Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména §60 - školní dílo.

Beru na vědomí, že Technická univerzita v Liberci TUL nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

Užiji-li diplomovou práci nebo poskytnu-li licenci k jejímu užití, jsem si vědom povinnosti informovat o této skutečnosti TUL; v tomto případě má TUL právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím diplomové práce a konzultantem.

Datum

Podpis

Poděkování

Rád bych poděkoval panu Prof. Ing. Janu Nouzovi, CSc. za cenné rady, připomínky, ochotu a čas věnovaný konzultacím mé diplomové práce a také panu Ing. Petru Davidovi za uvedení do problematiky rozpoznávání mluvčích.

Resumé

Diplomová práce se zabývá návrhem systému rozpoznávání mluvčích. Po stručném úvodu do problematiky rozpoznávání mluvčích a přehledu současného stavu je vysvětlena souvislost hlasových charakteristik a použitých melovských kepstrálních příznaků (MFCC). Práce se dále soustřeďuje na přístupy k reprezentaci modelů mluvčích v textově nezávislých systémech, vektorovou kvantizaci (VQ) a zejména směsi Gaussovských rozložení (GMM). Hlavní motivací práce je vytvoření modulu rozpoznávání mluvčích integrovatelného do systému kompletního automatického přepisu televizních a rozhlasových pořadů a tomu odpovídá formulace požadavků a návrh řešení.

Byla provedena řada experimentů zabývajících se mimo jiné různými způsoby vyhodnocení identifikace a verifikace mluvčích, porovnáním vlivu různých metod estimace parametrů modelů, nebo významem detekce hlasových framů. Na jejich základě se podařilo nalézt vhodnou kombinaci metod a jejich nastavení. Při poměrně vysokém počtu 306 referenčních řečníků se podařilo dosáhnout úspěšnosti rozpoznávání více než 81 %.

Klíčová slova: rozpoznávání mluvčích, televizní a rozhlasové pořady, směsi Gaussovských rozložení, vektorová kvantizace

Summary

The diploma thesis deals with design of speaker recognition system. After brief introducing to the field of speaker recognition and a summarization of the current state, the relationship between voice characteristics and mel cepstral coefficients (MFCC), used in proposed system, is explained. An attention of this thesis is then concentrated on approaches used to speaker modeling in text-independent systems, vector quantization (VQ) and particularly Gaussian mixture models (GMM). Main aim is to built a speaker recognition module integrable to the system for fully automated transcription of broadcast programmes, which impacts the demands and the proposed solution.

Performed experiments compare different approaches to evaluation of speaker identification and verification, different methods for estimation of model parameters or signification of voice frame detection. Accordingly to the acquired results, the appropriate combination of methods and their configuration was chosen. Using quite large population of 306 reference speakers, the recognition rate exceeded the level of 81 %.

Keywords: speaker recognition, broadcast programmes, Gaussian mixture models, vector quantization

OBSAH

1. ÚVOD

Rozpoznávání osob na základě jejich hlasu je jednou z biometrických metod. Biometrika představuje moderní způsob ověřování totožnosti a jednoznačné identifikace osob. K tomuto účelu využívá jedinečných fyziologických vlastností a charakteristik chování člověka. Mezi nejčastěji používané patří otisky prstů, tvar ruky, obličej, obraz sítnice, obraz duhovky a také hlas. V praxi lze najít využití například v různých přístupových systémech pro budovy, sklady, trezory, počítače nebo bankovní služby. Systémy založené na rozpoznávání hlasu jsou vhodné pro vzdálenou autentizaci a často uváděným nasazením je tak přístup k informačním systémům pomocí telefonu. Nespornou výhodou v porovnání s jinými technikami pro identifikaci osob využívajícími různé bezpečnostní karty, klíče a hesla je, že u biometrických charakteristik nehrozí odcizení, ztráta, nebo zapomenutí. Síla biometriky není v utajení informací používaných pro autentizaci, ale v jedinečnosti těchto informací.

Rozpoznávání mluvčích lze rozdělit na dvě základní úlohy – identifikaci a verifikaci. Cílem verifikace je ověřit řečníkem prohlašovanou totožnost na základě jeho hlasu. Neznámý řečník nejprve udá tvrzení o své totožnosti (například zadáním identifikačního čísla, nebo pomocí bezpečnostní karty) a verifikační systém následně musí rozhodnout, zda promluva neznámého řečníka je dostatečně podobná hlasu osoby, za kterou se vydává. Pokud se neznámý řečník vydává za někoho jiného, než skutečně je, označuje se jako podvodník. Úlohu identifikace lze dále rozdělit na identifikaci v uzavřené a otevřené množině. V obou případech není předkládáno žádné tvrzení o totožnosti řečníka a cílem je stanovit, kdo z množiny referenčních řečníků[1] promluvu vyslovil, případně rozhodnout, že nikdo z referenčních řečníků promluvu nevyslovil (v případě identifikace v otevřené množině). Identifikace v otevřené množině probíhá ve dvou fázích. Nejprve se provede identifikace v uzavřené množině, kdy se předpokládá, že nahrávka neznámého řečníka patří někomu ze skupiny referenčních řečníků. Totožnost, která je výsledkem této první fáze, pak vstupuje do verifikační fáze. Jejím cílem je ověřit, zda hlas v nahrávce neznámého řečníka je dostatečně podobný prohlašované totožnosti. V případě kladného rozhodnutí je neznámý řečník identifikován jako řečník, jehož totožnost je výsledkem identifikační fáze. V opačném případě je výsledkem prohlášení, že neznámý řečník není nikdo z referenčních řečníků. Podrobný popis základních úloh lze nalézt v [2].

[1] Řečníci, kteří jsou systému známí, tj. systém má uloženy reprezentace jejich hlasu (modely) v databázi.

Úlohy jsou dále rozlišovány na základě požadavků kladených na nahrávky použité pro trénování systému a rozpoznávání a na prostředí, ve kterém jsou tyto nahrávky pořizovány. V textově závislých systémech jsou promluvy použité k trénování systému a rozpoznávání omezeny na konkrétní slova, nebo fráze. V textově nezávislých systémech nejsou na trénovací a vyhodnocované nahrávky kladeny žádné požadavky týkající se jejich obsahu. Přechodný stupeň mezi oběma kategoriemi představují systémy s omezeným slovníkem. Slovník může například obsahovat číslice, ze kterých jsou vybírána testovací slova nebo fráze (například řetězec číslic).

Na základě použitého přístupu k reprezentaci modelů mluvčích lze provést rozdělení na systémy založené na metodách využívajících vzorové reprezentace nebo na metodách využívajících pravděpodobnostní modely.

1.1 Historie a současný stav problematiky

Velmi důkladně je dosavadní vývoj systémů pro automatické rozpoznávání řečníků a řeči zpracován v [7], nicméně pro úplnost jsou v této podkapitole zmíněny hlavní oblasti zájmu a dosažené pokroky zejména z pohledu použitých příznaků a modelů použitých pro reprezentaci mluvčích.

První snahy o automatické rozpoznávání mluvčích byly učiněny v 60. letech 20. století. Tyto experimenty byly založeny na použití bank filtrů a na měření korelace spektrogramů. Později byly banky filtrů nahrazeny analýzou formantů. Stejně jako dnes bylo již od počátku snahou získat příznaky, které by měly malou vnitřní variabilitu pro daného mluvčího a současně velkou variabilitu mezi jednotlivými mluvčími tak, aby byly schopné tyto mluvčí co nejlépe odlišit. Za účelem získání příznaků nezávislých na fonetickém obsahu byla provedena řada experimentů s různými statistickými a prediktivními parametry, jednalo se například o okamžitou kovariační matici spektra, průměrování autokorelace přes delší úseky, histogramy spektra a základní frekvence, lineární prediktivní koeficienty a jiné příznaky. Současně probíhal také vývoj textově závislých systémů, které díky vhodnému zarovnání dvou obsahově shodných promluv v časové oblasti provádí porovnání příznaků odpovídajících podobným fonetickým podmínkám a byly tak schopny poskytnout mnohem lepší výkon oproti textově nezávislým systémům.

Na konci 70. let bylo představeno použití kepstrálních příznaků (v kombinaci s jejich první a druhou derivací) za účelem zvýšení robustnosti rozpoznávacího systému vůči rušení

způsobenému přenosem po telefonním kanálu. Později se tyto příznaky staly standardem nejenom v systémech rozpoznávání mluvčích, ale také řeči.

Zřejmě největším pokrokem učiněným v průběhu 80. let bylo využití skrytých Markovských modelů (HMM) pro textově závislé i nezávislé systémy. Systémy založené na HMM využívaly modely mluvčích odvozené z posloupnosti několika slov, samostatných slov nebo fonémů. Kromě použití skrytých Markovských modelů, které patří mezi parametrické pravděpodobnostní modely, byla věnována pozornost také neparametrickým modelům, reprezentovaným vektorovou kvantizací (VQ). Při zjišťování vlivu počtu stavů a příslušného počtu komponent HMM (v úloze textově nezávislého rozpoznávání byly použity ergodické HMM, tzn. že jsou povoleny libovolné přechody mezi všemi stavy) byla dokázána invariantnost úspěšnosti rozpoznávání vůči počtu stavů modelu, zatímco úspěšnost silně závisela na počtu použitých komponent. Byl tak navržen jednostavový HMM, označovaný jako směs Gaussovských rozložení (GMM).

Jedním z témat, kterému byla věnována pozornost v 90. letech, byla normalizace skóre používaného k rozhodování verifikačních systémů. Za účelem vytvoření vhodné normalizace s ohledem na co nejnižší výpočetní náročnost byla navržena řada metod využívající univerzální modely okolí (UBM), nebo tzv. kohorty. Důležitým rysem je také snaha o využití systémů rozpoznávání mluvčích ve spolupráci se systémy rozpoznávání řeči, jejichž výkon se při použití modelů adaptovaných pro daného řečníka zvyšuje. V posledních letech je patrná snaha o využití vyšších příznaků, ty vycházejí například z ohodnocení typicky používaných slov, výslovnosti, prozódie, atd.

2. Řeč

2.1 Jak řeč vzniká?

Odlišnosti lidských hlasů mají dvojí původ. Hlas ovlivňují fyzické charakteristiky, související s anatomií hlasového ústrojí, a naučené charakteristiky, související s individuálními návyky, které si člověk osvojí v průběhu života.

Celý dýchací systém a všechny svaly od břicha až po nos hrají určitou roli při tvorbě zvuků, avšak nejdůležitější jsou hrtan, jazyk, rty a měkké patro. Hrtan je umístěn přibližně ve středu krku na vrcholu průdušnice v zadní části hrdla. Jde v podstatě o úsek vzduchové trubice s vnějším chrupavčitým obalem. Hrtan obsahuje hlasivky, které jsou upnuty na jedné straně k štítné chrupavce (součást ohryzku) a na druhé k páru pohyblivých arytenoidních (hlasivkových) chrupavek. Vibrace hlasivek vzniká v průběhu řeči tehdy, když se hlasivková štěrbina zúží a vzduch z plic je vytlačován hlasivkami a hrtanem. Tento jev se nazývá fonace a kmitání hlasivek vytváří kvaziperiodické pulsy stlačeného vzduchu, které budí vokální trakt. Frekvence oscilací se označuje jako základní hlasivkový tón a závisí na délce, velikosti a napětí hlasivek. Muži mají běžně větší hrtan a delší, ochablejší hlasivky, proto mají hlubší hlas než ženy a děti. Fonace je příčinou vzniku znělých zvuků, mezi ty patří všechny samohlásky a některé souhlásky.

Neznělé zvuky vznikají třením, turbulencí nebo explozí výdechového proudu vzduchu o tzv. koartikulační orgány – zuby, jazyk, rty. Z uvedeného je patrné, že souhlásky mohou být znělé i neznělé. To závisí na tom, zda je vzduch vycházející z plic modulován hlasivkami a nebo prochází volně. Souhlásky lze dále dělit na frikativy a explozívy. Frikativy vznikají protlačováním proudu vzduchu štěrbinami, které jsou vytvořeny v některém místě hlasového ústrojí artikulačními orgány, čímž vznikají vzduchové turbulence s produkováním charakteristického šumu. Při tvorbě explozív dochází k úplnému uzavření hlasového ústrojí (jazykem, rty, nebo patrem), vytvoření tlaku vzduchu a rychlému uvolnění. Výsledkem toho je ticho (trvající po dobu, která odpovídá hromadění tlaku) následované krátkou explozí šumu. Rezonanční kvality různých dutin v ústech a dýchacím systému dodávají hlasu individualitu. Například přesné vyslovení tzv. nosových hlásek závisí na volné rezonanci v nose. Efekt vznikající při mluvení se stlačeným nosem dokazuje, že vzduchový prostor nosu dodává řeči zvučnost a jasnost.

Protože lidé mají rozdílné tvary nosu, hrudníku a úst, mají různě znějící hlasy. Svůj vliv má také průdušnice, trubice spojující plíce s hrtanem, typicky 12 cm dlouhá s 2 cm v průměru.

Pokud hlasivky vibrují, vznikají nad i pod nimi rezonance a právě tyto podhlasivkové rezonance jsou velmi závislé na vlastnostech průdušnice a tím pádem na daném řečníkovi. Ostatní fyzické vlastnosti závislé na mluvčím jsou dechový objem (maximální objem vzduchu, který je člověk schopen vydechnout po maximálním nadechnutí), maximální délka fonace (maximální doba, po kterou může být hláska udržována), fonační koeficient (podíl dechového objemu a maximální délky fonace) a tok vzduchu hlasivkami. Nicméně, protože zvuk není na toku vzduchu nijak závislý, je obtížné tyto parametry odvodit pouze ze znalosti akustického signálu.

Jak bylo zmíněno v úvodu této podkapitoly, řeč ovlivňují také naučené charakteristiky. Těmi jsou například tempo řeči (projevuje se různou délkou konkrétních slov i v časovém kolísání uvnitř slov), prozódie (často se používá reprezentace pomocí posloupnosti sady symbolů, které odpovídají průběhu základního hlasivkového tónu a energie v čase), dialekt (ve spektru se projevuje systematickým posunem formantových frekvencí), intonace, volba slov, typická slovní spojení (k vyjádření těchto charakteristik se používají různé n-gramové modely). Automatická kvantifikace těchto rysů, a tím pádem jejich využití pro rozpoznávání, je však obtížnější ve srovnání s využitím vnitřních charakteristik souvisejících s anatomií hlasového ústrojí člověka. Na druhou stranu vnitřní charakteristiky jsou závislé na zdravotním nebo emocionálním stavu řečníka a navíc se mění v průběhu času. V [2] je zmiňována studie, která uvádí pokles rozpoznávacího skóre z 96 % na 52 % pokud interval mezi záznamem trénovacích a testovacích promluv byl 3 dny, resp. 3 měsíce.

2.2 Parametrizace příznaky MFCC

Velké množství experimentů prokázalo, že lidské ucho nevnímá frekvence obsažené v řeči v lineární stupnici, ale v tzv. Melovské stupnici. Tato stupnice je přibližně lineární v oblasti do 1 kHz a logaritmická v oblasti nad 1 kHz. Převod mezi frekvencí a mel-frekvencí je dán vztahem:

$$Mel(f) = 2595.\log_{10}\left(1 + \frac{f}{700}\right). \tag{2.1}$$

Výpočet příznakových vektorů (používá se také pojem parametrizace) je detailně popsán v [1] a zhruba probíhá následujícím způsobem. Číslicově zaznamenaný signál je nejprve segmentován na framy. Využívá se přitom faktu, že frekvenční parametry řečového signálu zůstávají téměř konstantní v časových úsecích několika desítek milisekund, což je způsobeno omezenou rychlostí přestavby hlasového traktu člověka. Řečový signál je tedy rozčleněn do segmentů (tzv. framů), jejichž délka bývá obvykle 20 ms. Framy se přitom částečně

překrývají. V dalším kroku se provádí preemfáze, která slouží ke zvýraznění vyšších kmitočtů, které mají obvykle nižší úroveň. Dále se aplikuje Hammingovo okénko, které slouží k potlačení váhy vzorků na okrajích framu za účelem eliminace vlivu náhlého oříznutí průběhu na okrajích. Vlastní výpočet příznaků probíhá tak, že se pomocí FFT vypočte amplitudové spektrum signálu framu. Použitím banky trojúhelníkových filtrů uspořádaných v Melovském měřítku se vypočte energie signálu v příslušných pásmech filtru. Získává se vektor energií, který je dále transformován na vektor kepstrálních příznaků užitím diskrétní kosinové transformace logaritmu vektoru energie. Posledním krokem je provedení tzv. liftrace (jde o filtraci, a protože je prováděna na kepstru, jehož název vznikl přesmyčkou slova spektrum, je název vytvořen přesmyčkou). Tímto způsobem se vypočítají statické kepstrální příznaky.

