylo využito (C.6) a také

$$\underline{\tilde{\boldsymbol{x}}}^T\underline{\boldsymbol{\Upsilon}}^{-1}\underline{\boldsymbol{V}} = \sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}. \tag{C.11}$$

Na základě (C.2), (C.8), (C.9) a (C.10) je zřejmé, že platí

$$\begin{aligned}\underline{\tilde{\boldsymbol{x}}}^T(\underline{\boldsymbol{W}}\underline{\boldsymbol{W}}^T + \underline{\boldsymbol{\Sigma}})^{-1}\underline{\tilde{\boldsymbol{x}}} = &\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r \\ &- \left(\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)\left(\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)^{-1}\left(\sum_{r=1}^{R}\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r\right).\end{aligned} \tag{C.12}$$

Konečně dosazením (C.7) a (C.12) dostáváme vyjádření (C.1) ve tvaru

$$\begin{aligned}\log P(\underline{\boldsymbol{x}}|\boldsymbol{\Theta}) = &-\frac{1}{2}RD\log(2\pi) - \frac{1}{2}R\log|\boldsymbol{\Upsilon}| - \frac{1}{2}\log|\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}| \\ &- \frac{1}{2}\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r \\ &+ \frac{1}{2}\left(\sum_{r=1}^{R}\tilde{\boldsymbol{x}}_r^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)\left(\boldsymbol{I} + R\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\boldsymbol{V}\right)^{-1}\left(\sum_{r=1}^{R}\boldsymbol{V}^T\boldsymbol{\Upsilon}^{-1}\tilde{\boldsymbol{x}}_r\right).\end{aligned} \tag{C.13}$$

Protože je zpravidla $D_{chan} < D$, kde D_{chan} hodnost matice $\boldsymbol{\Upsilon}$, je výhodné provedení výpočtu inverze $\boldsymbol{\Upsilon}$ s využitím Woodburyho maticové identity:

$$\boldsymbol{\Upsilon}^{-1} = (\boldsymbol{U}\boldsymbol{U}^T + \boldsymbol{\Sigma})^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{U}\left(\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}. \tag{C.14}$$

Připomeňme, že $\boldsymbol{\Sigma}$ je diagonální matice.