TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky a mezioborových
inženýrských studií

Disertační práce

2006 Petr David

TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových
inženýrských studií

IDENTIFIKACE AUDIOSEGMENTŮ PRO AUTOMATICKOU TRANSKRIPCI ZPRAVODAJSKÝCH POŘADŮ

DISERTAČNÍ PRÁCE

2006

PETR DAVID

IDENTIFIKACE AUDIOSEGMENTŮ PRO AUTOMATICKOU TRANSKRIPCI ZPRAVODAJSKÝCH POŘADŮ

DISERTAČNÍ PRÁCE

Disertant:	Ing. Petr David
Studijní program:	2612V Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Katedra elektroniky a zpracování signálů Fakulta mechatroniky a mezioborových inženýrských studií Technická univerzita v Liberci
Školitel:	Prof. Ing. Jan Nouza, CSc.

ROZSAH PRÁCE:

Počet stran:	148
Počet obrázků:	48
Počet vzorců:	104
Počet tabulek:	23
Počet příloh:	4

Anotace

Tato disertační práce pojednává o metodách identifikace audiosegmentů v úloze automatické transkripce zpravodajských pořadů. Vzhledem ke stanovenému tématu se jedná o práci, jež zasahuje do několika oblastí automatického zpracování řeči. Autor se postupně věnuje úlohám identifikace mluvčího, identifikace pohlaví, verifikace mluvčího a nalezení řečových segmentů v záznamech televizního zpravodajství. V návaznosti na tato témata je představen i real-time systém pro identifikaci a verifikaci mluvčích, jenž vznikl v průběhu vývoje jako funkční ukázková aplikace ověřující chování rozpoznávacího software v reálných podmínkách. Posledním tématem je evropská databáze televizního zpravodajství COST278-BN, na jejímž vzniku se autor aktivně podílel a jež je posléze využita v mezinárodní evaluační kampani.

Práce je strukturována následujícím způsobem:

Kapitola 1 je věnována úvodnímu slovu a obecnému popisu řešené úlohy.

Kapitola 2 stručně představuje systém pro automatický přepis zpravodajských pořadů vyvíjený na Technické univerzitě v Liberci, jeho jednotlivé části a v závěru vymezuje autorovu účast v tomto projektu.

Kapitola 3 pojednává o předzpracování, segmentaci a parametrizaci audiosignálu pro pozdější použití v jednotlivých úlohách rozpoznávání řeči.

Kapitola 4 se věnuje teoretickému rozboru úlohy rozpoznávání řečníka a popisu metod a algoritmů vhodných k jejímu řešení. V závěru je provedena řada experimentů k určení optimálního nastavení parametrů implementovaných metod.

Kapitola 5 se zabývá metodami identifikace řečových segmentů.

Kapitola 6 popisuje databázi COST278-BN a vyhodnocovací nástroje.

Kapitola 7 uveřejňuje dosažené výsledky testů se zpravodajskými daty.

Kapitola 8 hodnotí výsledky dosažené v této práci a naznačuje směr, kam by se mohl ubírat další výzkum navazující na současný stav.

Abstract

This thesis deals with the problem of automatic audio segment identification in Broadcast News (BN) transcription. Taking subject matter into account several speech recognition areas are discussed. The author subsequently addresses speaker identification and verification, gender identification and speech-nonspeech detection with regard to BN specificity. This thesis also discusses the real-time system for automatic speaker identification and verification which was developed during BN system development as a practical application of recognition software in real world conditions. The last subject matter is the pan-European BN database COST278-BN in the creation of which the author played an active role. The results of an international evaluation campaign with COST278-BN database are also discussed.

The thesis consists of the following parts:

Chapter 1 is dedicated to introductory words. General information about the solved problem is also brought to light within this chapter.

Chapter 2 briefly describes complete system for automatic BN transcription which is developed at TU of Liberec and also describes the author's participation in this project.

Chapter 3 contains the details about preprocessing, segmentation and feature extraction from an acoustic speech signal.

Chapter 4 deals with the theoretical background of speaker recognition tasks and outlines the principles of the state-of the-art methods and algorithms implemented therein.

Chapter 5 is dedicated to the speech-nonspeech detection problem.

Chapter 6 describes the COST278-BN database and the evaluation tools.