Příznakový vektor volitelně obsahuje také příznaky dynamické, ty vyjadřují změnu statických příznaků. Dynamické příznaky prvního řádu se označují jako delta příznaky a druhého řádu jako delta-delta příznaky. Výpočet dynamických příznaků se provádí pomocí numerické derivace průběhu statických příznaků v čase. Po parametrizaci je ještě možné provést normalizaci metodou CMS (cepstral mean substraction). Jde o odečtení střední hodnoty kepstra za účelem eliminace vlivu různých zdrojů (mikrofonů, přenosových kanálů, vliv hlasitosti atd.). Normalizace se provádí pouze pro statické příznaky, protože dynamické příznaky by vždy měly mít střední hodnotu nulovou.

Příznaky MFCC jsou tedy odvozeny ze spektra a jako takové zachycují zejména informaci o resonančních frekvencích. To znamená, že se vztahují k fyzickým charakteristikám řečníka. Příznaky MFCC jsou v současnosti nejčastěji používané příznaky v systémech automatického rozpoznávání řečníka.

3. Reprezentace mluvčích

V minulé kapitole bylo popsáno jak řeč vzniká, jaké jsou příčiny odlišností lidských hlasů a výpočet příznaků MFCC, které vyjadřují akustickou část informace obsažené v řeči. Nyní tedy máme k dispozici příznaky a abychom mohli provést rozpoznávání, je nutné vytvořit pro mluvčí modely. Existuje celá řada různých metod, které lze rozdělit podle dvou základních kritérií. Podle povahy modelů je možné rozlišovat metody využívající vzorové nebo statistické reprezentace a v závislosti na úloze metody určené pro textově nezávislé nebo závislé rozpoznávání.

V případě textově závislého rozpoznávání nebo rozpoznávání s pevným slovníkem je nejčastěji používáno rozpoznávání na základě časových funkcí nebo skrytých Markovských modelů. Rozpoznávání na základě časových funkcí vychází ze znalosti vzorové posloupnosti příznakových vektorů dané promluvy pro každého referenčního řečníka. Při rozpoznávání je posloupnost příznakových vektorů testovací promluvy porovnávána s těmito vzorovými posloupnostmi. Protože je velmi nepravděpodobné, že dvě posloupnosti stejných promluv budou shodně dlouhé a budou mít stejnou rytmizaci a to i pro téhož mluvčího, je nutné provést časovou normalizaci. Nejjednodušším způsobem je lineární normalizace, ale její nevýhodou je, že nedokáže postihnout časové kolísání uvnitř posloupností. Proto je vhodnější použít nelineární časovou transformaci. Algoritmus dynamického borcení času (dynamic time warping – DTW) je založený na dynamickém programování a jeho princip spočívá v nalezení nejmenší ze vzdáleností dvou posloupností podél všech přípustných transformačních cest.

Textově závislé systémy využívající pravděpodobnostních modelů jsou nejčastěji založeny na skrytých Markovských modelech (hidden Markov models – HMM). Základní myšlenka vychází z předpokladu, že řeč se skládá z kratších či delších stacionárních úseků, v nichž se parametry mění jen málo. Namísto reprezentace referenčních promluv posloupností framů se framy nahradí menším počtem stavů. Každému stavu je přiřazena informace o vlastnostech framů, které reprezentuje. Nejčastěji jsou tyto vlastnosti určeny statisticky pomocí normálního rozložení, tzn. že každý stav je pak reprezentován střední hodnotou a rozptylem příslušných framů. Současně je stavu přiřazena pravděpodobnost setrvání v tomto stavu, kterou je možné vyjádřit také statisticky na základě množství framů posloupnosti, které daný stav reprezentuje. Klasifikace v tomto případě spočívá ve vyhodnocení míry podobnosti testovací posloupnosti příznaků a referenčních modelů.

Pro tuto práci jsou významné metody využívané v textově nezávislém rozpoznávání a jim bude proto věnována pozornost v této kapitole. Případný zájemce o bližší informace o DTW a HMM je nalezne například v [1, 2, 3].

3.1 Vektorová kvantizace

Vektorová kvantizace (VQ) slouží jako základ mnoha kompresních algoritmů obrazu a zvuku. Z pohledu rozpoznávání řečníka a uvedených kritérií jde o metodu používanou v textově nezávislých systémech, založenou na vzorových reprezentacích.

Motivace pro použití vektorové kvantizace vychází z předpokladu, že akustický prostor odpovídající hlasovým možnostem konkrétního řečníka je tvořen množinou nepřekrývajících se akustických tříd. V [2] je ilustrativně předkládána zjednodušená představa, že tyto třídy odpovídají například kvaziperiodickým, šumovým nebo explozivním zvukům, v případě jemnějšího členění by však mohly odpovídat i jednotlivým fonémům nebo dokonce alofónům. Jednotlivé akustické třídy jsou přitom charakterizovány konfigurací hlasového traktu daného řečníka a nesou tak informaci o identitě řečníka. Ve skutečnosti je ale velmi obtížné předjímat jakékoliv informace o tom, zda se vektory opravdu uspořádají vzhledem k nějakým obecně definovatelným akustickým třídám. Velký vliv na uspořádání bude mít bezpochyby struktura příznakového vektoru (např. v případě příznaků MFCC, zda obsahuje dynamické příznaky) a složení trénovací databáze.

Vektorová kvantizace představuje proces, kdy jsou spojité vektorové hodnoty transformovány na konečný počet diskrétních vektorových hodnot. Princip spočívá v rozdělení *n*-dimenzionálního prostoru Q do L disjunktních podprostorů, kde každý podprostor je reprezentován jediným vektorem – tzv. centroidem – $c_l, l = 1,...,L$. Libovolný vektor $x \in Q$ se pak nahradí centroidem, pro který platí

$$\hat{c}_l = \arg\min_{c_l} d(x, c_l), \tag{3.1}$$

kde $d(x, c_l)$ je vzdálenost vektoru x a centroidu c_l. Zpravidla se používá Eukleidova nebo Mahalanobisova vzdálenost. Množina všech centroidů $\{c_1,...,c_L\}$ se označuje jako kódová kniha. Kódová kniha je navrhnuta tak, aby minimalizovala vzdálenost $d(x, c_l)$ pro všechna x z trénovací množiny. Pro výpočet centroidů kódové knihy se používá algoritmus K-mean, nebo LBG. Druhý jmenovaný byl použit k návrhu kódových knih v této práci.

3.1.1 LBG algoritmus

Algoritmus je pojmenován podle autorů, kteří ho navrhli (Linde, Buzo, Gray). Tato metoda je zobecněním Lloydova algoritmu z roku 1957. Algoritmus se skládá z konečného počtu kroků, kdy je v každém kroku vytvořena nová kódová kniha s dvojnásobným počtem kódových slov a s menším, nebo maximálně stejným celkovým zkreslením oproti předchozí kódové knize.

Nechť $\{x_1,...,x_T\}$ je množina příznakových vektorů získaných z nahrávek daného řečníka a nechť $\{c_1,...,c_L\}$ je množina centroidů kódové knihy tohoto řečníka.

Činnost lze vyjádřit pomocí následujících kroků:

1) *inicializace:* je vytvořena kódová kniha pouze s jedním centroidem, který se určí jako průměr ze sady trénovacích vektorů. Nastaví se $L = 1$ a

$$c_1^* = \frac{1}{T}\sum_{k=1}^{T} x_k \ . \tag{3.2}$$

Na základě zvolené vzdálenosti se vypočte celkové zkreslení D^* přes všechna trénovací data.

2) *rozdělení:* zdvojnásobení počtu centroidů kódové knihy. Nové shluky se vytvoří ze stávajících užitím následujících pravidel

$$c_{L+l}^{(0)} = c_l^*.(1+\varepsilon) \tag{3.3}$$

$$c_l^{(0)} = c_l^*.(1-\varepsilon) \tag{3.4}$$

pro $l = 1,...,L$. ε je iterační parametr, zpravidla nabývá hodnoty v intervalu $(0{,}01 \div 0{,}001)$. $c_l^{(0)}$ představuje *l*-tý centroid počáteční kódové knihy pro následující iterační výpočet.

Po výpočtu všech počátečních centroidů se nastaví $L = 2L$.

3) *optimalizace:* nejprve se nastaví čítač iterací $i = 0$ a hodnota celkového zkreslení pro počáteční kódovou knihu $D^{(0)} = D^*$. Následně se provede vnitřní iterační procedura odpovídající MacQueenovu algoritmu za účelem nalezení optimálních hodnot centroidů kódové knihy velikosti *L*.

i) vektory $x_k, k = 1,...,T$ se rozdělí do *L* tříd $Q_1^{(i)},...,Q_L^{(i)}$ tak, že

$$x_k \in Q_l^{(i)}, \text{ jestliže } d\left(x_k, c_l^{(i)}\right) < d\left(x_k, c_m^{(i)}\right), \forall m = 1,...,L, m \neq l \tag{3.5}$$

ii) aktualizují se hodnoty centroidů dle vztahu

$$c_l^{(i+1)} = \frac{\sum_{x \in Q_l^{(i)}} x}{n_l^{(i)}}, l = 1,..., L, \tag{3.6}$$

kde $n_l^{(i)}$ je počet vektorů *x* v podprostoru $Q_l^{(i)}$ v *i*-tém iteračním kroku.

iii) vypočítá se celkové zkreslení $D^{(i+1)}$ v tomto kroku iterace

iv) pokud platí

$$\frac{D^{(i)} - D^{(i+1)}}{D^{(i)}} > \sigma \tag{3.7}$$

a současně počet iterací nepřevýšil maximální povolený počet, nastaví se index iterace $i = i + 1$ a provede se návrat do kroku i). σ je zvolený práh minimální relativní změny celkového zkreslení pro pokračování v iteračním výpočtu centroidů.

v) nastaví se celkové zkreslení výsledné kódové knihy o velikosti *L* a ztotožní se centroidy této knihy s centroidy kódové knihy vygenerované v posledním iteračním kroku.

$$D^* = D^{(i+1)} \tag{3.8}$$

$$c_l^* = c_l^{(i+1)}, l = 1,..., L \tag{3.9}$$

4) pokud *L* neodpovídá žádané velikosti kódové knihy, provede se návrat do kroku 2).

Je nutné podotknout, že tento algoritmus sice vždy konverguje, ale jen k lokálnímu minimu kriteriální funkce (celkového zkreslení).

3.1.2 Klasifikace pomocí VQ

V průběhu klasifikace se míra příslušnosti posloupnosti vektorů příznaků ke třídě reprezentované kódovou knihou C^s určí následujícím způsobem. Nechť $X = \{x_1 x_2 ... x_T\}$ je posloupnost vektorů příznaků získaná parametrizací promluvy neznámého řečníka a $C^s = \{c_1^s,..., c_{L^s}^s\}$ je kódová kniha řečníka *s* o velikosti L^s. Míra vzájemné příslušnosti se pak určí jako vzdálenost dle vztahu

$$D(X, C^s) = \frac{1}{T} \sum_{t=1}^{T} \left(\min_{l=1,...,L^s} d(x_t, c_l^s) \right). \tag{3.10}$$

Uvedený vztah je možné interpretovat tímto způsobem. Posloupnost X se postupně kvantuje pomocí kódové knihy s-tého referenčního řečníka tak, že se každému vektoru x_t přiřadí nejbližší centroid a výsledná vzdálenost se spočítá jako průměrné kvantizační zkreslení pro kódovou knihu tohoto řečníka.

Při identifikaci v uzavřené množině je vybírán referenční řečník s^* s nejmenší vzdáleností kódové knihy a posloupnosti X, tzn. platí

$$s^* = \arg\min_{s=1,\ldots,S} D\left(X, C^s\right). \tag{3.11}$$

Při verifikaci je vzdálenost $D\left(X, C^s\right)$ porovnávána s verifikačním prahem θ. Předkládaná totožnost řečníka s je odmítnuta pokud $D\left(X, C^s\right) \geq \theta$. V opačném případě je totožnost akceptována.

3.2 Směs Gaussovských rozložení

V české literatuře se lze setkat také s označením Gaussovský mixturový model, nebo Gaussovské hustotní směsi [2]. Jak již bylo zmíněno v úvodu, směs Gaussovských rozložení představuje v současné době hlavní přístup k reprezentaci modelů mluvčích v textově nezávislých systémech. Podobně jako u vektorové kvantizace vychází motivace pro použití GMM z představy uspořádání trénovacích dat do akustických tříd charakteristických pro daného řečníka. Zde ovšem tyto třídy nejsou reprezentovány pouze centroidy, ale spojitým rozložením pravděpodobnosti. Přívlastek Gaussovský vyplývá z faktu, že toto rozložení je normální. Směs je tvořena lineární kombinací jedno-modálních vícerozměrných normálních rozložení akustických tříd a hustota pravděpodobnosti směsi je vyjádřena vztahem

$$P\left(x \mid \lambda\right) = \sum_{l=1}^{L} w_l P_l\left(x\right), \tag{3.12}$$

kde x je n-rozměrný příznakový vektor, λ označuje model, L je počet komponent modelu, w_l jsou váhy těchto komponent splňující podmínku

$$\sum_{l}^{L} w_l = 1 \tag{3.13}$$

a $P_l\left(x\right)$ je funkce hustoty pravděpodobnosti l-té komponenty. Je charakterizována vektorem středních hodnot μ_l o rozměru $n \times 1$, kovariační maticí Σ_l rozměru $n \times n$ a platí

$$P_l(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_l}} \exp\left\{-\frac{1}{2}(x-\mu_l)^T \Sigma_l^{-1}(x-\mu_l)\right\}. \quad (3.14)$$

Model řečníka je tedy jednoznačně určen hodnotami vah komponent, jejich středními hodnotami a kovariačními maticemi, používá se proto souhrnný zápis

$$\lambda = \{w_l, \mu_l, \Sigma_l\}, l = 1,...,L. \quad (3.15)$$

Přestože obecně může být kovariační matice plná, zpravidla se používá pouze diagonální. Hlavními důvody jsou mnohem nižší výpočetní náročnost (zejména při výpočtu inverzní matice), a jak je uvedeno např. v [4], zkušenost, že Gaussovské směsi s diagonálními kovariačními maticemi poskytují lepší výsledky rozpoznávání.

Pro vytvoření modelů mluvčích reprezentovaných směsí Gaussovských rozložení se nejčastěji používá EM algoritmus (Expectation-Maximization) založený na metodě maximalizace věrohodnosti (maximum likelihood estimation – MLE). Za určitých okolností však může být výhodné použít k vytváření modelů metodu maximální aposteriorní pravděpodobnosti (maximum a posteriori probability). Způsob vytváření modelů pomocí těchto metod bude vysvětlen v následujících kapitolách a bude také upozorněno na některá jejich úskalí.

3.2.1 EM algoritmus

Jak již bylo uvedeno, EM algoritmus slouží k estimaci parametrů GMM modelu metodou maximální věrohodnosti. Tato metoda je založena na předpokladu, že parametry modelu jsou pevné, ale neznámé. Parametry jsou estimovány s cílem maximalizace pravděpodobnosti, že příznakové vektory z trénovací množiny byly vygenerovány uvažovaným modelem. Problém odhadu parametrů spadá do obecné třídy problémů chybějících dat (missing data problem). Odhad parametrů by se totiž značně zjednodušil, pokud by bylo pro vektory z trénovací množiny známo, která komponenta je vygenerovala. Tato příslušnost ale není k dispozici.

