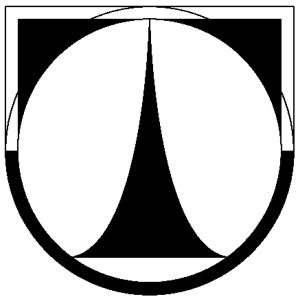
Technická univerzita v Liberci

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

Autoreferát disertační práce

Listopad 2011 Ing. Jan Silovský

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

---

Autoreferát disertační práce

Ing. Jan Silovský

Studijní program:	P2612 Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Ústav informačních technologií a elektroniky Fakulta mechatroniky, informatiky a mezioborových studií Technická univerzita v Liberci
Školitel:	Prof. Ing. Jan Nouza, CSc.

Listopad 2011

Abstrakt

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.

Abstract

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

1 Úvod

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 identifikace ale přináší několik komplikací při vyhodnocení systémů. Naštěstí nic nebrání formulaci úlohy identifikace mluvčího v dané nahrávce jako opakované verifikace všech zvažovaných mluvčích.

Cílem *verifikace* (bývá používán také termín *detekce*) 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.

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. Výhodou textově závislých systému je, že při rozpoznávání dochází k porovnávání shodných fonetických sekvencí. Nutným předpokladem je však dobrovolná spolupráce uživatele se systémem. V řadě praktických aplikací ale není možné řídit obsah promluv (multimediální archivy, telefonní odposlechy). Textově nezávislé systémy nekladou žádné nároky na obsah promluv použitých k zapsání (registraci) uživatele ani k rozpoznávání.

1.1 Současný stav problematiky

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. Toto jsou hlavní témata, kterým je v současnosti věnována pozornost:

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 [1]. Z později navržených metod zmiňme například metody označované jako *feature warping* [2] nebo *feature mapping* [3].

Byla také představena řada metod provádějících normalizaci v prostoru skóre. Největší popularitu si získaly H-norm [4] (tato metoda je dnes již zastoupena jinými technikami), Z-norm nebo T-norm [5].

Metody založené na faktorové analýze Metody založené na faktorové analýze [6, 7] byly také navrženy s cílem omezení vlivu variability akustických podmínek na rozpoznávání. Z pohledu výše uvedené klasifikace se jedná o metody pracující na úrovni modelů mluvčích. Jedním ze základních předpokladů metod založených na faktorové analýzy je možnost separace variability způsobené odlišností hlasů mluvčích a variability způsobené rozdílností akustických podmínek. Není přitom vyžadována žádná kategorizace akustických podmínek.

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 úloze 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. Speciálně pro úlohu rozpoznávání mluvčích byla navržena GLDS (Generalized Linear Discriminant Sequence) jádrová funkce [8] nebo Fisherova jádrová funkce [9]. V současnosti velké množství systémů založených na SVM využívá jádrovou funkci odvozenou na základě aproximace Kullback-Leiblerovy (KL) divergence GMM modelů mluvčích [10]. Důležitým tématem je opět kompenzace variability akustických podmínek. Systémy založené na SVM klasifikátorech nejčastěji využívají techniku nazvanou projekce rušivých atributů (NAP) [11, 12].

Využití automatických přepisů Jedním ze 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 například o fonémy [13] nebo slova [14]. Nevýhodou tohoto přístupu je fragmentace dat, která může zkomplikovat trénování některých modelů kvůli nedostatku dat.

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) [15]. Parametry této transformace je možné následně použít pro charakterizaci mluvčích a provést rozpoznávání např. pomocí SVM [16, 17]. Výhodou této metody přitom je, že nedochází k žádné fragmentaci dat.

Příznaky vyšší úrovně Hlas ovlivňují jak anatomické vlastnosti orgánů podílejících se na tvorbě řeči, které jsou vrozené, tak i naučené vlastnosti získané v průběhu osvojování si jazyka. 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 [18, 19, 20].

Idiolekt představuje charakteristický způsob užívání slov a frází specifických pro konkrétního mluvčího. Využitím idiolektu pro automatické rozpoznávání mluvčích se zabývá například [21]. Mluvčí jsou v tomto případě reprezentováni jazykovými n-gramy odvozenými od univerzálního jazykového modelu, stejného, jaký se používá při rozpoznávání řeči.

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. V současné době je velmi často využívaným způsobem odhadu těchto vah metoda založená na lineární logistické regresi [22].

1.2 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.

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.

2 Vyhodnocování systémů rozpoznávání mluvčích

V rámci vyhodnocení verifikačního systému jsou mu předkládány k posouzení dva druhy *soudů*. V případě *oprávněného soudu* (target trial) řečová data v testované nahrávce skutečně pochází od ověřovaného (detekovaného) mluvčího. V případě *neoprávněného soudu* (impostor trial) pak řeč v testované nahrávce nepochází od ověřovaného mluvčího. Pro každý z předložených soudů je stanoveno *skóre* a v závislosti na porovnání tohoto skóre s verifikačním prahem γ je provedena binární klasifikace soudu. Při tom 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:

$$P_{miss} = \frac{N_{miss}}{N_{tar}} \quad \text{a} \quad P_{FA} = \frac{N_{FA}}{N_{non}}, \tag{1}$$

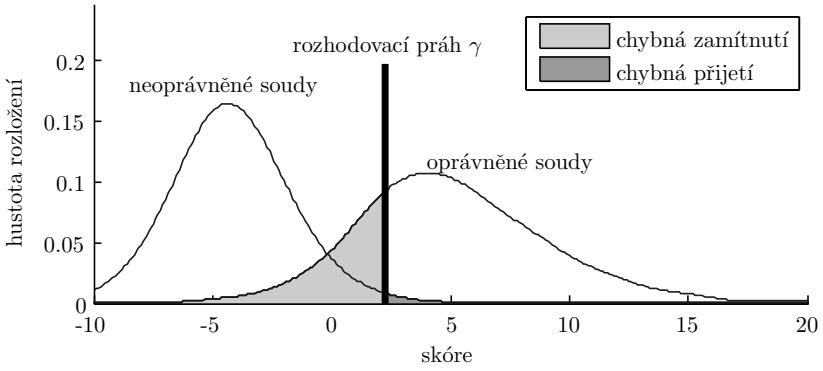
kde N_{miss} je počet soudů, ve kterých došlo k chybnému zamítnutí, a N_{tar} je počet předložených oprávněných soudů. Obdobně N_{FA} je počet soudů, ve kterých došlo k chybnému přijetí, a N_{non} je počet předložených neoprávněných soudů.

Obr. 1 ilustruje typické rozložení skóre určených detektorem mluvčích pro oprávněné soudy $P(s|tar)$ a rozložení skóre pro neoprávněné soudy $P(s|non)$. Rozhodnutí detektoru 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:

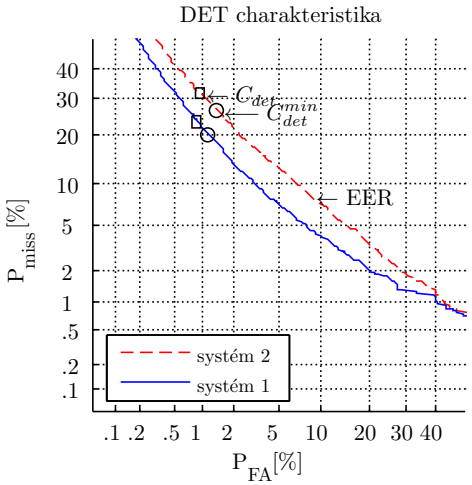
$$P_{miss}(\gamma) = P(s < \gamma|tar) = \int_{-\infty}^{\gamma} P(s|tar)ds \tag{2}$$

$$P_{FA}(\gamma) = P(s \geq \gamma|non) = \int_{\gamma}^{\infty} P(s|non)ds. \tag{3}$$

DET charakteristika 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 DET (Detection Error Trade-off) diagram [23], viz obr. 2. Měřítka jeho os odpovídají kvantilové funkci (inverzní distribuční funkci) standardního normálního rozložení. DET křivka tak vyjadřuje vzájemnou závislost probit(P_{FA}) a probit(P_{miss}).



Obrázek 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}*



Obrázek 2: *Příklad DET charakteristik vyhodnocených pro dva systémy*

Významným bodem DET křivky je průsečík s diagonálou. Tento bod udává hodnotu míry EER (Equal Error Rate) a lze jej definovat jako bod, pro který platí $(P_{FA}; P_{miss}) = (\text{EER}; \text{EER})$. Míra EER tak představuje souhrnnou reprezentaci DET charakteristiky jedinou hodnotou. DET diagramy i hodnota EER ale indikují pouze *rozlišovací schopnosti* detektoru (ve smyslu schopnosti rozlišovat oprávněné a neoprávněné soudy).