Chapter 7 discusses recognition results obtained from BN data (see chapter 6).

Chapter 8 concludes the thesis and implies new operating orientations which could take up again on the recent recognition system.

Obsah

Seznam obrázků

Seznam tabulek

Seznam použitých zkratek

zkratka	český název (pokud existuje)	anglický název
$\bar{x}$	vektor středních hodnot	vector of mean values
Σ	kovarianční matice	covariance matrix
ANN	umělé neuronové sítě	artificial neural networks
BIC	bayesovské informační kriterium	bayesian information criterion
BN	televizní a rozhlasové zpravodajství	broadcast news
BPNN	neuronová síť se zpětným šířením chyb	backpropagational neural net.
CMS	odečítání kepstrálního průměru	cepstral mean subtraction
COST	evropský projekt kooperace v oblasti vědy a technologií	cooperation in science and technology
DCT	diskrétní kosinová transformace	discrete cosine transformation
DET		detection error tradeoff
DFT	diskrétní Fourierova transformace	discrete Fourier transformation
DTW	dynamické borcení času	dynamic time warping
EER		equal error rate
EM	algoritmus stř. hodnota – maximalizace	expectation-maximization alg.
FFT	rychlá Fourierova transformace	fast Fourier transformation
Fs	vzorkovací frekvence	sample frequency
GD	na pohlaví závislý (model)	gender dependent
GMM	gaussovské modely směsí	gaussian mixture models
HMM	skryté markovské modely	hidden markov models
IDFT	inverzní diskrétní Fourierova transformace	inverse discrete Fourier transformation
IM	identifikace mluvčího	speaker identification
IP	identifikace pohlaví	gender identification
KK	kódová kniha	code book
*k*NN	metoda *k* nejbližších sousedů	*k* nearest neighborhoods
LBG	algoritmus k tvorbě KK	Linde, Buzo and Gray algorithm
LR	hodnota věrohodnosti	likelihood ratio
LVCSR	rozpoznávání spojité řeči s velkým slovníkem	large vocabulary continuous speech recognition

MFCC	melovské frekvenční kepstrální koeficienty	mel-frequency cepstral coefficients
MLP	vícevrstvé perceptronové sítě	multi-layer perceptron
NIST		national institute of standards and technology
NN	nejbližší soused	nearest neighborhood
PDF	hustota pravděpodobnosti	probability density function
ROC		receiver operating characteristic
SI	na mluvčím nezávislý	speaker independent
SNR	odstup signálu od šumu	signal to noise ratio
STM		sclite segment time mark
SVM		support vector machines
TRS	transkripční formát Transcriberu	transcriber transcription format
UBM	univerzální model mluvčích	universal background model
UNICODE	standard pro vícebytové kódování znaků národních abeced	unique number for every national character
UTF		universal test framework
VM	verifikace mluvčího	speaker verification
VQ	vektorová kvantizace	vector quantization
WER	míra chybně rozpoznaných slov	word error rate
WERR	relativní zlepšení míry chybně rozpoznaných slov	word error rate reduction
XML	rozšířitelný značkovací jazyk	extensible markup language

1. Úvod

1.1. Popis úlohy

V několika posledních letech byla značná část výzkumného úsilí v oblasti zpracování řeči věnována problematice automatické transkripce audiovizuálních pořadů. V následující kapitole bude popsán systém pro plně automatický přepis zpravodajství v českém jazyce. Využití takového systému je samozřejmě možné rozšířit na úlohy týkající se přepisu téměř jakéhokoliv pořadu, v němž je hlavní informace sdělována lidským hlasem (parlamentní debaty, záznamy rozhovorů, navigace v rozsáhlých audionahrávkách nebo přepis a indexace diskusních pořadů).