Při odvození algoritmu se vychází z následující úvahy. Nechť je dána posloupnost nezávislých příznakových vektorů $X = \{x_1,...,x_T\}$. Je známo, že hledaný model je směsí, pro tento okamžik obecných, rozložení a je zadán počet komponent modelu *L*. Úkolem je stanovit parametry rozložení, pro které platí $\lambda \in \Lambda$. Není známa příslušnost vektorů $x_k, k = 1,...,T$

ke komponentám modelu $y_l, l = 1,...,L$, tato informace představuje chybějící data. Rozložení $P(x|\lambda)$ lze vyjádřit jako

$$P(x|\lambda) = \sum_{l=1}^{L} P(x|y_l, \lambda).P(y_l|\lambda) \tag{3.16}$$

a odpovídá tedy směsi rozložení $P(x|y_l, \lambda)$ vážených $P(y_l|\lambda)$. Pravděpodobnost vygenerování posloupnosti *X* se označuje jako věrohodnost $l(\lambda)$ a díky předpokladu nezávislosti vektorů ji lze vyjádřit vztahem

$$l(\lambda) = P(X|\lambda) = \prod_{k=1}^{T} P(x_k|\lambda) = \prod_{k=1}^{T} \sum_{l=1}^{L} P(x_k|y_l, \lambda).P(y_l|\lambda). \tag{3.17}$$

Cílem je nalézt parametry $\lambda^* \in \Lambda$ takové, aby věrohodnost $l(\lambda)$ nabývala své maximální hodnoty. Řeší se tedy problém

$$\lambda^* = \arg\max_{\lambda \in \Lambda} l(\lambda|X) = \arg\max_{\lambda \in \Lambda} \prod_{k=1}^{T} \sum_{l=1}^{L} P(x_k|y_l, \lambda).P(y_l|\lambda). \tag{3.18}$$

Z výpočetních důvodů je výhodné maximalizovat logaritmus věrohodnosti $L(\lambda) = \log(l(\lambda))$. Úpravu lze provést, protože logaritmus je monotónně rostoucí funkce. Optimalizační problém přechází v následující tvar

$$\lambda^* = \arg\max_{\lambda \in \Lambda} L(\lambda|X) = \arg\max_{\lambda \in \Lambda} \sum_{k=1}^{T} \log\left(\sum_{l=1}^{L} P(x_k|y_l, \lambda).P(y_l|\lambda) \right). \tag{3.19}$$

Tento problém nemá analytické řešení, a proto vyžaduje použití numerických optimalizačních metod. Jednou z možností je iterační EM algoritmus, převádějící optimalizační problém na sérii jednodušších, majících pro požadované Gaussovo rozložení analytické řešení. Odhad parametrů GMM pomocí EM algoritmu probíhá následovně:

1) *inicializace:* je vytvořen počáteční odhad parametrů modelu $\mu_l, \Sigma_l, w_l, l = 1,...,L$ a nastaví se čítač iterací $i = 1$

2) *E-step:*

$$\alpha(k, y_l) = \frac{P\left(x_k|y_l, \lambda^{(i-1)}\right).P\left(y_l|\lambda^{(i-1)}\right)}{\sum_{l=1}^{L} P\left(x_k|y_l, \lambda^{(i-1)}\right).P\left(y_l|\lambda^{(i-1)}\right)}, k = 1,...,T, l = 1,...,L \tag{3.20}$$

3) *M-step:*

$$w_l = P(y_l) = \frac{1}{T} \sum_{k=1}^{T} \alpha(k, y_l), l = 1,...,L \tag{3.21}$$

$$\mu_l = \frac{\sum_{k=1}^{T} \alpha(k, y_l).x_k}{\sum_{k=1}^{T} \alpha(k, y_l)}, l = 1,...,L \tag{3.22}$$

$$\Sigma_l = \frac{\sum_{k=1}^{T} \alpha(k, y_l).(x_k - \mu_l)^2}{\sum_{k=1}^{T} \alpha(k, y_l)}, l = 1,...,L \tag{3.23}$$[2]

4) pokud platí

$$\frac{L(\lambda^{(i-1)}) - L(\lambda^{(i)})}{L(\lambda^{(i-1)})} > \varepsilon \tag{3.24}$$

a současně počet iterací nepřevýšil maximální povolený počet, nastaví se index iterace $i = i + 1$ a provede se návrat do kroku 2).

Různými způsoby volby počátečních parametrů modelu se zabývá např. [10], prováděný experiment sice nesouvisí s rozpoznáváním řečníka, ale ukazuje na úloze rozpoznávání dvou slov, že volba vhodného způsobu nastavení počátečních parametrů se může pro danou úlohu lišit v závislosti na použitém počtu komponent modelu nebo počtu příznaků. Bohužel neexistuje žádný obecný návod, jak co nejlépe volit počáteční parametry. V [2] je zmiňován experiment, při kterém byla trénovací data rozdělena do 50 předem definovaných akustických tříd a vektory středních hodnot a kovariační matice těchto tříd byly použity jako inicializační odhad pro EM algoritmus. Výsledky identifikace řečníka však nebyly nijak výrazně lepší ve srovnání s výsledky, kdy bylo za počáteční vektory středních hodnot dosazeno 50 náhodně vybraných příznakových vektorů a jako počáteční kovariační matice byly zvoleny identické matice. Kromě náhodné inicializace se často využívají také shlukové algoritmy.

Problém vhodné volby počtu komponent lze řešit natrénováním modelů s jejich různým počtem a výběrem modelu na základě zvoleného výběrového kritéria. To je běžně funkcí věrohodnosti daného modelu a počtu komponent. Snahou je maximální věrohodnost při co nejnižším počtu komponent a rostoucí počet komponent je tak penalizován zvyšující se hodnotou kritéria. Tento způsob je blíže popsán v [12]. Nevýhodou EM algoritmu je dále možnost konvergence ke hranicím prostoru příznaků a tím pádem vznik singulárních kovariačních matic pro některé komponenty (problém se označuje jako přeučení, v anglické

[2] x^2 slouží jako zkratka pro xx'

literatuře pak jako overfitting). K tomuto dochází v případech, je-li zvolen velký počet komponent a přitom je k dispozici malé množství trénovacích dat. Na druhou stranu pokud je zvolen nízký počet komponent, může se stát, že směs rozložení nebude schopna dostatečně věrohodně aproximovat skutečné rozložení. Nicméně existují varianty EM algoritmu, které se tento problém snaží řešit.

Snahou Figueiredo-Jain (FJ) algoritmu prezentovaného v [12] je odstranit nutnost stanovení pevného počtu komponent pro vytvářený model. Vhodný počet komponent není stanoven na základě výběru z několika různě velkých směsí, ale je stanoven přímo. V průběhu estimace modelu jsou prořezávány komponenty, které nejsou podporovány daty, nebo se jejich kovariační matice stanou singulárními. Počet komponent je tak snižován ze stanoveného počátečního maxima.

Hladový EM algoritmus (greedy EM algorithm, GEM) [13] také stanoví počet komponent samovolně. GEM začíná s jedinou komponentou a postupně po jedné přidává. Tím, že se začíná pouze s jedinou komponentou je značně zjednodušen problém odhadu počátečních parametrů modelu. Stačí spočítat střední hodnotu a kovariační matici přes všechna trénovací data. V [11] bylo provedeno porovnání standardního EM, FJ a GEM algoritmu na třech různých úlohách. Standardní EM algoritmus dokázal při vhodné volbě počtu komponent směsi překonat ostatní metody. Pomocí FJ algoritmu bylo dosaženo lepších výsledků ve srovnání s GEM, ale nevýhodou jsou vyšší nároky na minimální množství trénovacích dat v porovnání s ostatními algoritmy. Výhodou GEM algoritmu byla největší robustnost odhadu vzhledem k množství trénovacích dat.

3.2.2 Metoda MAP

Zatímco metoda maximální věrohodnosti předpokládá, že estimované parametry jsou pevné a neznámé, metoda maximální aposteriorní pravděpodobnosti (maximum a posteriori – MAP) vychází z předpokladu, že parametry jsou náhodné veličiny se známým apriorním rozložením. Tato metoda se často používá k estimaci parametrů modelů v případě nedostatku trénovacích dat, kdy je tento nedostatek kompenzován znalostí apriorního rozložení.

Nechť $X = \{x_1,...,x_T\}$ je posloupnost příznakových vektorů a $P(\lambda)$ je apriorní rozložení parametrů λ. Cílem metody MAP je nalézt hodnoty parametrů λ_{MAP} takové, aby platilo

$$\lambda_{MAP} = \arg\max_{\lambda} P(\lambda|X), \tag{3.25}$$

kde $P(\lambda|X)$ je aposteriorní rozložení estimovaných parametrů pro posloupnost *X*. S využitím Bayesova teorému je možné přepsat vztah (3.25) jako

$$\lambda_{MAP} = \arg\max_{\lambda} \frac{P(X|\lambda).P(\lambda)}{P(X)}. \tag{3.26}$$

Maximalizace aposteriorního rozložení pravděpodobnosti $P(\lambda|X)$ je dosaženo změnou parametru λ tak, aby byl maximalizován výraz $P(X|\lambda).P(\lambda)$. Detailní odvození vztahů pro výpočet nových hodnot parametrů lze nalézt v [6].

V oblasti rozpoznávání mluvčích se tato metoda nejčastěji používá za účelem odvození verifikačních modelů mluvčích adaptací univerzálního modelu okolí (universal background model – UBM). Tento pojem bude vysvětlen v následující kapitole. Pro tuto chvíli si vystačme s konstatováním, že UBM model je natrénován pomocí metody maximální věrohodnosti na velkém množství dat. Uváděný výpočet parametrů neodpovídá přesně obecné metodě MAP, ale jak je uvedeno v [5], bylo experimentálně zjištěno, že výsledky verifikace při použití těchto vzorců jsou lepší.

Nechť $\lambda = \{w_l, \mu_l, \Sigma_l\}$ jsou parametry UBM modelu reprezentující apriorní odhad. Hledané parametry modelu řečníka jsou $\hat{\lambda} = \{\hat{w}_l, \hat{\mu}_l, \hat{\Sigma}_l\}$. Adaptace parametrů modelu pro posloupnost *X* probíhá následovně:

1) výpočet okupační pravděpodobnosti *l*-té komponenty směsi pro posloupnost *X*

$$\alpha(k, y_l) = \frac{P_l(x_k).w_l}{\sum_{l=1}^{L} P_l(x_k).w_l}, k = 1,...,T, l = 1,...,L. \tag{3.27}$$

2) stanovení adaptačních váhových koeficientů pro váhy, střední hodnoty a kovariační matice komponent

$$\xi_l^{\rho} = \frac{\sum_{k=1}^{T} \alpha(k, y_l)}{\sum_{k=1}^{T} \alpha(k, y_l) + r^{\rho}}, \rho \in \{w, \mu, \Sigma\}, \tag{3.28}$$

kde váhový faktor r^{ρ} je konstantní pro všechny komponenty.

3) nové parametry modelu jsou určeny rovnicemi

$$\hat{w}_l = \left[\xi_l^w \frac{1}{T} \sum_{k=1}^{T} \alpha(k, y_l) + (1 - \xi_l^w) w_l \right] \gamma, l = 1,..., L \tag{3.29}$$

$$\hat{\mu}_l = \xi_l^\mu \frac{\sum_{k=1}^{T} \alpha(k, y_l).x_k}{\sum_{k=1}^{T} \alpha(k, y_l)} + (1 - \xi_l^\mu) \mu_l, l = 1,..., L \tag{3.30}$$

$$\hat{\Sigma}_l = \xi_l^\Sigma \frac{\sum_{k=1}^{T} \alpha(k, y_l).x_k^{\,2}}{\sum_{k=1}^{T} \alpha(k, y_l)} - \hat{\mu}_l^2 + (1 - \xi_l^\Sigma)(\Sigma_l + \mu_l^2), l = 1,..., L \tag{3.31}$$

Kde koeficient γ zajišťuje, že součet vah komponent směsi bude roven jedné.

Je tedy zřejmé, že parametry nového modelu jsou určeny jako vážený součet hodnot apriorních parametrů a hodnot parametrů odhadnutých metodou maximální věrohodnosti z dat posloupnosti *X*. Čím větší je hodnota váhového faktoru r^ρ, tím větší vliv na výsledek estimace mají apriorní parametry. Naopak s rostoucím množstvím dat určených pro estimaci roste vliv parametrů odhadnutých metodou maximální věrohodnosti.

3.2.3 Klasifikace pomocí GMM

Míra příslušnosti posloupnosti $X = \{x_1 ... x_T\}$ odvozené z promluvy neznámého řečníka a směsi Gaussovských rozložení řečníka *s* se určí jako aposteriorní pravděpodobnost, že tuto promluvu vyslovil řečník *s*, jde tedy o podmíněnou pravděpodobnost $P(\lambda^s | X)$. Pomocí směsi Gaussovských rozložení ale není možné určit tuto pravděpodobnost přímo, a proto se provede následující odvození na základě Bayesova vztahu.

$$P(\lambda^s | X) = \frac{P(X | \lambda^s) P(\lambda^s)}{P(X)}, \tag{3.32}$$

kde $P(X | \lambda^s)$ je pravděpodobnost, že řečník *s* vysloví promluvu, ze které byla posloupnost *X* získána způsobem odpovídajícím právě posloupnosti *X*. Tuto pravděpodobnost lze určit pomocí směsi Gaussovských rozložení. $P(\lambda^s)$ je apriorní pravděpodobnost, že promluvil řečník *s*, a pro $P(X)$ platí

$$P(X) = \sum_{s=1}^{S} P\left(X \middle| \lambda^s\right) P\left(\lambda^s\right). \tag{3.33}$$

Cílem identifikace v uzavřené množině je nalézt řečníka s^*, který s největší pravděpodobností vyslovil promluvu *X*. Tedy

$$s^* = \arg\max_{s=1,\ldots,S} P\left(\lambda^s \middle| X\right) = \arg\max_{s=1,\ldots,S} \frac{P\left(X \middle| \lambda^s\right) P\left(\lambda^s\right)}{P(X)}. \tag{3.34}$$

Obecně lze pravděpodobnost $P\left(\lambda^s\right)$ považovat za shodnou pro všechny řečníky a tím pádem může být ve vztahu zanedbána. Ze vztahu (3.33) je zřejmé, že ani pravděpodobnost $P(X)$ není závislá na konkrétním řečníkovi a také ji lze proto zanedbat. V praxi se z výpočetních důvodů často počítá s logaritmem pravděpodobnosti. Dostáváme tak kritérium maxima logaritmů věrohodnosti (maximum log-likelihood)

$$s^* = \arg\max_{s=1,\ldots,S} \left[\log P\left(X \middle| \lambda^s\right)\right]. \tag{3.35}$$

Při verifikaci je aposteriorní pravděpodobnost řečníka *s* porovnávána s verifikačním prahem θ. Pokud platí $P\left(\lambda^s \middle| X\right) \leq \theta$, je totožnost řečníka *s* odmítnuta. Totožnost je přijata, jestliže $P\left(\lambda^s \middle| X\right) > \theta$. Stejně jako v případě identifikace v uzavřené množině pravděpodobnosti $P\left(\lambda^s\right)$ a $P(X)$ nezávisí na konkrétním řečníkovi, zde je ovšem nutné zohlednit závislost pravděpodobnosti $P(X)$ na posloupnosti *X*. Při identifikaci tato závislost nepředstavovala problém, protože všechny pravděpodobnosti porovnávané při výběru maxima jsou závislé na $P(X)$ stejným způsobem a ve výsledku tedy tento člen nemá význam a může být skutečně zanedbán. Při verifikaci dochází k porovnání s prahem θ a tento práh by se při zanedbání členu $P(X)$ musel měnit v závislosti na *X*. To je v praxi realizovatelné pouze obtížně. Vhodnější řešení představuje normalizace aposteriorní pravděpodobnosti tak, aby verifikační práh θ mohl být stanoven nezávisle na *X*.

Verifikace používaná v této práci je založena na principu testování hypotéz a použití univerzálních modelů okolí (universal background model – UBM) [5]. Úkol odmítnutí, nebo přijmutí předkládané totožnosti je formulován jako test hypotéz vyjadřujících následující situace

H_0: *X* reprezentuje promluvu, kterou vyslovil prohlašovaný řečník s

H_1: *X* reprezentuje promluvu, kterou nevyslovil prohlašovaný řečník s.