Vyhodnocení binární klasifikace soudů Volba rozhodovacího prahu musí reflektovat aplikaci, ve které je detektor použit. Zpravidla specifikujeme aplikaci stanovením hodnot apriorní pravděpodobnosti výskytu oprávněného soudu P_{tar} a pokut chybné přijetí C_{FA} a chybného zamítnutí C_{miss}. Vhodná volba rozhodovacího prahu s ohledem na stanovené parametry zamýšlené aplikace je stejně významnou součástí návrhu detektoru jako vývoj klasifikačních metod. Na základě uvedených aplikačních parametrů je definována očekávaná pokuta detekčních chyb C_{det} jako

$$C_{det}(P_{miss}, P_{FA}) = C_{miss} P_{miss} P_{tar} + C_{FA} P_{FA} (1 - P_{tar}), \tag{4}$$

kde P_{miss} a P_{FA} představují chybové míry stanovené na základě počtu příslušných chybných rozhodnutí detektoru při zvoleném rozhodovacím prahu.

Význam kalibrace detektoru Pokutová funkce C_{det} charakterizuje současně rozlišovací schopnosti i *kalibraci* detektoru. Na základě znalosti referencí evaluačních dat je možné stanovit optimální hodnotu rozhodovacího prahu, pro kterou je C_{det} minimální, na základě kritéria

$$C_{det}^{min} = \min_{-\infty \leq \gamma \leq \infty} C_{miss} P_{miss}(\gamma) P_{tar} + C_{FA} P_{FA}(\gamma)(1 - P_{tar}). \tag{5}$$

Při vyhodnocení tak dochází k přizpůsobení se hodnotám skóre poskytovaných systémem a na rozdíl od C_{det} nejsou hodnocena pevná rozhodnutí detektoru.

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. Pokud ovšem výstup detektoru odpovídá *logaritmu poměru věrohodností* LLR, není nutné požadovat po detektoru, aby poskytoval pevná rozhodnutí o klasifikaci soudů. Rozhodovací práh lze pak totiž určit nezávisle na použitém detektoru na základě parametrů aplikace. Na základě provedených rozhodnutí lze pak stanovit míry P_{miss} a P_{FA} a vyhodnotit příslušnou hodnotu C_{det}. Tímto způsobem však provedeme vyhodnocení kvality poskytovaných LLR pouze z pohledu zvoleného pracovního bodu. Po systému však požadujeme, aby poskytoval správné hodnoty v celém možném rozsahu pracovních bodů. Autoři [24] tedy definují nové kritérium pro souhrnné *aplikačně nezávislé* vyhodnocení detektoru mluvčích na základě poskytovaných hodnot LLR

$$C_{llr} = C_0 \int_{-\infty}^{\infty} C_{det}\left(P_{miss}(\gamma), P_{FA}(\gamma), \gamma\right) d\gamma, \tag{6}$$

kde $C_0 > 0$ je normalizační konstanta. Protože C_{det} hodnotí rozhodnutí detektoru o klasifikaci soudů, reflektuje funkce C_{llr} jak rozlišovací schopnosti, tak kvalitu kalibrace detektoru. Kritérium C_{llr} je možné interpretovat jako celkovou chybovou míru vyhodnocenou přes celý rozsah možných pracovních bodů. Definice C_{llr} podle (6) je dobře interpretovatelná z hlediska významu, integrál ale představuje komplikaci z hlediska praktického výpočtu. Naštěstí je však možné ukázat [25], že C_{llr} lze na základě vyhodnocení provedených pro předložená evaluační data analyticky vyjádřit v uzavřené formě jako

$$C_{llr}(\{\mathcal{L}'_t\}) = \frac{1}{2 \log 2} \left(\frac{1}{N_{tar}} \sum_{t \in \text{tar}} \log\left(1 + e^{-\mathcal{L}'_t}\right) + \frac{1}{N_{non}} \sum_{t \in \text{non}} \log\left(1 + e^{\mathcal{L}'_t}\right) \right), \tag{7}$$

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.

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 uvedených kritérií, tak i proces hledání transformační funkce provádějící převod z hodnot skóre na hodnoty LLR.

Logistická regrese Pravděpodobně nejčastěji používaný způsob kalibrace detektorů mluvčích je založen na lineární logistické regresi [26, 27]. Snahou je nalézt parametry afinní transformace, které by zajistily optimální hodnotu účelové funkce. Transformační funkce má podobu

$$\mathcal{L}'_t = \alpha_0 + \alpha_1 s_t, \tag{8}$$

kde s_t je skóre určené detektorem pro verifikační soud t, $\{\alpha_0$ a $\alpha_1\}$ představují parametry transformační funkce a $\mathcal{L}'_t$ je kalibrované skóre ve formátu LLR. Protože je transformační funkce ryze monotónní, nedochází k ovlivnění rozlišovacích schopností detektoru. Účelová funkce je definována jako [27]

$$\begin{aligned} Q(\boldsymbol{\lambda}, P_{tar}) =& \frac{P_{tar}}{N_{tar}} \sum_{t \in tar} \log_2 \left(1 + e^{(-w_{lr}(s_t, \boldsymbol{\lambda}) - \operatorname{logit} P_{tar})}\right) \\ &+ \frac{(1 - P_{tar})}{N_{non}} \sum_{t \in non} \log_2 \left(1 + e^{(w_{lr}(s_t, \boldsymbol{\lambda}) + \operatorname{logit} P_{tar})}\right). \end{aligned} \tag{9}$$

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 opět lineární logistická regrese [27]. Výsledné skóre stanovené na základě kombinace skóre určených K dílčími systémy pomocí lineární logistické regrese představuje vážený součet

$$\mathcal{L}'_t = \alpha_0 + \sum_{k=1}^{K} \alpha_k s_{t,k}, \tag{10}$$

kde $s_{t,k}$ je skóre určené systémem k pro soud t a $\alpha_0, \ldots, \alpha_K$ parametry, které 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 (9).

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. 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 $P(y)$. 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 Směsi Gaussových rozložení

V úloze automatického rozpoznávání mluvčích (i řeči) je rozložení akustických příznakových vektorů obvykle modelováno směsí Gaussových (normálních) rozložení (GMM). Ta představuje model tvořený váženou kombinací C Gaussových rozložení charakterizovaných vahou w_c, vektorem středních hodnot $\boldsymbol{\mu}_c$ a kovarianční maticí $\boldsymbol{\Sigma}_c$, kde $c = 1, \ldots, C$. Logaritmus věrohodnosti pro sekvenci F-dimenzionálních příznakových vektorů $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ je vyjádřen jako

$$P(\boldsymbol{O}|\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \sum_{c=1}^{C} w_c \frac{1}{\sqrt{(2\pi)^F |\boldsymbol{\Sigma}_c|}} \exp\left(-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{o} - \boldsymbol{\mu}_c)\right). \tag{11}$$

Odhad parametrů GMM $\boldsymbol{\theta} = \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\}_{c=1}^{C}$ spadá do obecné třídy problémů chybějících dat (missing data problem). Pravděpodobně nejčastěji využívaným způsobem odhadu parametrů GMM je EM algoritmus. Nejčastěji používanými metodami odhadu parametrů jsou metoda *maximální věrohodnosti* (ML) a metoda *maximální aposteriorní pravděpodobnosti* (MAP).

UBM-GMM systém Využívá metodu maximálně věrohodného odhadu parametrů GMM pro natrénování univerzálního hlasového modelu 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. Pro odvození modelů zapisovaných mluvčích je využívána *relevanční* MAP adaptace UBM modelu. Významným parametrem této metody je relevanční faktor τ, který v rámci MAP odhadu řídí váhu ML odhadu parametrů stanoveného na základě předložených trénovacích dat mluvčího a parametrů původního UBM modelu (roste s rostoucím τ).

V procesu rozpoznávání je skóre pro předložený verifikační soud stanoveno na základě průměrné hodnoty logaritmu poměru věrohodností GMM ověřovaného mluvčího s a UBM (ten zastupuje neoprávněné mluvčí)

$$s = \frac{1}{T} \left(\log P(\boldsymbol{O}|\boldsymbol{\theta}_s) - \log P(\boldsymbol{O}|\boldsymbol{\theta}_{UBM})\right), \tag{12}$$

kde $\boldsymbol{O} = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ je sekvence příznakových vektorů reprezentující testovanou nahrávku.

4.2 Faktorová analýza

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 [28]

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{z} + \boldsymbol{\epsilon}, \tag{13}$$

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 $\boldsymbol{\Sigma}$.

GMM supervektor V rámci této práce je pojmem *supervektor* označován vektor vytvořený zřetězením vektorů středních hodnot komponent GMM. Za předpokladu F-dimenzionálních příznakových vektorů a modelu s C komponentami bude rozměr *GMM supervektoru* roven CF.