Příčina, proč je v dnešní době o tyto systémy stále větší zájem, je ve stále rostoucím objemu multimediálních dat, mezi kterými je třeba vyhledávat, třídit je a efektivně je archivovat. Zajímavou aplikací je i filtrace multimediálních dat podle požadavků uživatelů. Ti chtějí obdržet pouze setříděná data vztahující se k jejich požadavkům. Jako příklad by bylo možné uvést zařízení pro cílené nahrávání konkrétně zaměřených příspěvků, dále pak multimediální výukové záznamy týkající se daného tématu, nebo multimediální archívy a knihovny. Důvodem bránícím většímu rozšíření byla v minulých letech složitost řešeného problému. Vývojáři, kteří takový systém chtějí vyvinout, musí mít přístup k velkému množství pokročilých technologií z oboru zpracování řeči. Proto jsou nejznámější systémy v této oblasti vyvíjeny na platformě velkých světových laboratoří jako například AT&T, BBN, LIMSI nebo IBM. Systémy pro některé světové jazyky jako je angličtina, francouzština, španělština nebo japonština jsou popsány v literatuře [GAU99], [BAC03], [NGU02] nebo [MCT03]. Na druhou stranu je mnoho jazyků, u nichž jde vývoj mnohem pomaleji. To jsou především jazyky, u kterých je nutné použít rozsáhlejší slovníky[1]. Potřeba takto velkých slovníků vzniká u ohebných jazyků, které se vyznačují skloňováním a časováním (typicky to platí pro slovanské jazyky jako ruština, čeština, polština nebo slovenština). Dále to mohou být i jazyky zahrnující slučování (němčina) nebo aglutinaci (finština, turečtina nebo maďarština). Velikost slovníku může dosahovat u těchto jazyků až na milión slov. Takto rozsáhlý slovník neovlivňuje pouze rychlost a přesnost rozpoznávání, ale způsobuje problémy i při tvorbě a uložení odpovídajícího jazykového modelu.

Systém uvedený v této práci a vyvíjený Laboratoří počítačového zpracování řeči (SpeechLab) na Technické univerzitě v Liberci je založen na modulární architektuře, která

[1] Tzn. s obsahem více než obvyklých sedmdesát tisíc slov, což je typický limit u většiny komerčně nebo volně dostupných systémů pro automatický přepis zpravodajství.

umožňuje v případě potřeby přidání dalšího modulu (v poslední době to byl například modul pro zpracování výsledného textového výstupu) nebo nahrazení modulu jiným, rychlejším případně výkonnějším modulem. Takto zvolená architektura navíc umožňuje nasazení tohoto systému i pro další jazyky. To ovšem platí pouze v tom případě, že pro daný jazyk máme k dispozici modul pro rozpoznávání spojité řeči. Jaké jsou tedy úkoly v případě přepisu zpravodajských pořadů? V první řadě se jedná o rozčlenění kontinuálního audiosignálu na řečové/neřečové části a následné rozdělení řečových částí podle hovořících osob, případně podle změn v akustickém pozadí záznamu. Popis takto rozčleněného signálu je možný na základě identity mluvčího a jeho pohlaví. Předpokládá se, že segmenty jednotlivých mluvčích budou mít v rámci jednoho zpravodajství unikátní číslo. Další přidanou informací může být druh akustického pozadí (ruch, hudba) nebo akustická kvalita daného segmentu (studiová řeč, reportáž z rušného místa, rozhovor uskutečněný přes telefonní linku). V neposlední řadě je tu také nejdůležitější část přepisu zpravodajského pořadu, získání vlastního textu promluvy.

1.2. Cíle disertační práce

Základním cílem této disertační práce je navržení, implementace a reálné otestování metod pro identifikaci audiosegmentů[1] v úloze automatického přepisu zpravodajství. Na základě modulární architektury i s ohledem na aplikovatelnost problematiky byly vymezeny čtyři základní úlohy:

1. parametrizace signálu,
2. nalezení řečových segmentů,
3. identifikace mluvčího a pohlaví,
4. verifikace mluvčího.

Zpracování řeči poskytuje pro řešení vytýčených úloh celou řadu metod. V této práci budou použity přístupy založené na metodě gaussovských modelů směsí[2], vektorové kvantizace a skrytých markovských modelů.

Kromě úspěšnosti jednotlivých modulů budou výstupy těchto řešení sloužit ke zlepšení celkového rozpoznávacího skóre, tj. ke zkvalitnění výsledného textového přepisu zpravodajství. Z tohoto důvodu bude v práci provedena celá řada experimentálních ověření přínosu uvedených metod na kompletní systém.