Pokud budeme reprezentovat hypotézu H_0 pravděpodobností $P\left(\lambda^s \middle| X\right)$ a hypotézu H_1 pravděpodobností $P\left(\lambda^{\bar{s}} \middle| X\right)$, provede se rozhodnutí na základě poměru věrohodností následujícím způsobem. Hypotéza H_0 je přijata, jestliže platí

$$\frac{P\left(\lambda^s \middle| X\right)}{P\left(\lambda^{\bar{s}} \middle| X\right)} > \theta\,, \tag{3.36}$$

v opačném případě je hypotéza H_0 zamítnuta. S využitím Bayesova vztahu lze vztah (3.36) po provedení úprav přepsat jako

$$\frac{P\left(X \middle| \lambda^s\right)}{P\left(X \middle| \lambda^{\bar{s}}\right)} > \frac{P\left(\lambda^{\bar{s}}\right)}{P\left(\lambda^s\right)}\theta\,. \tag{3.37}$$

Označíme-li

$$\theta' = \frac{P\left(\lambda^{\bar{s}}\right)}{P\left(\lambda^s\right)}\theta\,, \tag{3.38}$$

lze namísto (3.37) psát

$$\frac{P\left(X \middle| \lambda^s\right)}{P\left(X \middle| \lambda^{\bar{s}}\right)} > \theta'\,. \tag{3.39}$$

Opět se zpravidla počítá s logaritmem věrohodností

$$\log P\left(X \middle| \lambda^s\right) - \log P\left(X \middle| \lambda^{\bar{s}}\right) > \Theta\,, \tag{3.40}$$

kde $\Theta = \log\theta'$. Vztah (3.40) vyjadřuje kritérium poměru logaritmů věrohodnosti (log-likelihood ratio). Člen $\log P\left(X \middle| \lambda^{\bar{s}}\right)$ bývá označován jako normalizační. Existují dva hlavní přístupy k určení tohoto členu. První vychází z použití modelů ostatních referenčních řečníků za účelem pokrytí akustického prostoru možných podvodníků a následného výběru z věrohodností určených pro tyto modely. Neprovádí se přitom vyhodnocení věrohodnosti pro modely všech referenčních řečníků, ale pouze pro skupinu řečníků označovanou jako kohorta. Situaci však komplikuje výběr řečníků kohorty. Experimenty prokázali, že k dosažení nejlepších výsledků je nutné sestavit pro každého mluvčího individuální kohortu [5]. Druhý přístup spočívá ve vytvoření jediného modelu z velkého množství dat od mnoha řečníků. Tento model se označuje jako univerzální model okolí (UBM) a je společný pro všechny řečníky. Výběr trénovacích dat musí odpovídat situacím, do kterých se systém může při svém provozu dostat. Pokud je předem známo, že systém bude zpracovávat telefonní nahrávky, všichni referenční řečnící jsou muži a možnými podvodníky jsou také pouze muži, budou trénovací data

tvořena pouze telefonními nahrávkami mužských hlasů. Pokud ale není předem známa žádná informace o pohlaví možných podvodníků, mikrofonech, kterými jsou nahrávky pořizovány a přenosových kanálech, kterými nahrávky procházejí, je nutné, aby trénovací množinu tvořila data zaznamenaná různými mikrofony, z různých přenosových kanálů a od řečníků obou pohlaví. Přitom je důležité, aby nahrávky odpovídající různým akustickým podmínkám, stejně jako nahrávky obou pohlaví, byly v trénovacích datech zastoupeny pokud možno rovnoměrně, aby nebyl model vychýlený směrem k převládající skupině. V praxi lze tento problém řešit trénováním modelů pro různé skupiny (např. muže a ženy) odděleně a následným sloučením modelů. Případně není nutné modely slučovat v jeden, věrohodnost se vyhodnotí samostatně pro dílčí modely a výsledný model se vybere dle vztahu [15]

$$P\left(X\middle|\lambda^{UBM}\right)=\max\left\{P\left(X\middle|\lambda^{1}\right),P\left(X\middle|\lambda^{2}\right)\right\}. \tag{3.41}$$

4. Popis navrhovaného systému

4.1 Formulace požadavků

Zadáním práce je provést rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů a je tedy zřejmé, že neznámým řečníkem může být v podstatě kdokoliv. Za správné rozpoznání jsou považovány následující situace:

- neznámý řečník je jeden z referenčních řečníků a systém určí správně jeho totožnost
- neznámý řečník není nikdo z referenčních řečníků a systém rozhodne, že řečník je neznámý.

To odpovídá úloze identifikace v otevřené množině. Snahou je, aby byla referenční množina tvořena co možná největším počtem řečníků (hlasatelů, moderátorů, politiků, mluvčích, atd.) objevujících se nejčastěji na českých televizních a rozhlasových stanicích se zaměřením na zpravodajské a diskusní pořady.

Systém je navrhován s ohledem na začlenění do systému automatického přepisu televizních a rozhlasových pořadů vyvíjeného již několik let Laboratoří počítačového zpracování řeči. Přínos rozpoznávání mluvčích pro tento systém je dvojí. Uživatel systému získává informaci o totožnosti řečníka nahrávky a navíc lze tuto informaci využít ke zlepšení úspěšnosti rozpoznávání řeči díky použití modelu adaptovaného pro daného řečníka (speaker adapted – SA). Nejnovější výsledky systému přepisu pořadů dosažené při použití adaptovaných modelů pro rozpoznávání řeči lze nalézt např. v [22].

Nahrávky předkládané modulu rozpoznávání řečníka jsou získány pomocí automatické segmentace pořadů. Při návrhu modulu je tak nutné počítat s tím, že některé nahrávky mohou být tvořeny hlukem nebo tichem a systém by je měl být schopen identifikovat.

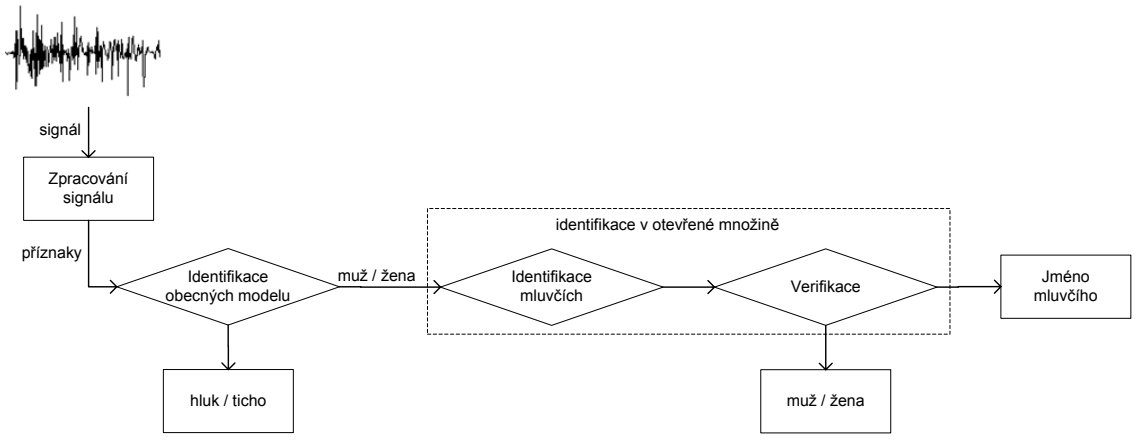
Posledním kvalitativním požadavkem na informace poskytované tímto modulem je identifikace pohlaví mluvčího. Pokud totiž modul o nahrávce rozhodne, že se jedná o řečový segment a že hlas nepatří nikomu z referenčních řečníků, je z pohledu uživatele i rozpoznávače řeči využívajícího adaptované modely přínosné, aby modul rozpoznávání řečníka poskytl informaci o pohlaví mluvčího.

4.2 Návrh řešení

Z požadavků jednoznačně vyplývá, že je logické provést jako první krok detekci, zda předložená nahrávka reprezentuje řečový nebo neřečový segment. Další zpracování je pak prováděno pouze s řečovými segmenty. Na základě uvedených požadavků lze intuitivně očekávat, že v dalším kroku proběhne identifikace v otevřené množině a v případě rozhodnutí, že neznámý řečník není nikdo z referenčních řečníků, je provedeno určení pohlaví. V navrhovaném systému je však určení pohlaví provedeno v prvním kroku společně s detekcí řeči. Myšlenka vychází ze snahy o maximální využití prostředků vytvořených pro rozpoznávání mluvčích. Činnost systému lze rozdělit do dvou hlavních kroků:

1) nejprve se vykoná identifikace v uzavřené množině s obecnými modely ticha, hluku a UBM modely pro muže a ženy. Pokud je výsledkem této fáze rozhodnutí, že předložený segment je řečový (mluvčím je žena nebo muž), pokračuje se krokem 2).
2) provede se identifikace v otevřené množině s modely referenčních řečníků. Při verifikaci prováděné v této fázi v rámci ověření totožnosti předložené identifikací v uzavřené množině se s ohledem na vztah (3.41) použije mužský nebo ženský UBM model v závislosti na výsledku kroku 1), přitom není nutné provádět opětovné vyhodnocení věrohodnosti.

UBM modely tak v tomto systému nacházejí vícenásobné uplatnění. Nejprve jsou použity k detekci řeči a určení pohlaví, v dalším kroku slouží k vytvoření normalizačního členu pro verifikaci. Schéma navrhovaného řešení je znázorněno na obr. 4-1.



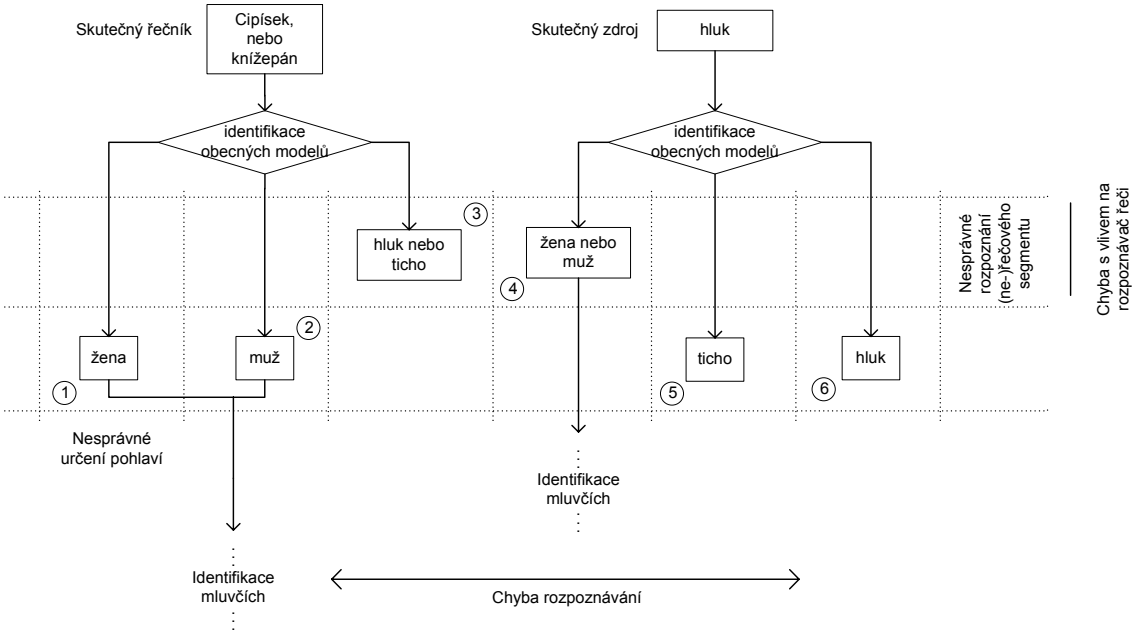
Obr. 4-1 – *schéma navrhovaného systému*

4.3 Ohodnocení systému

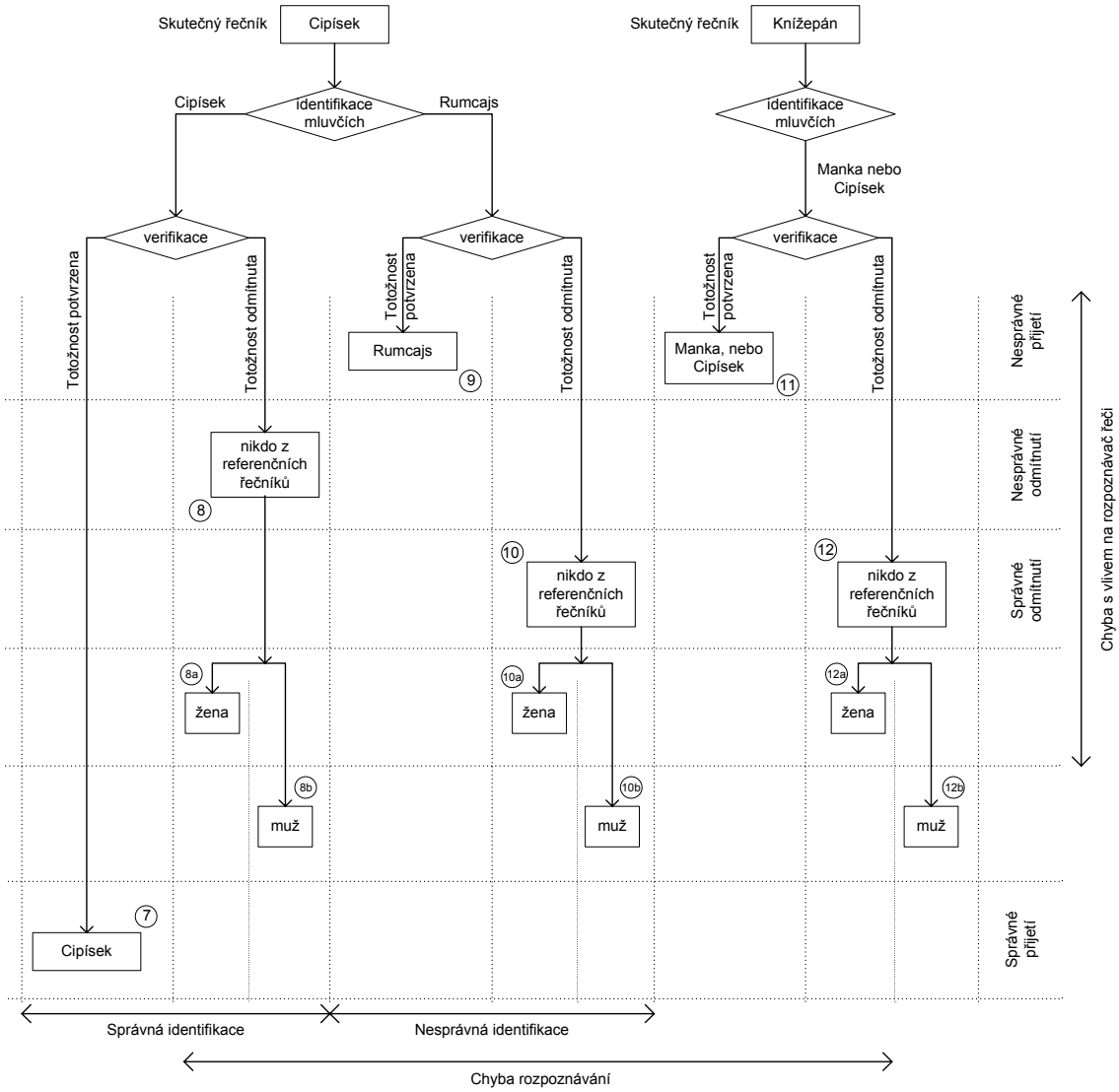
Navrhovaný systém se může dopustit různých chyb, které lze vnímat z odlišných pohledů. Systém je tvořen jako soubor dvou identifikací v uzavřené množině a verifikace. Z pohledu uživatele systému není významné, zda chyba nastala při identifikaci nebo verifikaci. Na druhou stranu nás může například zajímat s jakou úspěšností je identifikace používající obecné modely schopna určit řečové segmenty a z této znalosti pak učinit závěr o oprávněnosti použití tohoto jednoduchého způsobu identifikace řečových segmentů. Pro stanovení verifikačního prahu je důležité znát pravděpodobnost nesprávného odmítnutí nebo nesprávného přijmutí. Z pohledu vývoje systému tak potřebujeme mít k dispozici informace o chybách, ke kterým dochází v jednotlivých částech systému.

V [2] jsou pro názornost referenční řečníci i možní podvodníci pojmenováni po vzoru postav z pohádky o Rumcajsovi. Využijeme této ilustrativní představy pro následující výklad. Předpokládejme, že referenčními řečníky jsou Rumcajs, Cipísek a Manka. Podvodníky lze rozlišovat na cizí a vlastní. Cizími podvodníky jsou všichni řečníci, kteří nepatří do referenční množiny, například kněžna a knížepán. Vlastními podvodníky se rozumí referenční řečníci vydávající se za někoho jiného. V navrhovaném systému tato situace nastává pokud dojde ke špatnému vyhodnocení ve fázi identifikace mluvčích v uzavřené množině.

Stavy, které mohou v průběhu činnosti systému nastat, jsou pro přehlednost zmapovány na obr. 4-2 a 4-3, čísla přiřazená stavům slouží k vytváření odkazů v následujícím výkladu.