4.2.1 Sdružená faktorová analýza

Model *sdružené faktorové analýzy* (JFA) [29] představuje generativní model GMM supervektorů. Na rozdíl od modelu (13) nepředstavují supervektory přímo pozorovatelné proměnné, ty jsou představovány sekvencemi F-rozměrných příznakových vektorů. Supervektor $\boldsymbol{m}_{r,s}$ příslušný nahrávce r mluvčího s je modelován generativním procesem

$$\boldsymbol{m}_{r,s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s + \boldsymbol{U}\boldsymbol{w}_{r,s}. \tag{14}$$

Tento model je uvažován složený ze dvou složek. První část $\boldsymbol{s}_s = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{D}\boldsymbol{z}_s$ charakterizuje mluvčího s a je předpokládána jako neměnná pro všechny jeho nahrávky. Druhá část $\boldsymbol{c}_{r,s} = \boldsymbol{U}\boldsymbol{w}_{r,s}$ charakterizuje akustické podmínky, které jsou pro každou nahrávku specifické. Supervektor $\boldsymbol{\mu}$ je nezávislý na mluvčím i nahrávce. Sloupce matice $\boldsymbol{V}$ definují báze (nízkodimenzionálního) podprostoru charakteristického pro variabilitu hlasů mluvčích a sloupce matice $\boldsymbol{U}$ pak báze (nízkodimenzionálního) podprostoru charakteristického pro variabilitu akustických podmínek. Diagonální matice $\boldsymbol{D}$ charakterizuje variabilitu hlasů mluvčích, která není popsána v podprostoru vymezeném $\boldsymbol{V}$. Vektory $\boldsymbol{y}_s$, $\boldsymbol{w}_{r,s}$ a $\boldsymbol{z}_s$ jsou tvořené *faktory*, které odpovídají variabilitě charakterizované příslušnými zátěžovými maticemi. Pro faktory (které reprezentují skryté proměnné) je předpokládáno standardní normální apriorní rozložení $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Rozložení supervektorů $\boldsymbol{s}_s$ pak odpovídá normálnímu rozložení $P(\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{s}_s|\boldsymbol{\mu}, \boldsymbol{V}\boldsymbol{V}^T + \boldsymbol{D}^2)$ a rozložení supervektorů $\boldsymbol{m}_{r,s}$ při známém $\boldsymbol{s}_s$ normálnímu rozložení $P(\boldsymbol{m}_{r,s}|\boldsymbol{s}_s) = \mathcal{N}(\boldsymbol{m}_{r,s}|\boldsymbol{s}_s, \boldsymbol{U}\boldsymbol{U}^T)$.

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 $\mathcal{N}(\boldsymbol{o}_t|\boldsymbol{m}_{r,s,c}, \boldsymbol{\Sigma}_c)$.

Rozpoznávání Skóre pro předložené verifikační soudy je pak stanoveno jako logaritmus poměru věrohodností

$$s = \frac{1}{T}\left(\log P(\boldsymbol{O}|\boldsymbol{s}_s, \boldsymbol{\Lambda}) - \log P(\boldsymbol{O}|\boldsymbol{\mu}, \boldsymbol{\Lambda})\right), \tag{15}$$

kde ověřovaný mluvčí je reprezentován supervektorem $\boldsymbol{s}_s$ a UBM model supervektorem $\boldsymbol{\mu}$.

4.2.2 I-vektory

Autoři [30] 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

$$\boldsymbol{m}_{r,s} = \boldsymbol{\mu} + \boldsymbol{T}\boldsymbol{x}_{r,s}, \tag{16}$$

kde $\boldsymbol{T}$ je matice hodnosti D_{ivec} a $\boldsymbol{x}_{r,s}$ je D_{ivec}-dimenzionální náhodný vektor s apriorním rozložením $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Vektor $\boldsymbol{x}_{r,s}$ je označován jako *i-vektor*. Model (16) 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.

I-vektor představuje reprezentaci nahrávky ve formě MAP odhadu $\boldsymbol{x}_{r,s}$ stanoveného pro tuto nahrávku. Proces výpočtu i-vektorů lze pak interpretovat jako součást parametrizace řečového signálu.

Rozpoznávání [30] ukazuje, že klasifikaci pomocí i-vektorů lze provést jednoduše na základě jejich kosinové vzdálenosti. Skóre hodnotící hypotézu, že mluvčí v nahrávce reprezentované i-vektorem $\boldsymbol{x}_1$ je shodný s mluvčím v nahrávce reprezentované i-vektorem $\boldsymbol{x}_2$, je stanoveno jako

$$s(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle}{\|\boldsymbol{x}_1\| \, \|\boldsymbol{x}_2\|}. \tag{17}$$

Další možností je aplikace pravděpodobnostní lineární diskriminační analýzy (PLDA).

Kompenzace variability akustických podmínek Pro kompenzaci variability akustických podmínek v prostoru i-vektorů lze použít např. metodu normalizace kovariance uvnitř tříd (WCCN), metodu projekce rušivých atributů (NAP) nebo lineární diskriminační analýzu (LDA).

Lineární diskriminační analýza Lineární diskriminační analýza (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 rozměr $\boldsymbol{A}$ je $n \times m$.

4.2.3 Pravděpodobnostní lineární diskriminační analýza

Pravděpodobnostní lineární diskriminační analýza (PLDA) představuje metodu založenou na faktorové analýze, navrženou původně pro úlohu rozpoznávání tváří [31]. 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

$$\boldsymbol{x}_{r,s} = \boldsymbol{\mu} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{w}_{r,s} + \boldsymbol{\epsilon}_{r,s}. \tag{18}$$

Tento model je podobně jako JFA model uvažován složený ze dvou složek. Vektor $\boldsymbol{x}_{r,s}$ ale na rozdíl od GMM supervektoru $\boldsymbol{m}_{r,s}$ v případě JFA modelu (14) reprezentuje přímo pozorovatelná data. 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. 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 faktory $\boldsymbol{y}_s$ a $\boldsymbol{w}_{r,s}$ je opět předpokládáno standardní normální apriorní rozložení $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. 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 $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{V}, \boldsymbol{U}, \boldsymbol{\Sigma}\}$.

Rozpoznávání Proces rozpoznávání je možné formulovat jako Bayesovský výběr modelů [31]. Konkrétně pro úlohu detekce mluvčích předpokládejme, že máme k dispozici dvě sady nahrávek $\underline{\boldsymbol{O}}_1$ 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. 3), lze věrohodnost tohoto modelu rozepsat jako

$$P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_2) = P(\underline{\boldsymbol{O}}_1|\mathcal{M}_2)P(\underline{\boldsymbol{O}}_2|\mathcal{M}_2). \tag{19}$$

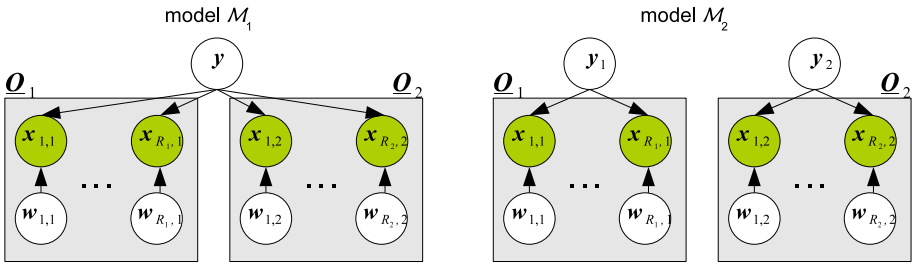
Výstup detektoru mluvčích pak odpovídá vyhodnocení logaritmu poměru věrohodností těchto modelů, tj.

$$\begin{aligned} s &= \log \frac{P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_1)}{P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_2)} \\ &= \log P(\underline{\boldsymbol{O}}_1, \underline{\boldsymbol{O}}_2|\mathcal{M}_1) - \log P(\underline{\boldsymbol{O}}_1|\mathcal{M}_2) - \log P(\underline{\boldsymbol{O}}_2|\mathcal{M}_2). \end{aligned} \tag{20}$$

5 SVM klasifikátory

Převážně používaným typem diskriminativních klasifikátorů v úloze rozpoznávání mluvčích jsou SVM (Support Vector Machine) klasifikátory.

SVM klasifikátory představují metodu strojového učení určenou pro binární klasifikaci dat. Důležitou součástí SVM klasifikátorů je *jádrová funkce* $K(.,.)$, která provádí transformaci z původního prostoru příznakových vektorů do prostoru transformovaných příznaků (typicky vyšší dimenze) a



Obrázek 3: *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*

umožňuje převést původně lineárně neseparovatelnou úlohu na úlohu lineárně separovatelnou. Klasifikace testovaného vektoru $\boldsymbol{x}$ do třídy $+1$ nebo -1 je pak provedena na základě hodnoty $\text{sign}(f(\boldsymbol{x}))$, kde

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \rho, \tag{21}$$

kde L je počet *podpůrných vektorů* $\boldsymbol{x}_i$ definujících SVM model, y_i představuje referenční klasifikaci $\boldsymbol{x}_i$ a konečně α_i jsou hodnoty Lagrangeových multiplikátorů příslušné podpůrným vektorům, které jsou společně s hodnotou ρ stanoveny trénovacím algoritmem. S ohledem na možnost provedení následné kalibrace je upřednostňován výstup SVM klasifikátoru formou skóre $f(\boldsymbol{x})$ před binární klasifikací. Kalibrace skóre je nezbytná, protože toto skóre nemá žádnou pravděpodobnostní/věrohodnostní interpretaci.