---

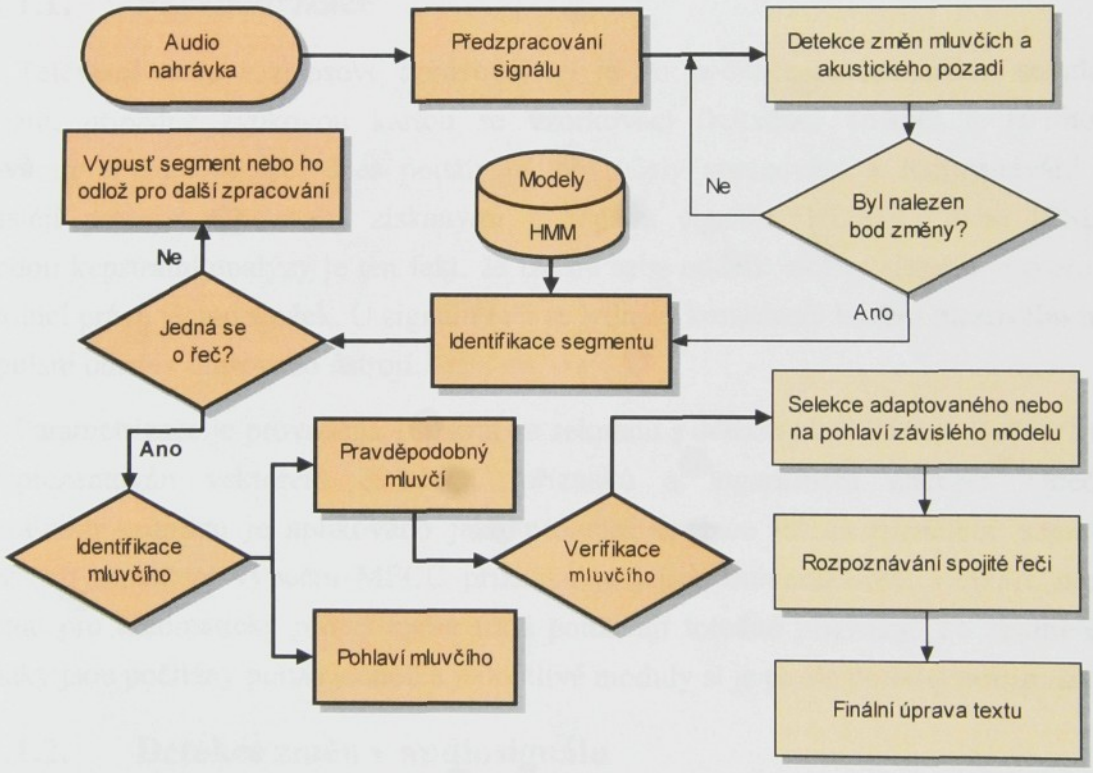
[1] Výraz „segment" má v této práci význam libovolně dlouhého signálu řeči, který vznikl manuálním či automatickým rozdělením původního spojitého audiosignálu. Jednotlivé segmenty na sebe bez překrývání navazují, což je spolu s jejich proměnnou délkou největší rozdíl oproti později používanému výrazu „frame".

[2] V českém jazyce zatím není zavedena jednotná terminologie, proto je možné setkat se v literatuře i s označením „směs gaussovských rozložení", „gaussovské mixturové modely" nebo „gaussovské složkové modely".

2. Systém pro automatický přepis zpravodajství

2.1. Blokové schéma

Na obrázku 2.1 je graficky znázorněno blokové schéma navrženého systému pro plně automatický přepis zpráv. Při splnění několika podmínek je tento systém připraven i na režim práce v reálném čase (reálným časem je myšleno zpoždění v řádu jednotek sekund). Konkrétně se jedná o modul pro předzpracování signálu, který zatím pracuje v off-line režimu, čímž ovlivňuje funkčnost i ostatních komponent systému.