Obr. 4-2 – *znázornění stavů, které mohou nastat při identifikace obecných modelů.*



Obr. 4-3 – *stavy, které mohou nastat při identifikaci mluvčích a příslušné ohodnocení.*

Činnost navrhovaného systému je ohodnocena následujícími chybami:

- míra neúspěšnosti identifikace řečových segmentů

$$R_{SNSE} = \frac{n_{s-ns} + n_{ns-s}}{n_{total}}, \tag{4.1}$$

kde n_{s-ns} je počet řečových segmentů (nahrávek), které byly označeny jako neřečové (situace ③), n_{ns-s} je počet neřečových segmentů, které byly označeny jako řečové (④) a n_{total} je celkový počet pokusů.

- míra neúspěšnosti identifikace pohlaví mluvčího

$$R_{GDE} = \frac{n_{m-f} + n_{f-m}}{n_{s-s}}, \tag{4.2}$$

kde n_{m-f} je počet nahrávek, kdy řečníkem byl muž a výsledkem identifikace byla žena (①), n_{f-m} vyjadřuje počet opačných chyb identifikace pohlaví, n_{s-s} je počet řečových segmentů, které byly správně rozpoznány jako řečové. Obě míry ohodnocení, R_{SNSE} i R_{GDE}, se tak vztahují k první identifikaci využívající obecné modely a je tak možné namítnout, že použití dvou ohodnocení je zbytečné a tato část navrhovaného systému mohla být ohodnocena jako prostá identifikace v uzavřené množině tak, jak je uvedeno např. v [2]. Navíc je patrné, že míra R_{GDE} je závislá na R_{SNSE}. Použitý způsob ohodnocení byl nicméně zvolen se zřetelem na význam modelů, který je dvojznačný. UBM modely reprezentují jednak řeč a jednak pohlaví. Navíc v případě implementace odlišného přístupu detekce řečových segmentů zůstává způsob ohodnocení neměnný a výsledky jsou porovnatelné.

- míra neúspěšnosti identifikace referenčních řečníků

$$R_{SIE} = \frac{n_{wspk}}{n_{reff}}, \tag{4.3}$$

kde n_{reff} je počet nahrávek, kdy byl řečníkem někdo z referenčních řečníků a n_{wspk} je počet případů, kdy byl řečník identifikován chybně (⑨, ⑩). Toto ohodnocení tedy přesně odpovídá běžnému ohodnocení identifikace v uzavřené množině.

- poměrný počet chyb nesprávného přijetí

$$R_{FA} = \frac{n_{FA}}{n_{wspk} + n_{nreff}}, \tag{4.4}$$

kde n_{nreff} je počet pokusů, při kterých neznámý řečník nebyl nikdo z referenčních řečníků a n_{FA} je počet případů, kdy byla přijata totožnost předložená identifikací v uzavřené množině, přestože byl výsledek identifikace chybný (⑨) nebo se nejednalo o nikoho z referenčních řečníků (⑪). Tato míra závisí na hodnotě verifikačního prahu θ.

- poměrný počet chyb nesprávného odmítnutí

$$R_{FR} = \frac{n_{FR}}{n_{rspk}}, \tag{4.5}$$

kde n_{rspk} je počet pokusů, kdy neznámý řečník je jeden z referenčních řečníků a byl správně identifikován, n_{FR} je počet případů, ve kterých verifikace tuto správnou totožnost odmítla (⑧). Stejně jako R_{FA} i R_{FR} závisí na volbě verifikačního prahu θ.

- míra neúspěšnosti rozpoznávání

$$R_E = 1 - \frac{n_{rpsk-CA} + n_{nreff-CR} + n_{rns}}{n_{total}}, \tag{4.6}$$

kde $n_{rspk-CA}$ je počet pokusů, kdy neznámý řečník byl jeden z referenčních řečníků, byl správně identifikován a totožnost byla správně potvrzena verifikací (⑦), $n_{nreff-CR}$ je počet případů, ve kterých neznámý řečník nebyl nikdo z referenčních řečníků, verifikace správně odmítla předkládanou totožnost a současně bylo správně určeno pohlaví řečníka (⑫b). n_{rns} představuje počet pokusů, kdy neřečová nahrávka byla správně označena jako neřečová a navíc byl správně rozpoznán konkrétní neřečový model, tzn. ticho nebo hluk (⑥). Toto ohodnocení je chápáno jako nejdůležitější z pohledu uživatele systému.

- míra neúspěšnosti rozpoznávání s vlivem na rozpoznávač řeči využívající adaptované modely řečníků

$$R_{SAE} = 1 - \frac{n_{rpsk-CA} + n_{rgd-R} + n_{ns-ns}}{n_{total}}, \tag{4.7}$$

kde n_{ns-ns} je počet neřečových segmentů, které byly správně označeny jako neřečové (⑤,⑥). Na rozdíl od R_E není důležité, zda byl neřečový model rozpoznán správně, protože neřečový segment je z dalšího zpracování vyřazen a je-li hluk rozpoznán jako ticho, případně naopak, nemá na rozpoznávač řeči žádný vliv. n_{rgd-R} je počet pokusů, kdy bylo správně rozpoznáno pohlaví řečníka a verifikace zamítla předkládanou totožnost, bez ohledu na to, zda došlo k správné nebo chybné identifikaci a nezávisle na tom zda neznámý řečník je nebo není někdo z referenčních řečníků (⑧b , ⑩b , ⑫b). Kladné ohodnocení zamítnutí totožnosti řečníka bez ohledu na správnost identifikace, ale za předpokladu správně určeného pohlaví, pro R_{SAE} vyplývá z faktu, že pokud

rozpoznávač řeči použije model adaptovaný pro nesprávného řečníka (⑨, ⑪), úspěšnost se sníží. Zatímco při použití modelu adaptovaného pro správné pohlaví řečníka (gender dependent – GD) se úspěšnost zvýší ve srovnání s použitím modelu zcela nezávislého na řečníkovi (speaker independent – SI). S ohledem na rozpoznávač řeči je tedy vhodné volit verifikační práh θ spíše přísněji, protože nesprávné odmítnutí nepředstavuje na rozdíl od nesprávného přijmutí žádnou komplikaci. Na druhou stranu zvýšený počet nesprávných odmítnutí negativně ovlivní R_E.

Je zřejmé, že uvedená ohodnocení jsou na sobě do značné míry závislá. To snižuje jejich porovnatelnost napříč různým experimentům se systémem. Pro porovnávání výkonu systému se tak v kapitole 6 používá míra R_E a R_{SAE}. Ostatní uvedená ohodnocení slouží jako informativní, zda se systém nedopouští v průběhu jedné z identifikací nebo verifikace výrazně vysokého počtu chyb, což by signalizovalo např. nevhodným způsobem natrénované modely. R_{FA} a R_{FR} poskytují důležitou informaci pro stanovení verifikačního prahu θ.

4.4 Volba verifikačního prahu

Často používaným způsobem ohodnocení systémů verifikace řečníka je míra R_{EER}, která umožňuje ohodnotit systém jediným číslem, a tzv. DET křivka, která slouží ke grafickému ohodnocení. Míra R_{EER}, používá se anglický název Equal Error Rate, je určena následujícím vztahem

$$R_{EER} = R_{FR}(\theta_{EER}) = R_{FA}(\theta_{EER}). \tag{4.8}$$

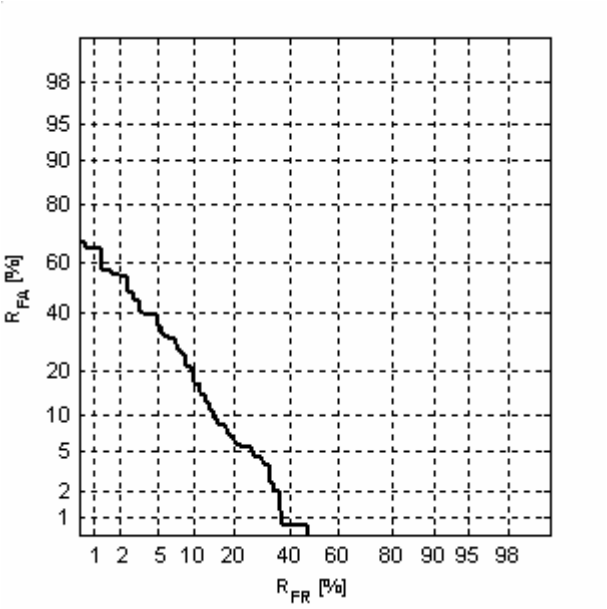
Je tedy nutné určit hodnotu verifikačního prahu θ_{EER}, pro kterou je poměrný počet chyb nesprávného přijetí roven poměrnému počtu chyb nesprávného odmítnutí. Přesné určení této hodnoty je značně obtížné, protože zpravidla se v praxi nejprve určí hodnota verifikačního prahu a až následně se zjišťují chyby, kterých se systém při daném nastavení dopustí. Řešením je určení pouze přibližné hodnoty R'_{EER} dle vztahu

$$R'_{EER} = \frac{R_{FR}(\theta'_{EER}) + R_{FA}(\theta'_{EER})}{2}, \tag{4.9}$$

kde

$$\theta'_{EER} = \arg\min_{\theta} \left| R_{FR}(\theta) - R_{FA}(\theta) \right|. \tag{4.10}$$

DET křivka vyjadřuje graficky závislost $R_{FR}(\theta)$ na $R_{FA}(\theta)$, přičemž parametrem závislosti je hodnota verifikačního prahu θ. Zvláštností této křivky je, že osy nejsou v lineárním měřítku vzhledem k vynášeným mírám $R_{FR}(\theta)$ a $R_{FA}(\theta)$, ale použité měřítko je lineární vzhledem ke kvantilům normálního normovaného rozložení. Díky tomu jsou lépe odlišitelné křivky vynášené pro různé systémy. Příklad DET křivky je znázorněn na obr. 4-4.



Obr. 4-4 – *DET křivka systému verifikace mluvčích.*

V předchozí kapitole bylo uvedeno, že s ohledem na rozpoznávač řeči je vhodnější, pokud je práh stanoven tak, aby spíše odmítl správného řečníka, než přijal nesprávného. Při reálném nasazení je systém skutečně nastaven v souladu s tímto požadavkem. Hodnota verifikačního prahu je určena víceméně empiricky na základě DET křivky. S ohledem na porovnatelnost provedených experimentů, je však nutné volit práh ve všech případech shodným způsobem, proto je v této práci pro experimenty se systémem verifikační práh volen jako θ_{EER} (resp. θ'_{EER}).

5. Popis databáze nahrávek

Navrhovaný systém byl vyvíjen na neveřejné databázi televizních a rozhlasových nahrávek shromažďovaných Laboratoří počítačového zpracování řeči za účelem vytvoření systému automatického přepisu pořadů. Zájem je soustředěn na zpravodajské a diskusní pořady. Databáze je soustavně rozšiřována a v době návrhu obsahovala více než 31 hodin nahrávek zaznamenaných v průběhu více než 3 let. V databázi je evidováno více než 800 řečníků s různým množstvím dat. Nicméně skutečný počet řečníků zastoupených v databázi je ještě vyšší. U nahrávek osob, o kterých lze předpokládat, že se ve sledovaných pořadech nebudou objevovat opakovaně (např. respondenti různých anket) není v databázi uloženo jméno (často není ani známé), ale pouze informace o pohlaví. To umožňuje použít nahrávky těchto osob při trénování UBM modelů, čímž se zvýší robustnost modelu. Obdobně má velké množství různých řečníků kladný přínos pro trénování robustních GD modelů pro rozpoznávání řeči.

Nahrávky jsou zaznamenávány s vzorkovací frekvencí 16 kHz a 16-ti bitovým rozlišením. Televizní a rozhlasový signál je zpracováván pomocí televizního tuneru v PC. Příjem signálu je anténní, nebo kabelový. O mikrofonech použitých k získání nahrávek nejsou informace dostupné. Nahrávky pocházejí z různých prostředí (studio, terén, telefonní vstupy, atd.) a obsahují různou úroveň hluků. Velká variabilita prostředí, mikrofonů, přenosových kanálů a různé způsoby příjmu signálu představují pro rozpoznávání vždy značnou komplikaci. Je nutné vzít v úvahu také dlouhou dobu, po kterou jsou nahrávky pořizovány.

K testovacím účelům byla použita sada dat tvořená 230 minutami nahrávek hlavních zpravodajských relací všech tří hlavních českých televizních stanic. Nahrávky jsou nezávislé na trénovací databázi, ale jejich charakteristika je shodná. I tyto nahrávky byly shromažďovány v průběhu více než 3 let. Testovací sada obsahuje nahrávky 255 řečníků, pouze částečně společných s trénovací databází.

Velké množství publikovaných experimentů je zpravidla vyhodnocováno s použitím veřejných řečových korpusů, jakými jsou například YOHO, NIST, KING, SPIDRE a jiné [17]. Korpus představuje soubor řečových záznamů doplněných o anotaci a dokumentaci. Dokumentace poskytuje informace o počtu řečníků zastoupených v korpusu, množství promluv, intervalu mezi nahrávkami, definuje použité mikrofony, přenosové kanály, akustické podmínky, omezení slovníku, nebo zda se jedná o čtené či spontánní promluvy. Veřejné korpusy umožňují objektivní srovnání výsledků dosažených použitím různých přístupů k identifikaci, verifikaci, detekci změny mluvčího atd. V navrhovaném systému bylo vyzkoušeno několik přístupů, jejichž prospěšnost byla prokázána s použitím veřejných korpusů.

6. Provedené experimenty

S navrhovaným systémem byly provedeny experimenty zaměřené na porovnání výsledků dosažených při použití modelů řečníků založených na VQ a GMM, byly testovány různé úpravy vyhodnocení identifikace a verifikace, vliv množství trénovacích dat, význam detekce řečových framů atd. Většina těchto experimentů je odvozena od základního experimentu, ten představuje referenční nastavení systému a jím dosažené výsledky jsou použity jako srovnávací napříč experimenty.

6.1 Základní experiment

Definování základního experimentu nám mimo jiné usnadní popis dalších experimentů, protože již budou zmiňovány pouze odlišnosti v použitých metodách a v jejich konfiguraci. Pokud při popisu ostatních experimentů nebude uvedeno jinak, byly pro modely řečníků i obecné modely použity směsi Gaussovských rozložení s 256 komponentami. Byly použity příznakové vektory tvořené 12 MFCC příznaky (nebyl použit koeficient c_0 ani dynamické příznaky). K jejich výpočtu byl použit již existující program, vytvořený Laboratoří počítačového zpracování řeči. Výběr trénovacích a testovacích dat byl proveden způsobem popsaným v kapitole 6.1.1 a vyhodnocení identifikace a verifikace odpovídá způsobu uvedenému v kapitole 6.1.2.

6.1.1 Výběr trénovacích a testovacích dat

Jak již bylo uvedeno v kapitole 5, množství dat v trénovací databázi je pro každého řečníka různé. Minimální množství dat pro zařazení mezi referenční řečníky bylo stanoveno na 75 s, toto kritérium bylo splněno pro 306 osob. Nižší nároky na minimální množství dat by byly samozřejmě splněny pro více osob, ale při trénování modelů by mohlo docházet k jejich přeučení. Nekvalitní modely by pak negativně ovlivnily úspěšnost rozpoznávání. Maximální množství dat nebylo při trénování modelů mluvčích žádným způsobem omezeno, některé modely tak byly trénovány s více než 1000 s nahrávek. Při trénování UBM modelů pro muže a ženy bylo nutné maximální množství dat omezit tak, aby nedošlo k vychýlení modelu vlivem převážení mluvčích s největším množstvím dat, výběru nahrávek je v tomto případě věnována

zvláštní pozornost. Pokud je pro daného řečníka dostupných více než 200 s nahrávek, probíhá výběr nahrávek náhodně, s cílem zajistit co možná největší pokrytí různých mikrofonů, akustických podmínek, přenosových kanálů a také časového intervalu, po který byly nahrávky pořizovány. Minimální množství dat v tomto případě nebylo stanoveno a UBM modely tak obsahují hlasy všech osob příslušného pohlaví z trénovací databáze.

Sada dat určená k testovacím účelům byla rozdělena zhruba v poměru 1:2. 75 minut bylo použito ke stanovení verifikačního prahu (viz. kapitola 4.4) a zbylých 155 minut bylo použito k vyhodnocení úspěšnosti rozpoznávání. Nahrávky byly různě dlouhé, nejkratší měla 0,50 s a nejdelší 56 s, průměrná délka testovací nahrávky byla 9 s.