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. Jádrová funkce odvozená na základě (horní meze) symetrické KL divergence GMM λ_1 a λ_2 (získaných MAP adaptací UBM pro nahrávky reprezentované sekvencemi příznakových vektorů $\boldsymbol{O}_1$ a $\boldsymbol{O}_2$) má pak podobu [10]

$$K_{KL-GMM}(\boldsymbol{O}_1, \boldsymbol{O}_2) = \sum_{c=1}^{C} \left(\sqrt{w_c}\boldsymbol{\Sigma}_c^{-\frac{1}{2}}\boldsymbol{\mu}_{1,c}\right)^T \left(\sqrt{w_c}\boldsymbol{\Sigma}_c^{-\frac{1}{2}}\boldsymbol{\mu}_{2,c}\right). \tag{22}$$

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 [16] usiluje právě o potlačení tohoto vlivu. 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 reprezentován přímo koeficienty MLLR transformace. Zpravidla je přitom využívána lineární jádrová funkce.

Kompenzace variability akustických podmínek pro SVM Dvěma základními metodami používanými pro potlačení vlivu variability akustických podmínek v systémech založených na SVM jsou metody označované jako projekce rušivých atributů (NAP) [12] a normalizace kovariance uvnitř tříd (WCCN) [32].

6 Rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů

Rozpoznávání mluvčích v záznamech televizních a rozhlasových (TVR) pořadů 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 [33]. V rámci tohoto systému bylo později analyzováno několik základních řešení založených jak na generativním, tak diskriminativním přístupu [34]. Konkrétně byly implementovány tři systémy:

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

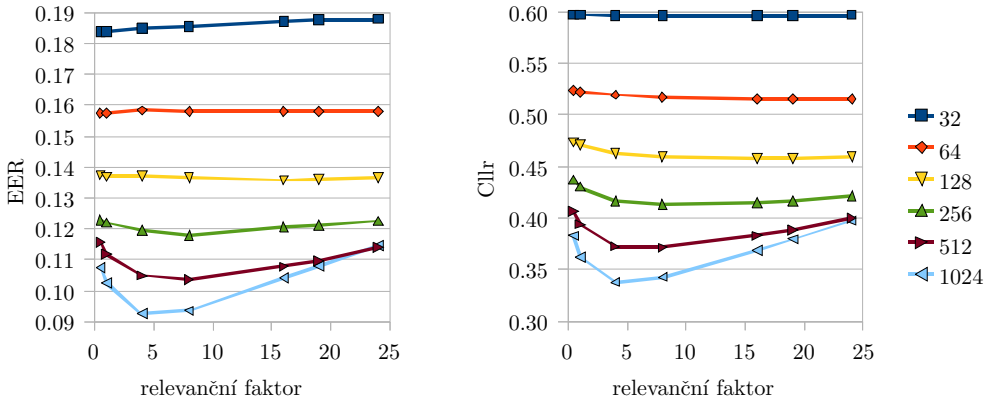
Jedním z přínosů této části práce je návrh vlastní evaluační databáze sloužící k jednotnému porovnání implementovaných řešení.

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. Testovací data byla 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). Celkem bylo



(a) (b)

Obrázek 4: *Vliv počtu komponent a hodnoty relevančního faktoru na úspěšnost rozpoznávání v případě UBM-GMM systému*

v rámci první sady provedeno 24 966 verifikačních soudů, z toho bylo 2 067 soudů oprávněných. 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.

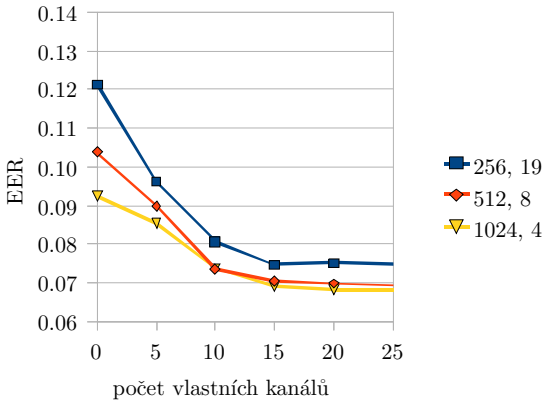
6.2 Experimentální vyhodnocení systémů

6.2.1 UBM-GMM systém

Základními parametry UBM-GMM systému 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.

Nejprve byl analyzován vliv hodnoty relevančního faktoru a počtu Gaussovských komponent. Obr. 4 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. Lze to přisuzovat shodnému charakteru dat v obou testovacích sadách (vytvořených rozdělením jedné databáze). 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



Obrázek 5: *Vliv počtu vlastních kanálů při aplikaci metody ECA v případě UBM-GMM systému*

supervektor $\boldsymbol{s}_{r,s}$, tvořený zřetězením vektorů středních hodnot komponent modelu, které jsou navíc *normalizovány* příslušnými směrodatnými odchylkami, je adaptován dle vztahu

$$\boldsymbol{m}_{r,s} = \boldsymbol{s}_s + \boldsymbol{U}\boldsymbol{w}_{r,s}, \tag{23}$$

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. Složky vektoru $\boldsymbol{w}_{r,s}$ jsou označovány jako kanálové faktory a jsou stanoveny na základě maximalizace věrohodnosti

$$P(\boldsymbol{O}_{r,s}|\boldsymbol{s}_s + \boldsymbol{U}\boldsymbol{w}_{r,s})P(\boldsymbol{w}_{r,s}). \tag{24}$$

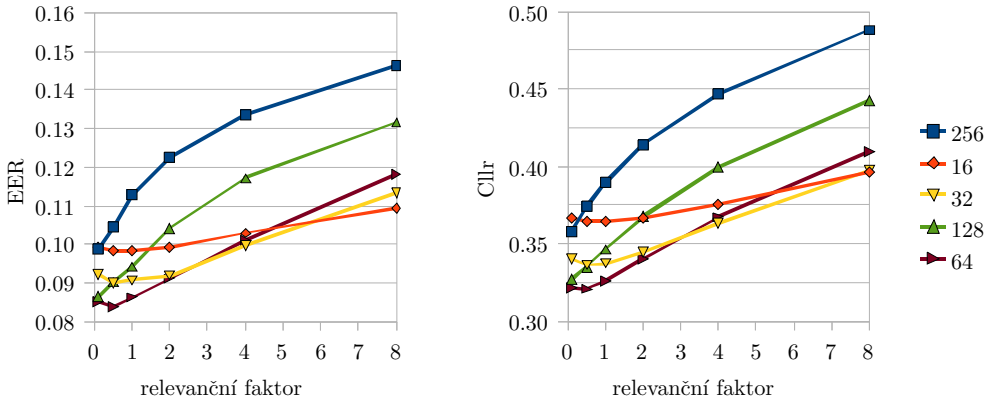
Vlastní kanály jsou stanoveny metodou analýzy hlavních komponent (PCA) obdobně jako v případě metody NAP [27]. Apriorní rozložení faktorů $\boldsymbol{w}_{r,s}$ v (24) je předpokládáno jako standardní normální, tedy $P(\boldsymbol{w}_{r,s}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Metoda ECA byla aplikována pro systémy s 256, 512 a 1024 komponentami. Obr. 5 shrnuje nejlepší dosažené výsledky vzhledem k testovaným hodnotám relevančního faktoru pro systémy s různým počtem komponent.

6.2.2 GMM-SVM systém

Tento SVM systém využívá jádrovou funkci odvozenou na základě KL divergence GMM modelů odvozených MAP adaptací UBM. Základní parametry GMM-SVM systému jsou shodné jako v případě UBM-GMM systému, přestože princip obou klasifikátorů je odlišný. Parametry jsou tedy a) počet

[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.



(a) (b)

Obrázek 6: *Vliv počtu komponent a hodnoty relevančního faktoru na úspěšnost rozpoznávání v případě GMM-SVM systému*

Gaussovských komponent, b) hodnota relevančního faktoru pro MAP adaptaci a c) dimenze NAP podprostoru.

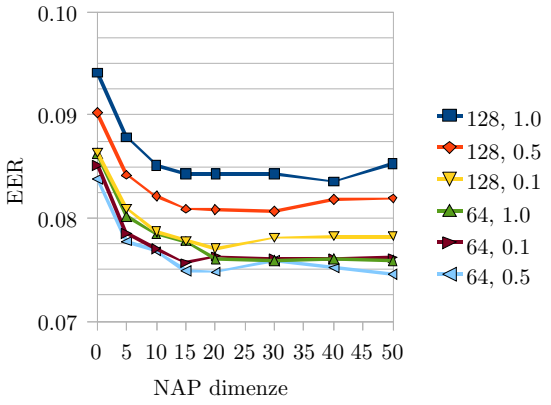
Nejprve byl opět analyzován vliv hodnoty relevančního faktoru a počtu Gaussovských komponent pro systém nevyužívající kompenzaci variability akustických podmínek, viz obr. 6. Z výsledků je patrný velký vliv jak počtu Gaussovských komponent, tak hodnoty relevančního faktoru. Přitom rostoucí počet komponent nevede nutně ke zvýšení úspěšnosti rozpoznávání. Zajímavý je 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 již velmi blíží ML odhadu parametrů modelu. 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 výsledků lze usuzovat, že diskriminativní povaze SVM vyhovuje, nejsou-li GMM supervektory příslušné mluvčím příliš koncentrovány blízko jednoho supervektoru (příslušného UBM modelu).

Obr. 7 zobrazuje výsledky systémů s aplikací metody NAP pro kompenzaci variability akustických podmínek.

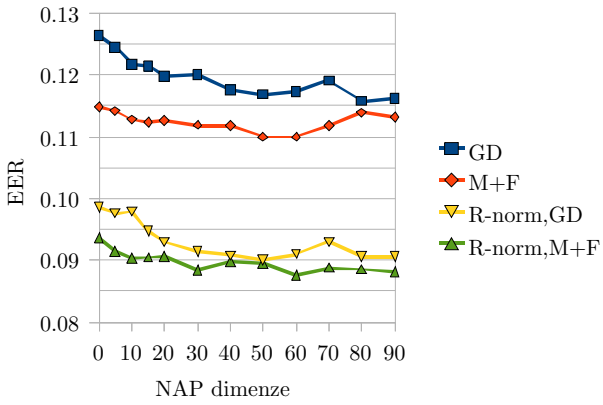
6.2.3 MLLR-SVM systém

MLLR-SVM systém je založen na SVM klasifikaci na základě MLLR adaptačních koeficientů. 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.

Pro SVM klasifikaci byla použita lineární jádrová funkce. První systém používal SVM vektory tvořené MLLR transformačními parametry s 1 560 koeficienty (byly použity 39-dimenzionální akustické příznakové vektory). Použitý systém pro rozpoznávání řeči používá modely specifické pro



Obrázek 7: *Efekt aplikace metody NAP na výsledky GMM-SVM systému v závislosti na dimenzi NAP podprostoru*



Obrázek 8: *Výsledky dosažené v případě MLLR-SVM systému*

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. Možným řešením, umožňujícím odstranit problém identifikace pohlaví, je využití MLLR transformací odvozených pro oba GD modely [17], tedy mužský (M) a ženský (F). GD modely jsou navíc trénovány 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. 8 poskytuje přehled dosažených výsledků. Výrazné zvýšení úspěšnosti rozpoznávání přináší aplikace metody R-norm (rank normalization) [17].

Výsledky 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.

Tabulka 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	-

6.3 Shrnutí výsledků

Tab. 1 shrnuje sledované metriky a výpočetní čas[2] nutný pro zpracování testovacích dat (v násobku reálného času, výpočet akustických příznaků MFCC není zahrnut) pro nejúspěšnější konfigurace porovnávaných systémů. Z provedených srovnání vyplývá, že UBM-GMM systém poskytuje nejvyšší úspěšnost. Výhodou GMM-SVM systému je především jeho rychlost. Nejnižší úspěšnosti dosáhl MLLR-SVM systém. Pravděpodobným důvodem je velmi krátká délka nahrávek, která neumožňuje robustní odhad MLLR transformačních parametrů.

7 Systémy TUL pro NIST evaluaci rozpoznávání mluvčích 2010

V letech 2008 a 2010 se autor zúčastnil [35, 36] evaluací systémů pro rozpoznávání mluvčích (SRE) pořádaných americkým Úřadem pro standardy a technologii (NIST). Účast autora v roce 2008 představovala první zapojení Laboratoře počítačového zpracování řeči TUL v sérii těchto evaluací.

V rámci účasti TUL v NIST SRE 2010 byly implementovány tři systémy:

- 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. Experimentální vyhodnocení systémů v rámci jejich vývoje bylo prováděno podle základního zadání evaluace pořádané v roce 2008.

[2]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 2: *Vyhodnocení systémů pro evaluační podmínky základního zadání NIST SRE 2008 specifikované v závislosti na typu dat použitých pro trénování modelů a rozpoznávání (např. „tel – tel“)*

	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

7.1 Vyhodnocení systémů na datech NIST SRE 2008

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.

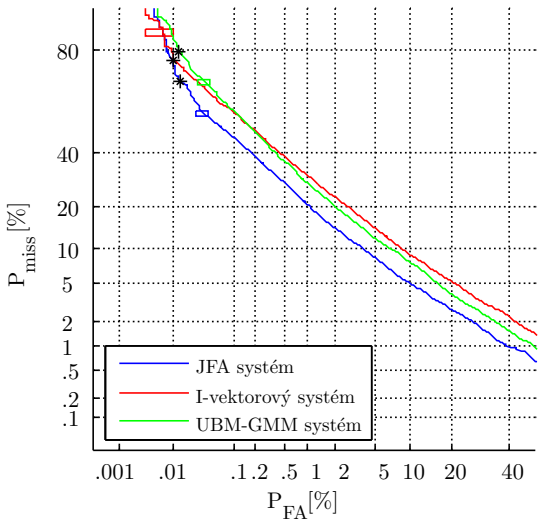
Tab. 2 shrnuje dosažené výsledky pro srovnatelné evaluační podmínky ročníků 2008 a 2010 (označení „tel“ je použito pro záznamy telefonních hovorů a „int“ pro mikrofonní záznamy interview tazatele a dotazovaného v záznamové místnosti, tyto záznamy jsou prováděny souběžně na 14 mikrofonech). V případě ročníku 2008 byly kromě anglických nahrávek použity nahrávky telefonních hovorů v dalších 25 jazycích. Hodnota C_{norm}^{min} v tab. 2 byla stanovena s aplikačními parametry specifikovanými pro NIST SRE 2008 [37], tedy $C_{miss} = 10$, $C_{FA} = 1$ a $P_{tar} = 0{,}01$. Z uvedených výsledků je dobře patrné, že nejhorší úspěšnosti dosahují systémy zpravidla v evaluační podmínce zahrnující rozdílný typ záznamu pro trénování modelů a pro rozpoznávání. Při souhrnném posouzení lze konstatovat, že nejlepších výsledků dosáhl JFA systém.

7.2 Vyhodnocení systémů na datech NIST SRE 2010

Výsledky dosažené pro evaluační data NIST SRE 2010, která nebyla v rámci vývoje systémů viděná, shrnuje tab. 3 (označení „mtel“ je použito pro mikrofonní záznamy telefonních hovorů zaznamenané v místnosti volajícího). Hodnota C_{norm}^{min} byla v tomto případě stanovena s novými aplikačními parametry specifikovanými pro ročník 2010 [38], tedy $C_{miss} = 1$, $C_{FA} = 1$ a $P_{tar} = 0{,}001$. Pro možnost porovnání s výsledky dosaženými v rámci vývojových experimentů je v tab. 3 uváděna i hodnota $C_{norm,old}^{min}$, která byla stanovena pro předešlé hodnoty aplikačních parametrů.

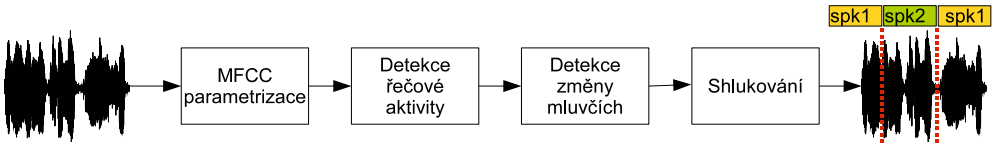
Tabulka 3: *Vyhodnocení systémů pro všechny verifikační soudy základního zadání NIST SRE 2010*

	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



Obrázek 9: *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*

Z porovnání výsledků uvedených v tab. 2 a 3 je patrné, že v případě tří ze čtyř srovnatelných evaluačních podmínek došlo ke snížení úspěšnosti rozpoznávání. Jedinou výjimkou je podmínka „int – tel“.



Obrázek 10: *Schéma diarizačního systému*

Přestože je úspěšnost (soudě dle hodnoty míry EER a normalizované pokutové funkce C_{norm}^{min}) dosažená na evaluačních datech NIST SRE 2008 pro většinu evaluačních podmínek základního zadání blízká úspěšnosti nejlepších systémů navržených účastníky této evaluace, nebyly na datech NIST SRE 2010 dosaženy příliš dobré výsledky. 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 v NIST SRE 2010. Obr. 9 zobrazuje srovnání DET charakteristik vyhodnocených pro implementované systémy pro evaluační podmínku „int-int“, 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.

8 Metody založené na faktorové analýze v úloze diarizace mluvčích

Úlohu diarizace mluvčích 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.

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.

8.1 Schéma diarizačního systému

Základní návrh vyvinutého diariazačního systému je standardní [39] a sestává ze tří hlavních modulů, jak ilustruje obr. 10. 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.

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.

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. V našem systému je používána tradiční míra pro srovnání uvedených hypotéz odvozená na základě Bayesovského informačního kritéria (BIC) [40].

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. 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.

Za základní přístup pro stanovení vzdálenosti shluků lze považovat přístup založený na BIC kritériu. Tento přístup je zde použit v rámci *referenčního systému*.