6.1.2 Vyhodnocení identifikace a verifikace

V rámci základního experimentu je identifikace i verifikace vyhodnocována běžným způsobem uvedeným v kapitole 3.2.3. Nejprve se vybere jeden z obecných modelů (ticha, hluku, ženský nebo mužský UBM) na základě kritéria maxima logaritmů věrohodnosti. Nechť je nahrávka reprezentována posloupností příznakových vektorů $X = \{x_1 ... x_T\}$. Věrohodnosti příslušné modelům se vypočítají jako součet pravděpodobností $P(x_t | \lambda^s), t = 1,...,T$. Obecný model se tedy určí jako

$$s^*_{bkg} = \arg\max_s \left[\sum_{t=1}^{T} \log P(x_t | \lambda^s) \right], s = noise, silence, UBM_{male}, UBM_{female}. \quad (6.1)$$

V případě, že model s^*_{bkg} je neřečový, rozpoznávání končí a za výsledek se prohlásí model s^*_{bkg}, v opačném případě se v dalším kroku provede identifikace mluvčího v uzavřené množině, tedy

$$s^* = \arg\max_{s=1,...,S} \left[\sum_{t=1}^{T} \log P(x_t | \lambda^s) \right] \quad (6.2)$$

a následně se totožnost ověří verifikací založenou na testování hypotéz a poměru logaritmů věrohodností. Jako normalizační člen pro verifikaci je použita věrohodnost modelu s^*_{bkg} vypočtená při identifikaci s obecnými modely. Totožnost s^* je přijata, pokud platí

$$\frac{1}{T} \sum_{t=1}^{T} \left(\log P(x_t | \lambda^{s^*}) - \log P\left(x_t \middle| \lambda^{s^*_{bkg}} \right) \right) > \theta. \quad (6.3)$$

Jestliže podmínka (6.3) není splněna, je výsledkem rozpoznávání mluvčích prohlášení, že řečník je neznámý a jeho pohlaví je určeno podle s^*_{bkg}.

6.1.3 Výsledky základního experimentu

V tabulce 6-1 jsou uvedeny výsledky základního experimentu. Z výsledků experimentu vyplývá, že identifikace obecných modelů je schopna spolehlivě určit, zda je předložený segment řečový, nebo neřečový. Řečový segment byl vyhodnocen jako řečový a neřečový jako neřečový ve více než 99 % pokusů. Také pohlaví řečníka bylo určeno správně s podobně vysokou úspěšností. DET křivka na obrázku 4-4 odpovídá vyhodnocení verifikace pro základní experiment.

míra neúspěšnosti identifikace řečových segmentů R_{SNSE}	0,78%
míra neúspěšnosti identifikace pohlaví R_{GDE}	0,99%
míra neúspěšnosti identifikace referenčního řečníka R_{SIE}	10,03%
equal error rate R_{EER}	12,50%
míra neúspěšnosti rozpoznávání R_E	**19,18%**
míra neúspěšnosti rozpoznávání s vlivem na rozpoznávač řeči R_{SAE}	**7,34%**

Tab. 6-1 – *výsledky základního experimentu*

6.2 Vliv množství dat použitého k rozpoznávání

Cílem tohoto experimentu je prověřit, zda není možné zkrátit dobu rozpoznávání výpočtem věrohodnosti pouze přes omezené množství framů, bez výrazného dopadu na úspěšnost rozpoznávání. Běžně se věrohodnost $P\left(X\middle|\lambda^s\right)$ vyhodnocuje přes všechny framy nahrávky, tzn. dle vztahu

$$P\left(X\middle|\lambda^s\right)=\sum_{t=1}^{T}\log P\left(x_t\middle|\lambda^s\right). \tag{6.4}$$

V rámci tohoto experimentu je pravděpodobnost počítána maximálně přes F framů, rovnoměrně rozložených po celé nahrávce, což lze vyjádřit vztahem

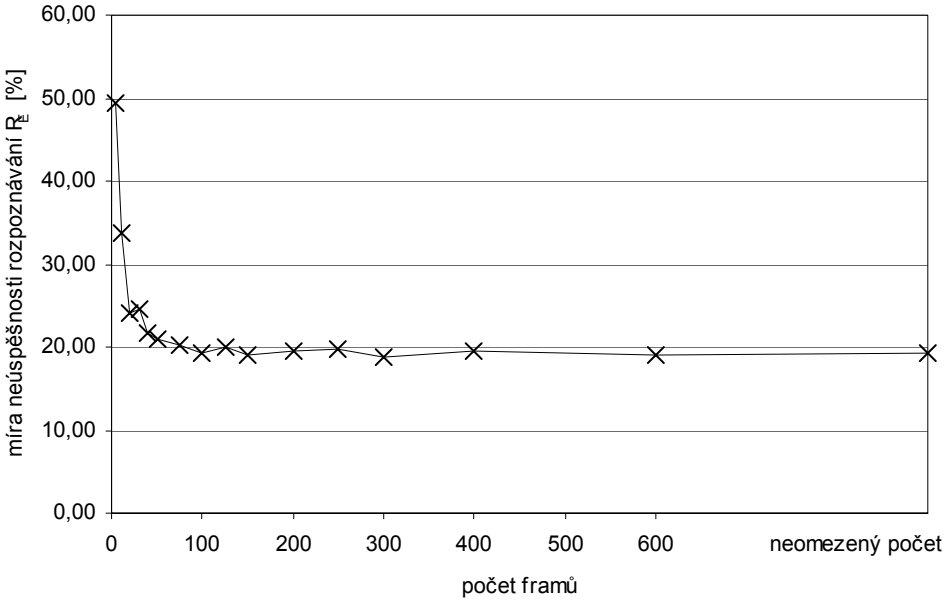
$$P\left(X\middle|\lambda^s\right)=\sum_{f=1}^{F}\log P\left(x_{f\frac{T}{F}}\middle|\lambda^s\right). \tag{6.5}$$

Pokud je nahrávka reprezentována posloupností příznakových vektorů kratší než *F*, není počet framů použitých k výpočtu pravděpodobnosti $P\left(X\middle|\lambda^s\right)$ žádným způsobem omezen a jsou použity všechny framy, pravděpodobnost se tedy určí dle vztahu (6.4). Výsledky tohoto experimentu jsou shrnuty v tabulce 6-2 a znázorněny na obr. 6-1. Průměrné délce 9 s nahrávek v testovací databázi odpovídá počet 900 framů, proto bylo omezení maximálního množství framů použitého k vyhodnocení pravděpodobnosti *F* sledováno do počtu 600 framů.

max. počet framů F	R_{SNSE} [%]	R_{GDE} [%]	R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
5	1,86	5,23	47,96	27,12	49,32	22,02
10	1,17	2,50	26,98	16,99	33,66	11,45
20	0,78	1,99	19,05	13,24	24,07	9,30
30	1,08	1,90	15,02	14,45	24,46	9,10
40	0,68	1,19	13,93	12,79	21,62	7,63
50	0,59	1,49	11,62	12,65	21,04	7,24
75	0,78	1,00	12,32	13,89	20,35	7,63
100	0,59	1,09	10,74	12,40	19,37	7,24
125	0,68	1,19	10,56	12,90	19,96	7,63
150	0,78	1,19	10,21	12,50	**19,08**	**7,05**
200	0,68	1,09	10,39	12,85	19,47	6,95
250	0,68	1,09	9,86	12,40	19,86	7,83
300	0,78	0,90	9,68	13,77	**18,88**	7,14
400	0,78	1,00	9,51	13,77	19,57	8,02
600	0,78	1,00	10,21	12,50	19,08	7,53
neomezený počet	0,78	1,00	10,04	12,50	19,18	7,34

Tab. 6-2 – *závislost měr ohodnocení systému na omezení počtu framů použitých k vyhodnocení věrohodnosti.*

Navzdory očekávání, že čím vyšší je počet framů, přes které je pravděpodobnost vyhodnocována, tím lepší bude úspěšnost rozpoznávání, je výsledkem překvapivé zjištění, že při použití 100 a více framů již trend úspěšnosti rozpoznávání přestává být monotónně se zlepšující a kolísá kolem konstantní hladiny. Tohoto pozorování lze využít ke značnému zkrácení doby rozpoznávání. Při výpočtu pravděpodobnosti přes maximální počet 150 framů je doba rozpoznávání zkrácena více než 6 krát a současně je úspěšnost rozpoznávání ve srovnání se základním experimentem dokonce mírně lepší.



Obr. 6-1 – *graf závislosti míry neúspěšnosti rozpoznávání v závislosti na maximálním množství framů použitém k rozpoznávání*

6.3 Porovnání VQ a GMM

V rámci tohoto experimentu bylo provedeno srovnání výkonu systémů využívajících modely mluvčích reprezentované kódovými knihami s různým počtem centroidů a směsmi Gaussovských rozložení s různým počtem komponent. Identifikace i verifikace byly v případě GMM modelů vyhodnocovány shodně se základním experimentem. Pravděpodobnost $P(X|\lambda)$ byla počítána nejvýše přes 150 framů nahrávky dle vztahu (6.5).

Výpočet celkové vzdálenosti framů nahrávky od centroidů kódové knihy daný vztahem (3.10) je v rámci experimentu upraven v souladu s vyhodnocením pravděpodobnosti. Výsledná vzdálenost je počítána přes omezený maximální počet framů dle vztahu

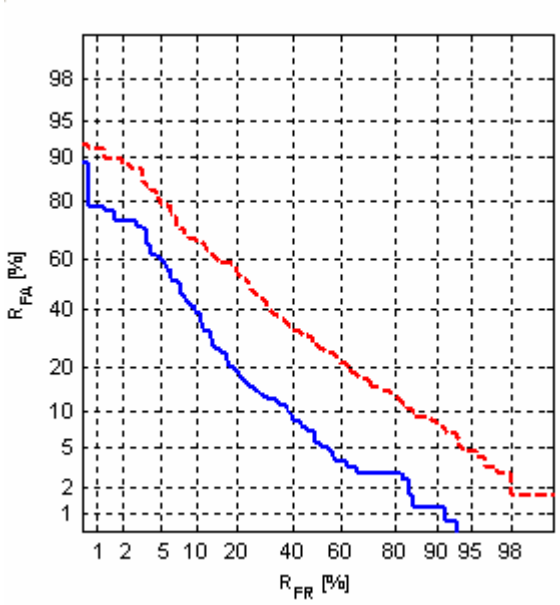
$$D(X,C^s)=\frac{1}{F}\sum_{f=1}^{F}\left(\min_{l=1,\ldots,L^s} d\left(x_{f\frac{T}{F}},c_l^s\right)\right). \tag{6.6}$$

Vzdálenost pro kódové knihy byla počítána pro stejné framy jako pravděpodobnost pro směsi Gaussovských rozložení. Přestože je v kapitole 3.1.2 uvedeno, že rozhodnutí verifikace při použití VQ vychází z prostého porovnání vzdálenosti $D(X,C^s)$ s verifikačním prahem θ, ukázalo se jako užitečné, pokud je tato vzdálenost normována podobným způsobem

jako pravděpodobnost při použití GMM. Normalizace se provede odečtením vzdálenosti vypočtené pro model řečníka od vzdálenosti odpovídající obecnému modelu mužského nebo ženského hlasu. Předkládaná totožnost řečníka je přijata pokud platí

$$D\left(X, C^{s^*_{bkg}}\right) - D\left(X, C^{s^*}\right) > \theta, \tag{6.7}$$

kde $\theta > 0$. V opačném případě je totožnost odmítnuta. Pro ilustraci je na obr. 6-2 znázorněno porovnání DET křivek systému využívajícího pro verifikační rozhodnutí normalizovanou vzdálenost a systému využívajícího přímé porovnání vzdálenosti modelu řečníka s verifikačním prahem. V obou případech byla použita kódová kniha s 64 centroidy.

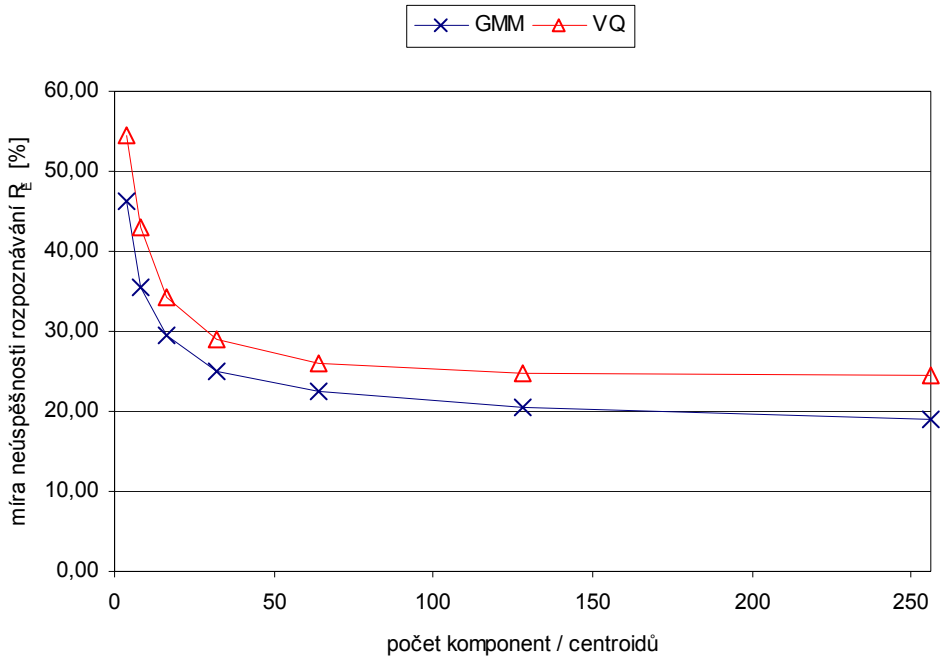


Obr. 6-2 – *DET křivky systémů založených na VQ. Plná modrá čára = verifikace využívající normalizovanou vzdálenost; čerchovaná červená čára = bez normalizace vzdálenosti.*

Výsledky porovnání VQ a GMM jsou uvedeny v tabulce 6-3 a na obr. 6-3. Dosažené výsledky odpovídají očekávání, úspěšnost rozpoznávání při použití směsí Gaussovských rozložení s určitým počtem komponent je ve srovnání s použitím kódových knih se shodným počtem centroidů ve všech případech vyšší. Navíc rozpoznávání při použití GMM je rychlejší. Přestože je výpočet pravděpodobnosti pro jednotlivé komponenty modelu náročnější oproti výpočtu vzdálenosti pro centroidy, při kvantování framu reprezentovaného příznakovým vektorem je nutné určit centroid s minimální vzdáleností, což vyžaduje provedení porovnání, které není nutné při výpočtu pravděpodobností u GMM provádět.

počet komponent / centroidů	R_{SNSE} [%]	R_{GDE} [%]	R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
GMM						
4	2,74	2,03	27,37	34,38	46,18	20,55
8	1,96	1,51	19,50	24,25	35,42	15,07
16	1,96	1,21	14,21	20,54	29,55	13,01
32	1,76	1,11	12,77	18,47	25,05	11,15
64	1,66	1,00	10,64	17,00	22,60	9,10
128	1,17	1,00	9,70	13,25	20,55	8,12
256	0,78	1,19	10,21	12,50	19,08	7,05
VQ						
4	3,03	3,16	43,24	38,51	54,50	26,61
8	2,35	2,02	27,66	30,51	42,95	18,88
16	2,05	1,21	19,61	26,14	34,15	15,85
32	2,05	1,21	13,60	24,41	28,96	12,43
64	1,96	1,31	12,06	19,43	25,93	11,35
128	1,66	1,31	10,09	19,18	24,85	10,76
256	1,76	1,41	8,51	18,75	24,56	11,64

Tab. 6-3 – *porovnání výsledků systémů založených na VQ a GMM pro různé počty centroidů kódových knih a komponent směsí*



Obr. 6-3 –*neúspěšnost rozpoznávání při využití modelů reprezentovaných kódovými knihami s různým počtem centroidů a směsmi Gaussovských rozložení s různým počtem komponent.*