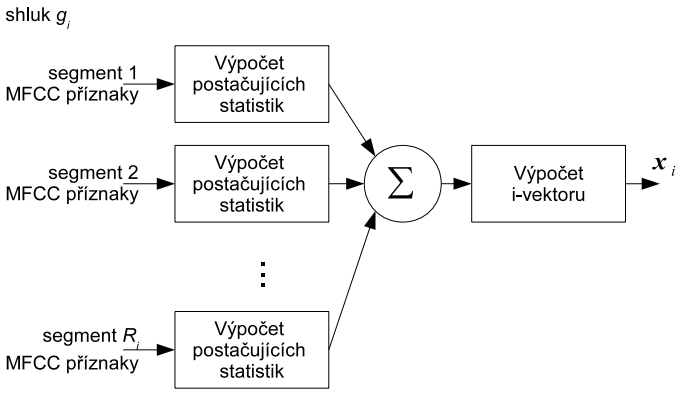
8.2 Metody pro shlukování mluvčích

V práci byly navrženy tři přístupy pro shlukování mluvčích založené na metodách běžně používaných v úloze rozpoznávání mluvčích. Všechny 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) [41]. 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 [40]:

$$\Delta BIC(g_1, g_2) = (N_1 + N_2) log\,|\mathbf{\Sigma}| - N_1 log\,|\mathbf{\Sigma}_1| - N_2 log\,|\mathbf{\Sigma}_2| - \alpha P, \tag{25}$$



Obrázek 11: *Popis způsobu odvození reprezentace shluků v případě CDS systému*

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

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right) log(N_1 + N_2). \tag{26}$$

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).

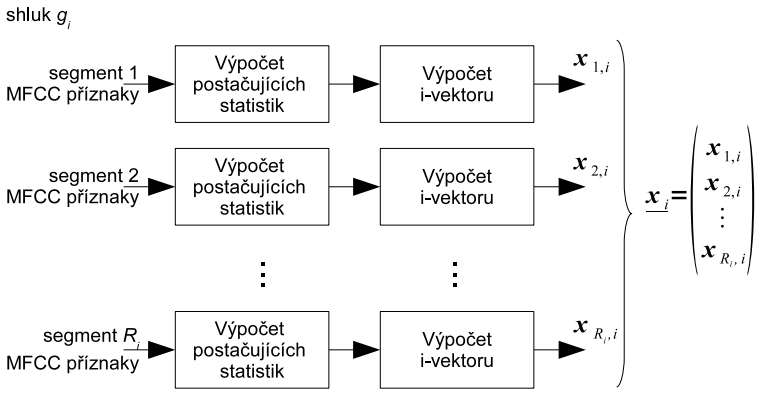
8.2.2 Shlukování na základě kosinové vzdálenosti i-vektorů

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. 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 (17).

Při sloučení dvou shluků je 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 (viz obr. 11).

V průběhu shlukování je v každém cyklu algoritmu sloučen pár shluků s nejvyšší hodnotou kosinové vzdálenosti. Ukončovací podmínka je splněna pokud je v daném cyklu maximální hodnota nižší než práh λ stanovený na vývojových datech.

Pro potlačení variability akustických podmínek je využívána lineární diskriminační analýza (LDA). Při trénování transformační matice byly jednotlivé třídy tvořeny všemi segmenty mluvčího v rámci daného audiozáznamu.



Obrázek 12: *Popis způsobu odvození reprezentace shluků v případě systému založeného na vícesložkovém PLDA přístupu*

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 (18).

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 obr. 3). V průběhu shlukování jsou v každém cyklu sloučeny shluky s nejvyšším hodnotou logaritmu poměru věrohodností stanovenou na základě (20). 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 daného shluku, tedy stejným způsobem jako v případě CDS přístupu (viz obr. 11).

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. 12).

8.2.4 Dvoufázové shlukování

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). Ve druhé fázi je aplikováno shlukování využívající jednu z metod založených na i-vektorech.

8.3 Specifikace evaluační databáze a vyhodnocení systémů

Experimentální vyhodnocení bylo provedeno s využitím databáze COST278 [42]. Tato databáze zahrnuje záznamy televizních zpravodajských pořadů v 9 evropských jazycích.

Pro účely našeho vyhodnocení byly definovány tři různé sady dat. První sada obsahovala celkem 11,5 hod. dat. 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.). 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.). Celková délka těchto záznamů je 6,3 hod.

8.3.1 Evaluační metriky

Systémy pro diarizaci mluvčích jsou obvykle porovnávány na základě míry DER (Diarization Error Rate). Míru DER je možné vyjádřit na základě součtu $DER = SPKE + FA + MISS$, 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. Protože všechny systémy sdílejí shodný modul pro detekci řečové aktivity a segmentaci mluvčích, je jako primární metrika použita míra SPKE.

8.4 Experimentální vyhodnocení systémů

8.4.1 Parametrizace řečového signálu

Všechny moduly využívají MFCC akustické příznaky. V rámci segmentace nebyla aplikována CMS normalizace a efekt její aplikace v rámci shlukování bude diskutován zvlášť pro jednotlivé přístupy.

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.

8.4.2 Vyhodnocení činnosti detektorů řeči a změny mluvčích

Na testovací sadě byly vyhodnoceny míry FA 0,8 % a MISS 3,2 %. Vyšší míra MISS je způsobena nepřesností referenčních anotací. 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. Výsledkem jsou však poměrně krátké segmenty. Průměrná délka segmentů je 3,6 s (zatímco u referenčních anotací je to 17,8 s).

8.4.3 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 (25). Bylo zjištěno, že systémy, pro které byla použita nenulová hodnota ukončovacího prahu λ, optimalizovaná na základě vývojových dat, dosahují lepších výsledků než systémy s nulovou hodnotou prahu.

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 %.

Dále byl zkoumán vliv aplikace CMS normalizace. Bylo přitom zjištěno, že aplikace CMS normalizace způsobuje výrazné zhoršení míry SPKE, zejména pro vyšší hodnoty penalizační váhy α.

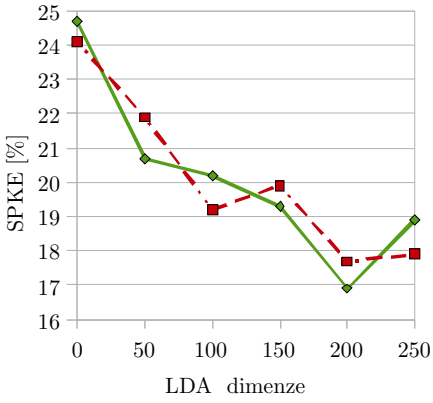
8.4.4 Výsledky systému založeného na CDS přístupu

Obr. 13 ilustruje vliv dimenze LDA transformace pro systémy pracující s 300 a 400-dimenzionálními i-vektory. Nulová dimenze LDA v obr. 13 odpovídá systému bez aplikace LDA transformace.

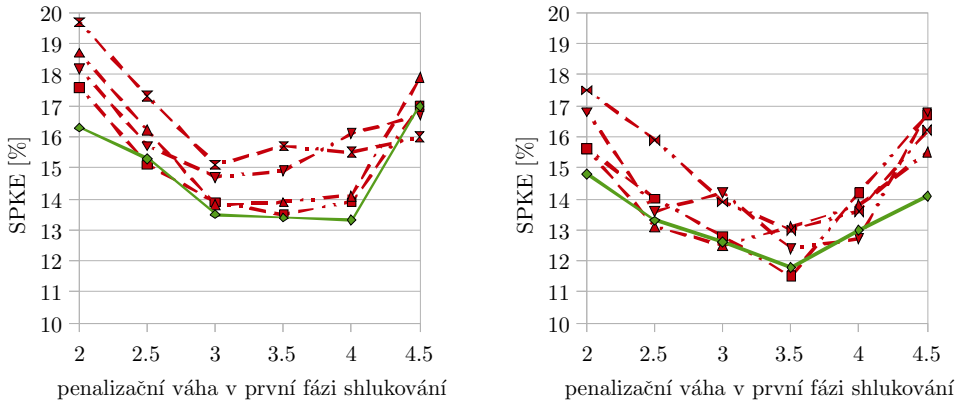
Bez aplikace LDA poskytuje systém založený na CDS přístupu výsledky srovnatelné s referenčním systémem. Odstranění nežádoucí variability prostřednictvím LDA však přináší významné zlepšení úspěšnosti. Nejnižší hodnota SPKE 16,9 % byla dosažena systémem pracujícím s 400-dimenzionálními i-vektory a 200-dimenzionální LDA transformací. Na rozdíl od systému založeného na BIC shlukování nebyl pozorován výrazný efekt aplikace CMS.

Dalšího snížení míry SPKE bylo dosaženo použitím dvoufázového shlukování. Počet shluků vstupujících do druhé fáze ovlivňuje hodnota penalizační váhy α použitá při vyhodnocení ΔBIC kritéria v první fázi. Čím nižší je tato hodnota, tím dříve je první fáze přerušena a tím více shluků vstupuje do druhé fáze. Obr. 14 dokládá vliv hodnoty penalizační váhy α použité v první fázi.

Aplikace CMS normalizace ve druhé fázi shlukování přináší snížení míry SPKE. S ohledem na výsledky dosažené pro různé hodnoty penalizační váhy α, lze jako nejlepší hodnotit systém pracující



Obrázek 13: *Efekt dimenze LDA transformace v případě CDS systémů pracujících s 300- (červená čára) a 400-dimenzionálními (zelená čára) i-vektory*



(a) bez aplikace CMS ve druhé fázi (b) s aplikací CMS ve druhé fázi

Obrázek 14: *Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování. Nejlepší konfigurace CDS přístupu je zvýrazněna plnou čarou*

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).

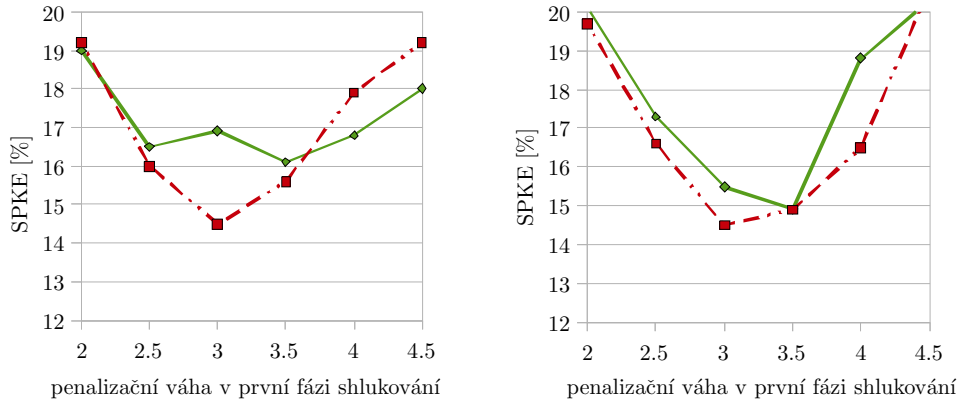
8.4.5 Výsledky systému založeného na jednosložkovém PLDA přístupu

Tab. 4 shrnuje vybrané výsledky dosažené v rámci testovaných konfigurací jednosložkového PLDA přístupu. 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).

Výrazného snížení SPKE bylo dosaženo v případě dvoufázového shlukování, jak dokládá obr. 15. Nejnižší míry SPKE 14,5 % (41,5% rel. zlepšení) dosáhl systém s 300-dimenzionálními i-vektory ($\mathrm{rk}(\boldsymbol{V}) = 150$,$\mathrm{rk}(\boldsymbol{U}) = 150$) v případě použití penalizační váhy ΔBIC kritéria $\alpha = 3{,}0$ v první fázi shlukování. Aplikace CMS normalizace ve druhé fázi shlukování založené na jednosložkovém PLDA přístupu přináší ve většině případů pouze nepatrné snížení míry SPKE.

Tabulka 4: *Výsledky systémů využívajících jednosložkové PLDA shlukování*

			bez aplikace CMS		s aplikací CMS	
rk($\boldsymbol{T}$)	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) bez aplikace CMS ve druhé fázi (b) s aplikací CMS ve druhé fázi

Obrázek 15: *Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování při aplikaci jednosložkového PLDA přístupu 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*)*

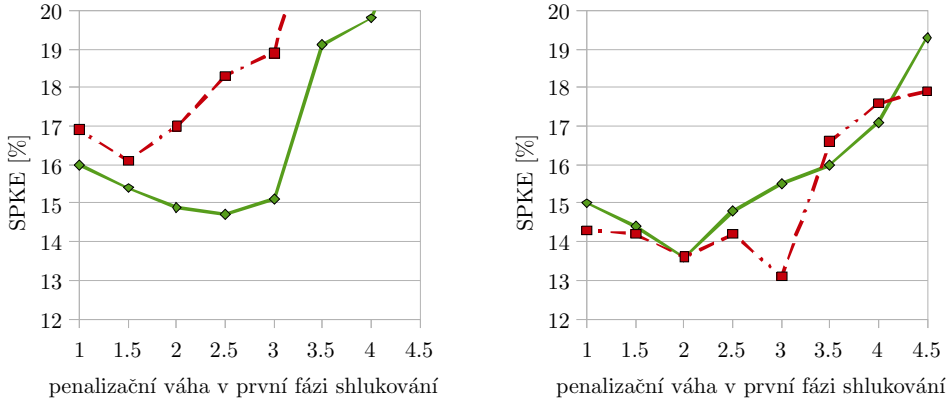
8.4.6 Výsledky systému založeného na vícesložkovém PLDA přístupu

Nejlepší výsledky dosažené pro vícesložkový PLDA přístup jsou uvedeny v tab. 5. 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. 16. Ve srovnání s předešlými přístupy je v tomto případě dosaženo pouze mírného zlepšení míry SPKE. 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. Zejména pro systém s 300-dimenzionálními i-vektory, který

Tabulka 5: *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



(a) bez aplikace CMS ve druhé fázi (b) s aplikací CMS ve druhé fázi

Obrázek 16: *Efekt penalizační váhy ΔBIC kritéria použité v první fázi dvoufázového shlukování při aplikaci vícesložkového PLDA přístupu 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*)*

pro penalizační váhu $\alpha = 3{,}0$ dosáhl míry SPKE 13,1 % (47,2% rel. zlepšení).

8.5 Shrnutí výsledků

Tab. 6 zprostředkovává přehled nejlepších výsledků dosažených pro prezentované přístupy při aplikaci jednofázového shlukování včetně výpočetní náročnosti uváděné jako násobek reálné délky zpracovávaných záznamů[3].

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.

Jednoznačně pozitivní přínos měla aplikace dvoufázového shlukování, kdy je v první fázi provedeno „předshlukování“ založené na BIC kritériu. Vybrané výsledky shrnuje tab. 7. 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

[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 6: *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

Tabulka 7: *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

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.

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.

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.

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 těchto 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

[1] Hynek Hermansky and Nelson Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[2] Jason Pelecanos and Sridha Sridharan. Feature Warping for Robust Speaker Verification. In *2001: A Speaker Odyssey - The Speaker Recognition Workshop*, Crete, Greece, 2001.

[3] Douglas A. Reynolds. 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.

[4] Douglas A. Reynolds. Comparison of Background Normalization Methods for Text-Independent Speaker Verification. In *5th European Conference on Speech Communication and Technology - Eurospeech '97*, Rhodes, Greece, 1997.

[5] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1–3):42–54, 2000.

[6] Patrick Kenny and Pierre Dumouchel. Disentangling speaker and channel effects in speaker verification. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2004*, pages 37–40, Montreal, Canada, 2004.

[7] Robbie Vogt, Brendan Baker, and Sridha Sridharan. Modelling Session Variability in Text-Independent Speaker Verification. In *9th European Conference on Speech Communication and Technology - Interspeech 2005*, Lisboa, Portugal, 2005.

[8] William M. Campbell. Generalized Linear Discriminant Sequence Kernels for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002*, Orlando, FL, USA, 2002.

[9] Shai Fine, Jiri Navratil, and Ramesh A. Gopinath. 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.

[10] William M. Campbell, Douglas E. Sturim, Douglas A. Reynolds, and Alex Solomonoff. 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, 2006.

[11] Alex Solomonoff, Carl Quillen, and William M. Campbell. Channel Compensation for SVM Speaker Recognition. In *Odyssey: The Speaker and Language Recognition Workshop*, Toledo, Spain, 2004.

[12] Alex Solomonoff, William M. Campbell, and Ian Boardman. Advances in channel compensation for SVM speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2005*, Philadelphia, PA, USA, 2005.

[13] Alex Park and Timothy J. Hazen. ASR Dependent Techniques for Speaker Identification. In *International Conference on Spoken Language Processing*, Denver, CO, USA, 2002.

[14] Douglas E. Sturim, Douglas A. Reynolds, Robert B. Dunn, and Thomas F. Quatieri. Speaker Verification using Text-Constrained Gaussian Mixture models. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2002*, Orlando, FL, USA, 2002.

[15] Mark J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12(2):75–98, 1998.

[16] Andreas Stolcke, Luciana Ferrer, Sachin S. Kajarekar, Elizabeth Shriberg, and Anand Venkataraman. MLLR transforms as features in speaker recognition. In *9th European Conference on Speech Communication and Technology - Interspeech 2005*, Lisboa, Portugal, 2005.

[17] Andreas Stolcke, Luciana Ferrer, and Sachin S. Kajarekar. Improvements in MLLR-transform-based speaker recognition. In *IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.

[18] Andre G. Adami, Radu Mihaescu, Douglas A. Reynold, and John J. Godjirey. Modeling Prosodic Dynamics for Speaker Recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP 2003*, Hong Kong, 2003.

[19] Barbara Peskin, Jiri Navratil, Joy Abramson, Douglas Jones, David Klusacek, Douglas A. Reynolds, and Bing Xiang. 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.

[20] Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Anand Venkataraman, and Andreas Stolcke. Modeling prosodic feature sequences for speaker recognition. *Speech Communication*, 46(3–4):455–472, 2005.

[21] George Doddington. Speaker Recognition based on Idiolectal Differences between Speakers. In *7th European Conference on Speech Communication and Technology - Eurospeech 2001*, Aalborg, Denmark, 2001.