6.4 Vliv množství trénovacích dat

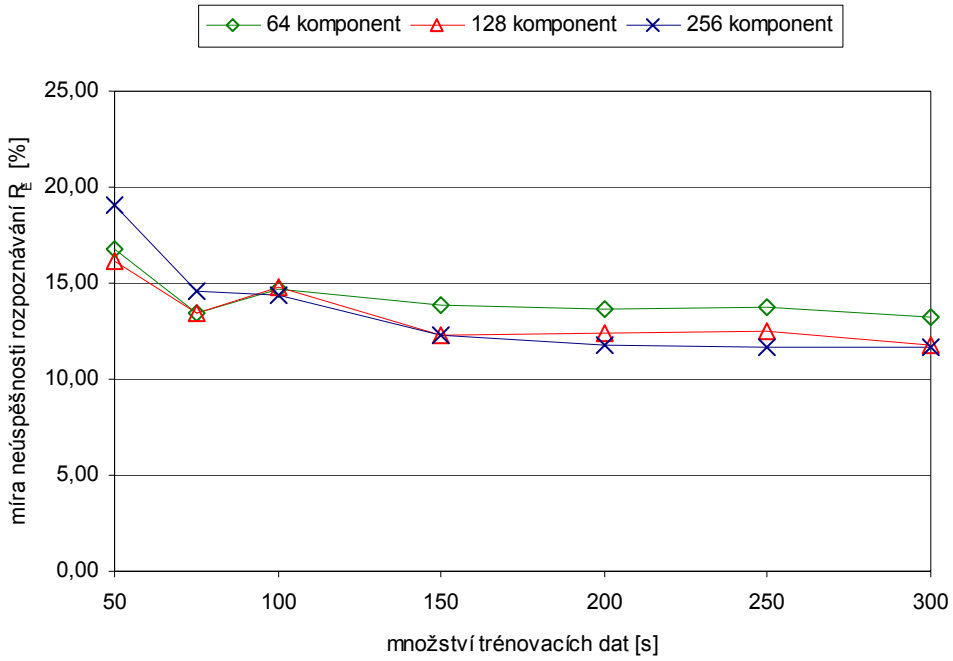
V kapitole 3.2.1 bylo uvedeno, že pokud je směs Gaussovských rozložení trénována na nedostatečném množství dat, může docházet ke vzniku singularit a tedy k tzv. přeučení. Pro 107 řečníků, pro které bylo v trénovací databázi k dispozici více než 300 s nahrávek, byl proveden experiment zabývající se vlivem množství trénovacích dat na úspěšnost rozpoznávání s cílem stanovit alespoň přibližně minimální množství dat nutné pro natrénování kvalitních modelů. Narozdíl od základního experimentu bylo maximální množství dat použité k trénování modelů mluvčích omezeno. Modely všech řečníků tak byly trénovány na základě shodného množství dat, voleného v rámci experimentu v intervalu 50 až 300 s. Přehled výsledků je uveden v tabulce 6-4 a na obr. 6-4.

množství trénovacích dat [s]	**50**	**75**	**100**	**150**	**200**	**250**	**300**
64 komponent							
$\mathbf{R_{SNSE}}$ **[%]**	1,66						
$\mathbf{R_{GDE}}$ **[%]**	1,00						
$\mathbf{R_{SIE}}$ **[%]**	8,44	6,07	6,86	5,28	5,54	5,01	5,01
$\mathbf{R_{EER}}$ **[%]**	10,97	10,76	10,65	9,71	9,65	9,39	10,26
$\mathbf{R_{E}}$ **[%]**	16,73	**13,41**	14,68	13,89	13,60	13,70	13,21
$\mathbf{R_{SAE}}$ **[%]**	10,57	**7,14**	9,30	8,61	9,20	8,32	8,32
128 komponent							
$\mathbf{R_{SNSE}}$ **[%]**	1,17						
$\mathbf{R_{GDE}}$ **[%]**	1,00						
$\mathbf{R_{SIE}}$ **[%]**	8,12	6,81	5,50	6,02	4,71	4,19	4,97
$\mathbf{R_{EER}}$ **[%]**	11,67	9,06	10,61	8,39	9,32	10,06	9,39
$\mathbf{R_{E}}$ **[%]**	**16,14**	**13,41**	14,77	12,33	12,43	12,52	11,74
$\mathbf{R_{SAE}}$ **[%]**	**9,49**	**7,05**	9,59	6,56	7,93	7,44	6,85
256 komponent							
$\mathbf{R_{SNSE}}$ **[%]**	0,78						
$\mathbf{R_{GDE}}$ **[%]**	1,19						
$\mathbf{R_{SIE}}$ **[%]**	9,92	8,62	7,05	4,96	5,22	4,70	4,44
$\mathbf{R_{EER}}$ **[%]**	15,53	9,94	9,52	8,68	9,62	10,13	8,79
$\mathbf{R_{E}}$ **[%]**	19,08	14,58	14,38	12,33	11,74	**11,64**	11,64
$\mathbf{R_{SAE}}$ **[%]**	11,45	7,44	8,32	6,75	6,85	**6,75**	6,56

Tab. 6-4 – *vliv množství trénovacích dat pro směsi Gaussovských rozložení s různým počtem komponent*

Přestože závislost neúspěšnosti rozpoznávání na rostoucím množství trénovacích dat znázorněná na obr. 6-4 je monotónně klesající pouze pro směsi s 256 komponentami

i u ostatních směsí lze pozorovat zlepšující se trend. Výsledky směsí s 256 komponentami jsou při použití malého množství dat výrazně horší ve srovnání s výsledky dosaženými použitím směsí s 128 a 64 komponentami. To přesně potvrzuje očekávání, že pro natrénování kvalitních modelů s vyšším počtem komponent je nutné dostatečné množství trénovacích dat. S rostoucím množstvím dat se kvalita směsí s 256 komponentami zvyšuje a výsledky rozpoznávání jsou ve srovnání s výsledky dosaženými při použití směsí s nižším počtem komponent lepší.



Obr. 6-4 – *závislost neúspěšnosti rozpoznávání na množství trénovacích dat pro GMM s různým počtem komponent*

Na základě výsledků tohoto experimentu se pro reprezentaci modelů mluvčích jeví jako vhodné použití směsí se 128 komponentami a lze předpokládat, že pro trénování kvalitních modelů stačí pouhých 75 s nahrávek. Víceméně na základě tohoto experimentu bylo stanoveno minimální množství dat potřebné pro zařazení mezi referenční řečníky v rámci základního experimentu. Je však nutné si uvědomit, že zatímco v rámci tohoto experimentu bylo množství trénovacích dat mluvčích úmyslně omezeno, aby zůstalo pro trénování modelů všech řečníků shodné, při běžném trénování nemá toto omezení žádný důvod a je vhodné použít k trénování veškerá dostupná data mluvčích. Modely některých mluvčích jsou tak trénovány na mnohem větším množství dat a to umožňuje trénování kvalitních směsí s vyšším počtem komponent. Výsledky uvedené v kapitole 6.3 potvrzují, že při použití směsí s 256 komponentami bylo

dosaženo nejlepších výsledků. Lze předpokládat, že vyšší počet komponent je výhodný jednak proto, že jsou modely velkého množství řečníků trénovány na výrazně více než 75 s nahrávek (pro některé mluvčí je v trénovací databázi k dispozici více než 1000 s), a jednak z důvodu vyššího počtu referenčních řečníků.

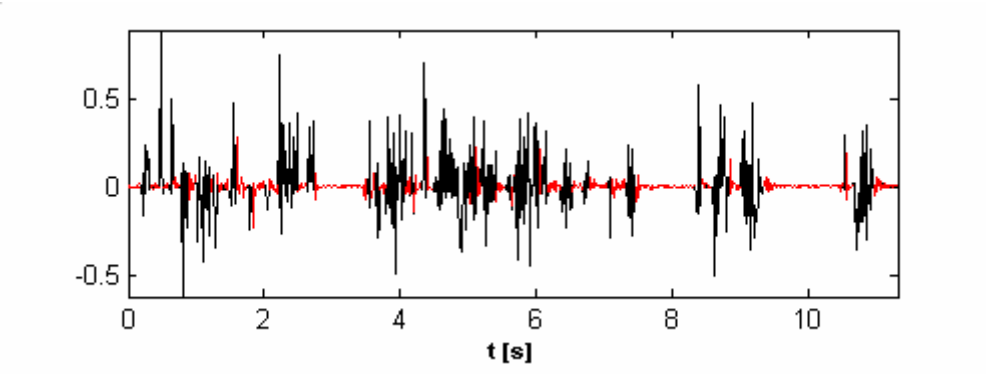
6.5 Detekce hlasových framů

Tiché úseky, které se v nahrávkách objevují například mezi slovy, nenesou žádnou informaci charakteristickou pro mluvčího. Navíc při trénování modelů mluvčích některé komponenty směsi mohou konvergovat do prostoru odpovídajícího těmto nehlasovým příznakovým vektorům. Mnoho systémů rozpoznávání řečníka proto využívá detekci nehlasových framů, které jsou vyřazeny a neúčastní se trénování modelů ani rozpoznávání.

Běžně využívaný přístup zjišťování nehlasových framů je založen na porovnávání energie framů, ke stanovení prahu je využito několik počátečních framů nahrávky. V rámci tohoto experimentu je detekce nehlasových framů prováděna na základě přístupu prezentovaného v [18]. Byla natrénována bi-modální směs Gaussovských rozložení, tedy směs 2 komponent, na základě principu učení bez učitele a s předpokladem, že rozložení hlasových a nehlasových framů bude zcela odlišné a jedna komponenta tak bude zastupovat hlasové framy a druhá nehlasové. Jako nehlasové jsou označeny framy, pro které platí

$$P\left(x_t \middle| G_{speech}\right) < P\left(x_t \middle| G_{silence}\right), \tag{6.8}$$

kde G_{speech} je komponenta představující hlasové framy, zatímco $G_{silence}$ nehlasové. Na rozdíl od modelů mluvčích a obecných modelů byl v tomto případě použit úplný příznakový vektor tvořený 39 příznaky MFCC.



Obr. 6-5 – *detekce nehlasových framů. Hlasové framy jsou znázorněny černě, nehlasové červeně.*

Na obr. 6-5 je znázorněn výsledek detekce nehlasových framů při použití bi-modální GMM. Nové modely mluvčích i obecné modely byly natrénovány s využitím této detekce, pouze na základě hlasových dat. Překvapivě však došlo ve srovnání s výsledky základního experimentu k poklesu úspěšnosti rozpoznávání. Přestože se zdá, že použitý způsob detekce nehlasových framů je schopen tyto framy spolehlivě označit. Lze se pouze domnívat, že problémem je označení více framů, než by ve skutečnosti odpovídalo, jako nehlasových. V průměru byla označena 43 % framů nahrávek jako nehlasová. Výsledky experimentu jsou shrnuty v tabulce 6-5.

detekce nehlasových framů	R_{SNSE} [%]	R_{GDE} [%]	R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
✓	0,49	1,09	11,01	12,89	20,94	7,34
×	0,78	1,00	10,04	12,50	19,18	7,34

Tab. 6-5 – *vliv detekce nehlasových framů*

6.6 Verifikace s adaptovanými modely mluvčích

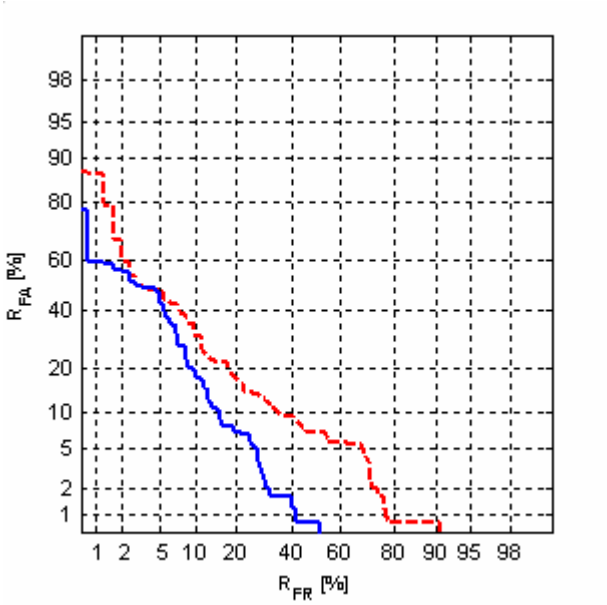
Jak již bylo několikrát uvedeno, k natrénování kvalitních GMM pomocí EM algoritmu je vyžadováno dostatečné množství dat, pokud je dat málo, využívají se k trénování modelů zpravidla metody založené na maximalizaci aposteriorní pravděpodobnosti, která nedostatek dat nahrazuje znalostí apriorního rozložení. V případě trénování modelů mluvčích je apriorní rozložení představováno UBM modelem, který je trénován na velmi velkém množství dat. Dosud však bylo snahou vybírat pouze mluvčí, pro které je v trénovací databázi dostatek dat pro natrénování modelů pomocí EM algoritmu, proč tedy používat metodu MAP? V [5] jsou zmiňovány experimenty, které prokázaly vhodnost použití adaptovaných modelů při verifikaci používající normalizaci věrohodnosti pomocí UBM modelů a je nabízeno vysvětlení výhod spřažení UBM modelů a z nich adaptovaných modelů mluvčích. Vzájemná vazba totiž zaručuje, že výsledek verifikace nebude ovlivněn akustickými jevy, které se neobjevily při trénování modelu mluvčího a mohou se objevit při provozu systému. V průběhu adaptace se totiž adaptují pouze parametry odpovídající akustickým třídám, které jsou zastoupeny v trénovacích datech mluvčího a parametry ostatních akustických tříd UBM modelu se přebírají beze změny. To znamená, že pokud se v průběhu rozpoznávání objeví data z akustických tříd, v průběhu adaptace nezastoupených, je v důsledku použitého kritéria poměru logaritmů věrohodnosti bezvýznamné pro verifikační rozhodnutí. Dobré výsledky verifikace používající modely mluvčích získané metodou MAP potvrzují experimenty publikované v [19].

V [5] je dále prezentován modifikovaný způsob vyhodnocení kritéria poměru logaritmů věrohodnosti, použitelný za předpokladu využití adaptovaných modelů mluvčích a UBM modelů. Vyhodnocení vychází z předpokladu, že GMM pokrývá rozlehlý příznakový prostor a při výpočtu pravděpodobnosti $P(x_t|\lambda)$ pro GMM s velkým množstvím komponent je v blízkém okolí vektoru x_t pouze několik komponent, které přispívají významným způsobem k výsledné hodnotě pravděpodobnosti a lze proto uvažovat o možné aproximaci výpočtem pravděpodobnosti pouze přes C komponent, které přispívají k výsledné hodnotě nejvíce. Dále je využito vazby, která existuje mezi UBM a adaptovanými modely. Vektory blízké určitým komponentám UBM modelu jsou blízké odpovídajícím komponentám adaptovaného modelu. Na základě těchto vlastností probíhá vyhodnocení verifikace tak, že se nejprve určí C nejlépe ohodnocených komponent UBM modelu a následně se provede vyhodnocení pravděpodobnosti pro model mluvčího pouze na základě odpovídajících komponent. Pokud má UBM model L komponent je nutné provést $L+C$ vyhodnocení pravděpodobnosti vícerozměrného Gaussovského rozložení ve srovnání s provedením $2L$ těchto vyhodnocení při běžném přístupu. V rámci tohoto experimentu bylo zvoleno $C=5$.

Výpočet parametrů adaptovaných modelů odpovídá způsobu uvedenému v kapitole 3.2.2. Byly adaptovány pouze střední hodnoty komponent, jejich váhy i kovariační matice byly převzaty z UBM modelu ($\xi_l^w = \xi_l^\Sigma = 0, \forall l$). V tabulce 6-6 jsou shrnuty výsledky související s ohodnocením verifikace a úspěšnosti rozpoznávání pro modely adaptované s různými hodnotami váhového faktoru r^μ. Z výsledků je zřejmé, že různé volby váhového faktoru nemají na hodnotu EER, ani úspěšnost rozpoznávání výrazný vliv. Zajímavější však je, že ve srovnání se základním experimentem došlo k výraznému zvýšení EER a poklesu úspěšnosti rozpoznávání. Slabý výsledek verifikace využívající adaptované modely mluvčích je dobře patrný z obr. 6-6, kde je provedeno srovnání DET křivky pro verifikaci s modely adaptovanými s váhovým faktorem $r^\mu = 16$ a DET křivky verifikace základního experimentu.

váhový faktor r^μ	R_{EER} [%]	R_E [%]	R_{SAE} [%]
2	18,59	24,85	10,08
6	18,59	25,15	10,27
10	18,59	25,34	10,57
16	18,79	25,15	10,47

Tab. 6-6 – *vybrané míry ohodnocení systému v závislosti na váhovém faktoru r^μ při aplikaci metody MAP za účelem adaptace středních hodnot modelů mluvčích*



Obr. 6-6 – *Porovnání DET křivek verifikace využívající modely mluvčích trénované metodou MAP – červená čerchovaná čára a metodou ML (EM algoritmem) – modrá plná čára.*

Navzdory očekávání dobrých výsledků, jakých bylo dosaženo např. v [19], se použití tohoto přístupu v navrhovaném systému neosvědčilo. Možným vysvětlením je výrazně odlišné množství dat pro jednotlivé referenční řečníky v trénovací databázi. Střední hodnoty komponent mluvčích s menším množstvím trénovacích dat tak budou bližší hodnotám UBM modelu. Pokud bude neznámý řečník správně identifikován, ale parametry jeho adaptovaného modelu budou v důsledku malého množství trénovacích dat blízké parametrům UBM modelu, bude vyhodnocen malý rozdíl věrohodnosti modelu mluvčího a UBM modelu. Pokud by byly modely mluvčích adaptovány na základě shodného množství dat, lze předpokládat, že vzdálenost (ve smyslu středních hodnot adaptovaných v rámci tohoto experimentu) všech těchto modelů bude při vyhodnocení přes všechny komponenty zhruba srovnatelná a verifikační práh tak bude přizpůsoben odpovídajícím nízkým rozdílům věrohodností. Pokud jsou ovšem modely některých mluvčích adaptovány s nesrovnatelně (v případě použité databáze až 10 krát) větším množstvím dat, je zřejmé, že tyto modely budou výrazně vzdálenější UBM modelu a vyhodnocený rozdíl věrohodností bude větší. Poměr věrohodností vyhodnocovaný pro verifikační rozhodnutí je mimo jiné závislý na množství dat dostupném při adaptaci modelu mluvčího. Souvislost mezi adaptovanými modely a UBM modely tak nepřináší pouze výhody zmíněné na začátku této kapitoly, ale v případě rozdílného množství adaptačních dat pro mluvčí situaci spíše komplikuje. Pouze pro úplnost uveďme, že v [19] byly modely všech mluvčích adaptovány na základě shodného množství dat.