[22] Niko Brummer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech and Language*, 20(2):230–275, 2006.

[23] Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. The DET Curve In Assessment Of Detection Task Performance. In *Proc. Eurospeech '97*, pages 1895–1898, Rhodes, Greece, 1997.

[24] David A. van Leeuwen and Niko Brummer. An Introduction to Application-Independent Evaluation of Speaker Recognition Systems. *Lecture Notes in Computer Science*, 4343/2007:330–353, 2007.

[25] Niko Brummer. *Measuring, refining and calibrating speaker and language information extracted from speech*. PhD thesis, University of Stellenbosch, October 2010.

[26] Stephane Pigeon, Pascal Druyts, and Patrick Verlinde. Applying logistic regression to the fusion of the nist'99 1-speaker submissions. *Digital Signal Processing*, 10(1-3):237–248, 2000.

[27] Niko Brummer, Lukas Burget, Jan Cernocky, Ondrej Glembek, Frantisek Grezl, Martin Karafiat, David A. van Leeuwen, Pavel Matejka, Petr Schwarz, and Albert Strasheim. 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*, 15(7):2072–2084, 2007.

[28] C. M. Bishop. *Pattern Recognition and Machine Learning*. 2006.

[29] Patrick Kenny. 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.

[30] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(4):788 –798, May 2011.

[31] S.J.D. Prince and J.H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings ICCV 2007*, Rio de Janeiro, Brazil, October 2007.

[32] Andrew O. Hatch and Andreas Stolcke. Generalized linear kernels for one-versus-all classification: Application to speaker recognition. In *Proc. IEEE ICASSP*, volume 5, pages 585–588, Toulouse, May 2006.

[33] Jan Silovsky and Jan Nouza. Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream. *Radioengineering*, 15(3):42–48, 2006.

[34] Jan Silovsky, Petr Cerva, and Jindrich Zdansky. Comparison of Generative and Discriminative Approaches for Speaker Recognition with Limited Data. *Radioengineering*, 18(3):307–316, 2009.

[35] Jan Silovsky. TUL System for the NIST 2008 Speaker Recognition Evaluation. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

[36] Jan Silovsky. TUL NIST 2010 SRE System Description. In *Proc. 2010 NIST Speaker Recognition Evaluation Workshop*, 2010.

[37] 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.

[38] 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.

[39] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1557 –1565, sept. 2006.

[40] S.S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proceedings 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pages 127–132, 1998.

[41] Jan Silovsky, Jan Prazak, Petr Cerva, Jindrich Zdansky, and Jan Nouza. PLDA-based clustering for speaker diarization of broadcast streams. In *INTERSPEECH'11*, pages 2909–2912. ISCA, August 2011.

[42] An Vandecatseye, Jean-Pierre Martens, Joao Neto, Hugo Meinedo, Carmen Garcia-Mateo, Javier Dieguez, France Mihelic, Janez Zibert, Jan Nouza, Petr David, Matus Pleva, Anton Cizmar, Harris Papageorgiou, and Christina Alexandris. The COST278 pan-European broadcast news database. pages 873–876, 2004.

Seznam vlastních publikací

[1] Petr Cerva, Jan Nouza, and Jan Silovsky. 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.

[2] Petr Cerva, Jindrich Zdansky, Jan Silovsky, and Jan Nouza. 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, pages 277–284, Berlin, Heidelberg, 2008. Springer-Verlag.

[3] Petr Cerva, Karel Palecek, Jan Silovsky, and Jan Nouza. An investigation into VTLN for improved transcription of czech broadcast programs. In *53rd International IEEE Symposium ELMAR-2011*, pages 201–204, September 2011.

[4] Petr Cerva, Karel Palecek, Jan Silovsky, and Jan Nouza. Using unsupervised feature-based speaker adaptation for improved transcription of spoken archives. In *INTERSPEECH*, pages 2565–2568. ISCA, August 2011.

[5] Josef Chaloupka, Jan Nouza, Jindrich Zdansky, Petr Cerva, Jan Silovsky, and Martin Kroul. Voice technology applied for building a prototype smart room. In *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 104–111. Springer-Verlag, Berlin, Heidelberg, 2009.

[6] Ramon Lopez-Cozar, Zoraida Callejas, Martin Kroul, Jan Nouza, and Jan Silovsky. 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, pages 617–624, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] Ramon Lopez-Cozar, Jan Silovsky, and David Griol. Enhancement of spoken dialogue systems by means of user emotion recognition. *Spanish Journal On Natural Language Processing (SEPLN)*, 45:191–198, September 2010. ve španělšině.

[8] Ramon Lopez-Cozar, Jan Silovsky, and David Griol. 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, pages 281–288. Association for Computational Linguistics, 2010.

[9] Ramon Lopez-Cozar, Jan Silovsky, and Martin Kroul. Enhancement of emotion detection in spoken dialogue systems by combining several information sources. *Speech Communication*, 53:1210–1228, November 2011.

[10] Jan Nouza, Josef Chaloupka, Jindrich Zdansky, Jan Silovsky, Martin Kroul, and Zbynek Mader. Voice Controlled Center for Homes of Motor-Handicapped Persons. In *Speech and Computer International Conference - Specom 2007*, pages 714–719, Moscow, Russia, 2007.

[11] Jan Nouza, Jan Silovsky, Jindrich Zdansky, Petr Cerva, Martin Kroul, and Josef Chaloupka. Czech-to-Slovak Adapted Broadcast News Transcription System. In *9th Annual Conference of the International Speech Communication Association - Interspeech 2008*, pages 2683–2686, Brisbane, Australia, 2008.

[12] Jan Nouza and Jan Silovsky. Fast Keyword Spotting in Telephone Speech. *Radioengineering*, 18(4):665–670, 2009.

[13] Jan Nouza, Jindrich Zdansky, Petr Cerva, and Jan Silovsky. Challenges in speech processing of slavic languages (case studies in speech recognition of czech and slovak). In Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt, editors, *COST 2102 Training School*, volume 5967 of *Lecture Notes in Computer Science*, pages 225–241. Springer, 2009.

[14] Jan Nouza and Jan Silovsky. 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, pages 377–384, Berlin, Heidelberg, 2010. Springer-Verlag.

[15] Jan Nouza, Karel Blavka, Marek Bohac, Petr Cerva, Jindrich Zdansky, Jan Silovsky, and Jan Prazak. Voice technology to enable sophisticated access to historical czech radio audio archive. In *MM4CH 2011*, volume 247 of *Communications in Computer and Information Science*, pages 27–38. Springer, 2011.

[16] Jan Prazak and Jan Silovsky. Comparison of segmentation and clustering methods for speaker diarization of broadcast stream audio. In *Analysis of Verbal and Nonverbal Communication and Enactment*, volume 6800 of *Lecture Notes in Computer Science*, pages 214–222. Springer, 2011.

[17] Jan Prazak and Jan Silovsky. Speaker diarization using plda-based speaker clustering. In *The 6th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems*, pages 347–350, September 2011.

[18] Jan Silovsky and Jan Nouza. Speech, Speaker and Speaker's Gender Identification in Automatically Processed Broadcast Stream. *Radioengineering*, 15(3):42–48, 2006.

[19] Jan Silovsky and Petr Cerva. Study on Speaker Recognition aided Broadcast Stream Transcription. In *16th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2006.

[20] Jan Silovsky, Petr Cerva, and Jindrich Zdansky. Text-Independent Speaker Verification Supported by ASR. In *17th Czech-German Workshop Speech Processing*, Prague, Czech Republic, 2007.

[21] Jan Silovsky. TUL System for the NIST 2008 Speaker Recognition Evaluation. In *Proc. 2008 NIST Speaker Recognition Evaluation Workshop*, 2008.

[22] Jan Silovsky, Petr Cerva, and Jindrich Zdansky. Comparison of Generative and Discriminative Approaches for Speaker Recognition with Limited Data. *Radioengineering*, 18(3):307–316, 2009.

[23] Jan Silovsky, Petr Cerva, and Jindrich Zdansky. MLLR Transforms Based Speaker Recognition in Broadcast Streams. In Anna Esposito and Robert Vich, editors, *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, volume 5641 of *Lecture Notes in Computer Science*, pages 423–431. Springer Berlin / Heidelberg, Prague,Czech Republic, 2009.

[24] Jan Silovsky and Petr Cerva. 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.

[25] Jan Silovsky. TUL NIST 2010 SRE System Description. In *Proc. 2010 NIST Speaker Recognition Evaluation Workshop*, 2010.

[26] Jan Silovsky, Petr Cerva, and Jindrich Zdansky. Assessment of speaker recognition on lossy codecs used for transmission of speech. In *53rd International IEEE Symposium ELMAR-2011*, pages 205–208, September 2011.

[27] Jan Silovsky, Jan Prazak, Petr Cerva, Jindrich Zdansky, and Jan Nouza. PLDA-based clustering for speaker diarization of broadcast streams. In *INTERSPEECH'11*, pages 2909–2912. ISCA, August 2011.

Ing. Jan Silovský

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

Autoreferát disertační práce

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

Náklad 10 výtisků

Listopad 2011