6.7 Upravené vyhodnocení identifikace mluvčích

Cílem tohoto experimentu je upravit vyhodnocení identifikace mluvčích tak, aby došlo ke zrychlení a současně zvýšení úspěšnosti rozpoznávání. Za účelem zrychlení byl použit přístup, který je výsledkem předešlých zkušeností Laboratoře počítačového zpracování řeči v oblasti rozpoznávání mluvčích, publikovaný např. v [20]. Spočívá v provedení identifikace ve dvou krocích. Nejprve se použijí modely mluvčích s nižším počtem komponent a do druhé fáze identifikace, kdy se použijí modely s vyšším počtem komponent, postupuje pouze omezený počet mluvčích. Tento počet není stanoven pevně, ale určí se na základě relativních rozdílů věrohodností vyhodnocených pro modely mluvčích a nejvyšší dosažené věrohodnosti. Konkrétně byl zvolen relativní odstup nejvýše 4 % od nejúspěšnějšího mluvčího.

V tabulce 6-7 jsou uvedeny výsledky pro několik konfigurací identifikace mluvčích. Věrohodnost byla vyhodnocována přes 150 framů. Při srovnání s výsledky základního experimentu je zřejmé, že při použití 256 komponent v druhé fázi nedošlo k poklesu úspěšnosti a výsledek odpovídá vyhodnocení přes omezený počet framů (viz tabulka 6-2). Jako nejlepší se zdá být použití 32 komponent v první fázi, kdy byla doba rozpoznávání nejkratší. Přestože vyhodnocení modelů s 16 komponentami je rychlejší, modely s 32 komponentami se pravděpodobně chovají více diskriminačně a do druhé fáze tak postupuje menší počet mluvčích. Při použití 32 komponent v první a 256 v druhé fázi byla doba rozpoznávání zkrácena více než 2 krát ve srovnání s použitím 256 komponent v jedné fázi.

počet komponent modelů v první / druhé fázi identifikace	**R_{SIE} [%]**	**R_{EER} [%]**	**R_E [%]**	**R_{SAE} [%]**
16 / 128	9,86	14,46	20,74	8,32
16 / 256	10,21	12,50	19,08	7,05
32 / 256	10,21	12,50	19,08	7,05
64 / 256	10,21	12,50	19,08	7,05

Tab. 6-7 – *porovnání výsledků pro několik konfigurací identifikace mluvčích*

Problémem klasického výpočtu pravděpodobnosti $P(X|\lambda)$ jako součtu (6.2) je, že pokud některý z framů nahrávky představuje příznakový vektor ležící na okraji rozložení definovaného modelem, je pravděpodobnost vyhodnocená pro tento frame velmi malá a je-li v nahrávce takových hodnot více, mohou velmi výrazně ovlivnit výslednou pravděpodobnost. Podrobněji se tímto efektem zabývá [21]. Snahou je zabránit tomuto dominantnímu vlivu několika špatných framů na chybný výsledek identifikace. Je nutné, aby vliv všech framů byl

více vyvážený. Řešením může být použití pravidla rozhodovacího hlasu každého framu (frame voting, FV). Každý frame se chová jako samostatný klasifikátor

$$s_t^* = \arg\max_{s=1,\ldots,S} P\left(x_t \middle| \lambda^s\right). \tag{6.9}$$

Při klasickém vyhodnocení se výsledky klasifikace framů kombinují násobením pravděpodobností (resp. sčítáním jejich logaritmů). Přístup založený na FV přiřazuje mluvčím na základě výsledků klasifikace jednotlivých framů hlasy a na závěr je vybrán mluvčí, který získal nejvyšší počet hlasů. Pro uvažovaný systém není možné přímé použití tohoto přístupu, protože hlasů je pouze omezené množství odpovídající počtu framů. V kapitole 6.1.1 bylo uvedeno, že průměrná nahrávka je tvořena zhruba 900 framy. Referenčních řečníků je více než 300. Při takto vysokém počtu hrozí, že se hlasy rozdělí mezi mnoho řečníků a žádný nebude počtem hlasů výrazně převyšovat ostatní, může se dokonce stát, že maximální počet hlasů bude shodný pro několik řečníků. Analýzou dosavadních výsledků identifikace založené na klasickém vyhodnocení bylo zjištěno, že v případech, kdy správný řečník není vyhodnocen jako nejpravděpodobnější, umísťuje se téměř vždy mezi několika prvními řečníky. Tohoto poznatku využijeme k omezení počtu mluvčích, pro které bude vyhodnocován FV. Identifikace tak bude probíhat celkem ve třech krocích. V prvních dvou krocích se provádí klasické vyhodnocení věrohodnosti jako součtu logaritmů pravděpodobností pro framy. Nejprve se použijí modely s 32 komponentami a dále s 256 komponentami. Tyto věrohodnosti se opět vyhodnocují přes omezený počet nejvýše 150 framů. Prvních N mluvčích pak postupuje do poslední fáze, ve které se použije pravidlo rozhodujícího hlasu každého framu. N je v tomto případě stanoveno pevně, aby nedocházelo k rozdělování hlasů mezi nekontrolovatelně velký počet mluvčích. V rámci tohoto experimentu bylo zvoleno $N = 5$. S cílem zajistit co největší robustnost se této fáze účastní všechny framy nahrávky. Výsledky jsou uvedeny v tabulce 6-8.

R_{SIE} [%]	R_{EER} [%]	R_E [%]	R_{SAE} [%]
11,44	12,05	18,69	6,56

Tab. 6-8 – *výsledky dosažené při použití identifikace kombinující klasické vyhodnocení věrohodnosti GMM (zde ve 2 fázích) a pravidlo rozhodujícího hlasu každého framu.*

Přestože došlo dokonce k mírnému zhoršení úspěšnosti identifikace mluvčích, identifikace chybuje v situacích, které jsou příhodnější z pohledu verifikace (hodnota EER se snížila) a ve výsledku dochází k zvýšení celkové úspěšnosti rozpoznávání. Při použití FV se doba rozpoznávání mírně prodloužila, oproti základnímu experimentu je však uvedený způsob rozpoznávání více než 9 krát rychlejší. Výsledky uvedené v tabulce 6-8 jsou nejlepší, jakých se podařilo s navrhovaným systémem dosáhnout.

7. Závěr

7.1 Provedené výzkumné a vývojové práce

V souladu se zadáním byla na základě studia současné světové literatury a sborníků mezinárodních vědeckých konferencí provedena analýza současného stavu problematiky textově nezávislého rozpoznávání mluvčích. V jazyce C byla implementována řada metod založených na VQ a GMM, provádějících identifikaci a verifikaci mluvčích. Na základě těchto metod byl vytvořen systém rozpoznávání mluvčích v záznamech televizních a rozhlasových pořadů popsaný v kapitole 4. Dále byla vytvořena sada programů sloužící k trénování modelů mluvčích, v případě kódových knih pomocí LBG algoritmu a v případě směsí Gaussovských modelů metodou maximální věrohodnosti (EM algoritmem) a metodou maximální aposteriorní pravděpodobnosti. Nakonec byla provedena řada experimentů s cílem nalézt vhodnou metodu identifikace a verifikace a jejich vhodné nastavení.

7.2 Stručný přehled dosažených výsledků

Většina pozornosti byla soustředěna na metody založené na směsích Gaussovských rozložení. Výsledky uvedené v kapitole 6.3 potvrzují, že úspěšnost rozpoznávání při použití modelů založených na VQ je výrazně nižší, pro různé velikosti kódových knih a odpovídající velikosti GMM v průměru zhruba o 5 %.

Jako velmi užitečné se osvědčilo vyhodnocení věrohodnosti pro GMM přes omezený počet framů. Výsledky uvedené v kapitole 6.2 ukazují, že při vyhodnocení věrohodnosti přes 150 framů nedochází k žádnému poklesu úspěšnosti rozpoznávání a současně je dosaženo více než 6-ti násobného zrychlení.

Dále byl proveden experiment zabývající se odstraněním nehlasových framů. K označení nehlasových framů byla použita bi-modální Gaussovská směs. Přestože se zdá, že použitý způsob detekce nehlasových framů je schopen tyto framy spolehlivě označit, došlo překvapivě ke zhoršení rozpoznávacího skóre. Pravděpodobným problémem může být označení více framů, než by ve skutečnosti odpovídalo, jako nehlasových. V průměru byla označena 43 % framů nahrávek jako nehlasová.

V mnoha publikacích je uváděno zlepšení výsledků verifikace, dosažené aplikací metody MAP. Výsledky experimentu zabývajícího se použitím tohoto přístupu v navrhovaném

systému ale ukazují, že při použití modelů mluvčích adaptovaných metodou MAP došlo k poklesu úspěšnosti rozpoznávání. Možným vysvětlením je, že modely mluvčích jsou adaptovány na základě výrazně odlišného množství dat.

Zlepšení úspěšnosti rozpoznávání a dalšího zrychlení bylo dosaženo úpravou vyhodnocení identifikace mluvčích. Úprava vedoucí ke zrychlení spočívá ve vyhodnocení ve dvou krocích, v první fázi s menším počtem komponent a ve druhé s větším, avšak pro menší počet mluvčích. V kapitole 6.7 je ukázáno, že tato úprava nemá žádný negativní vliv na úspěšnost rozpoznávání a umožňuje přitom zkrácení doby rozpoznávání. Zvýšení úspěšnosti rozpoznávání je dosaženo použitím pravidla rozhodovacího hlasu každého framu. Toto vyhodnocení se provádí jako další krok, pro mluvčí vyhodnocené v předchozích fázích jako nejpravděpodobnější. Aplikací této úpravy identifikace mluvčích se podařilo docílit nejlepších výsledků.

Při poměrně vysokém počtu 306 referenčních řečníků a za víceméně obecných akustických podmínek televizních a rozhlasových záznamů se podařilo dosáhnout úspěšnosti rozpoznávání více než 81 %. Navrhovaný systém také poskytuje informaci využitelnou rozpoznávačem řeči pro výběr vhodného adaptovaného modelu, úspěšnost rozpoznávání z tohoto pohledu dosahuje více než 93 %.

7.3 Využití výsledků

Na základě navrhovaného systému byl v rámci této práce vytvořen modul rozpoznávání řečníka začleněný do systému kompletního automatického přepisu televizních a rozhlasových pořadů vyvíjeného Laboratoří počítačového zpracování řeči.

7.4 Náměty na další práci

Zvýšení úspěšnosti rozpoznávání lze navzdory dosaženým výsledkům očekávat od použití detekce hlasových framů, založené na odlišném principu.

Přestože existuje řada různých modifikací vyhodnocení věrohodnosti GMM, které nebyly pro navrhovaný systém ověřeny, je nutné si uvědomit, že možnosti zlepšení vycházející z těchto modifikací jsou omezené, protože nepřináší žádnou novou informaci. Nevýhodou použitých příznaků MFCC je závislost na akustických podmínkách. Jako inspirující se proto jeví především využití vyšších příznaků, přinášejících informaci např. o prosodii nebo intonaci.

Seznam literatury

[1] NOUZA J. (editor): Počítačové zpracování řeči. TUL Liberec 2001.

[2] RADOVÁ, V.: Rozpoznávání řečníka. Habilitační práce. ZČU Plzeň 2004.

[3] CAMPBELL, J.P.: Speaker Recognition: A Tutorial. Proceedings of the IEEE, vol. 85, no. 9, September 1997.

[4] REYNOLDS, D.A.: Speaker Identification and Verification using Gaussian Mixture Speaker Models. MIT Lincoln Laboratory, March 1995.

[5] REYNOLDS, D.A., QUATIERI, T.F., DUNN, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10, 19-41, 2000.

[6] GAUVAIN, J., LEE, CH.: Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, IEEE Trans. on Speech & Audio, April 1994.

[7] FURUI, S.: 50 Years of Progress in Speech and Speaker recognition, SPECOM 2005.

[8] REYNOLDS, D.A.: An Overview of Automatic Speaker Recognition Technology, in Proc. Int. Conf. on Acoustic, Speech, and Signal Processing, 2002, vol. 4, pp. 4072-4075

[9] LOURADOUR, J., ANDRÉ-OLBRECHT, R., DAOUDI, K.: Segmentation and Relevance Measure for Speaker Verification, Interspeech 2004, 1401-1404.

[10] SOMOL, P.: Different approaches to initialization of the EM algorithm for use in Gaussian mixture modelling methods. In: Multidisciplinární přístupy k podpoře rozhodování v ekonomii a managementu, Workshop '97 grantu VS 96063. (Plešingr, J. ed.), FM JU, Jindřichův Hradec 1997, pp. 85–91.

[11] PAALANEN, P.: Bayesian Classification Using Gaussian Mixture Model and EM Estimation: Implementation and Comparisons. Information Technology Project. Lappeenranta, 2004.

[12] FIGUEIREDO, M.A.T., JAIN, A.K.: Unsupervised Learning of Finite Mixture Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):381-396, 2002.

[13] VLASIS, N., LIKAS, A.: A Greedy EM Algorithm for Gaussian Mixture Learning. Neural Processing Letters, 2000.

[14] PADRTA, A., RADOVÁ, V.: On the Amount of Speech Data Necessary for Successful Speaker Identification. In: Proceedings of the 8th European Conference on Speech Communication and Technology EUROSPEECH 2003, Geneva, Switzerland, September 2003, pp. 3021-3024, ISSN 1018-4074.

[15] PADRTA, A., RADOVÁ, V.: On the Background Model Construction for Speaker Verification Using GMM. Text, Speech and Dialogue TSD 2004. Lecture Notes in

Artificial Intelligence 3206. Springer-Verlag, Berlin, Heidelberg, 2004, pp. 425-432, ISBN 3-540-23049-1, ISSN 0302-9743.

[16] PADRTA, A., RADOVÁ, V.: Comparison of Several Speaker Verification Procedures Based on GMM. In: International Conference on Spoken Language Processing ICSLP 2004, Korea, 2004 1777-1780.

[17] REYNOLDS, D.A., CAMPBELL, J.P.: Corpora for the Evaluation of Speaker Recognition Systems. In: Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999.

[18] MARIÉTHOZ, J., BENGIO, S.: An Alternative to Silence Removal for Text-Independent Speaker Verification. IDIAP-RR 03-51, 2003.

[19] MARIÉTHOZ, J., BENGIO, S.: A Comparative Study of Adaptation Methods for Speaker Verification. IDIAP-RR 01-34, 2001.

[20] DAVID, P.: Presentation of Real-time System for Automatic Speaker Identification and Verification. In Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics – SCI 2003. Orlando-USA, July 2003. Volume IV. pp. 372-376

[21] NARAYANASWAMY, B.: Improved Text-Independent Speaker Recognition using Gaussian Mixture Probabilities. A Report in Candidacy for the Degree of Master of Science. Carnegie Mellon University, May 2005.

[22] NOUZA, J., ŽĎÁNSKÝ, J., DAVID, P., ČERVA, P.,KOLORENČ, J., NEJEDLOVÁ, D.: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In: Interspeech 2005, September, 2005, Lisboa, Portugal, pp. 1681-1684, ISSN 1018-4074.