Obr. 2.1: *Blokové schéma systému pro přepis zpráv s vyznačeným vývojem v čase. Bloky, jež jsou tématem této disertační práce, jsou tučně zvýrazněny.*

Pomineme-li úvodní krok, kterým je nahrání a digitalizace požadovaného audiozáznamu, je prvním funkčním blokem předzpracování signálu. V něm je převáděn kompaktní 16 kHz audiosignál na vektor příznaků (prvních 13 MFCC, delta, delta-delta, 10/25 ms). Takto reprezentovaný signál je dále přiveden do bloku provádějícího detekci změn akustického pozadí a mluvčích, který tento kontinuální vektor příznaků rozčlení na jednotlivé segmenty. Na vytvořených, akusticky homogenních segmentech je aplikována

operace odečítání kepstrálního průměru (CMS), kterou používáme pro potlačení vlivů přenosové cesty. Vzniklé segmenty (nyní již v podstatě můžeme hovořit o větách) jsou vyhodnocovány v bloku pro detekci řečových částí. Zde jsou z dalšího zpracování odděleny ty segmenty, které nenesou žádnou řečovou informaci. V tuto chvíli je u zbylých řečových segmentů určen mluvčí a jeho pohlaví. Tato informace neslouží pouze k prostému označení segmentů jménem a pohlavím mluvčího (například pro účely indexace záznamu), ale v dalším průběhu přepisu určuje způsob adaptace řečového modelu pro každý jednotlivý segment. Tato adaptace probíhá právě na základě jména, pohlaví mluvčího nebo i skupiny mluvčích. S takto zvolenými modely pro automatické rozpoznávání řeči je proveden přepis segmentů do textové podoby. Textový výstup je nakonec finálně upraven přidáním teček, čárek a velkých písmen ve vzniklých větách.

2.1.1. Parametrizace

Televizní nebo rozhlasové zpravodajství je do počítače zaznamenáno standardní televizní, případně zvukovou kartou se vzorkovací frekvencí 16 kHz a 16 bitovým kódováním. Parametrizace, dnes používaná pro účely zpracování a rozpoznávání řeči, nejčastěji pracuje s příznaky získanými z kepstra signálu [HUA01] nebo [PSU95]. Výhodou kepstrální analýzy je ten fakt, že lze od sebe oddělit složky signálu vytvořeného konvolucí právě těchto složek. U signálu řeči se jedná o kombinaci buzení hlasového traktu a impulsní odezvy hlasového ústrojí.

Parametrizace je prováděna 100 krát za sekundu s délkou okna 25 ms. Každý frame[1] je reprezentován vektorem 39 MFCC příznaků a logaritmem energie. Odečítání kepstrálního průměru je aplikováno jako poslední operace již na rozdělené segmenty. Důležitým aspektem výpočtu MFCC příznaků je jejich univerzálnost. Veškeré moduly systému pro automatický přepis zpráv totiž používají totožné příznaky. To znamená, že příznaky jsou počítány pouze jednou a jednotlivé moduly si je podle potřeby pouze načítají.

2.1.2. Detekce změn v audiosignálu

Cílem tohoto kroku je rozčlenění audiozáznamu na menší, více či méně homogenní části s ohledem na jejich akustické charakteristiky. Konkrétně se snažíme nalézt takové změny v charakteristice signálu, které reprezentují změnu mluvčího, změnu akustického pozadí z řeči na hudbu, z řeči na hluk a naopak. Dále to mohou být i změny v kvalitě přenosové cesty signálu, například přechod od čtené studiové řeči k telefonnímu rozhovoru.

[1] V české literatuře je kromě termínu „frame" používán i ekvivalentní výraz „rámec"

Jedna z posledních verzí tohoto modulu používá ke zjišťování výše uvedených přechodů metodu s adaptivním oknem [ŽĎÁ05]. Celý algoritmus funguje přibližně následujícím způsobem: je inicializováno okno o velikosti B a je umístěno na počátek signálu. V daném okně je nalezen takový bod $\hat{t}$, jenž bude maximalizovat zisk $G(\hat{t}|a,b)$. Tento bod je označen jako *kandidát na bod změny*. Poté mohou nastat dva různé případy. Kandidát není potvrzen a analyzující okno je rozšířeno. Je-li kandidát $\hat{t}$ potvrzen, je velikost okna zmenšena na $[a, \hat{t}]$ a v tomto novém okně je hledán další kandidát. Bude-li tento kandidát potvrzen, okno bude opět zmenšeno. V opačném případě byl nalezen bod změny, jenž byl z hlediska původního okna nejvíce vlevo, analyzující okno je posunuto do tohoto bodu a jeho inicializační velikost je opět nastavena na B.

Detekce změn v audiosignálu má velký dopad na finální výstup celého rozpoznávacího řetězce. Pokud je tento krok proveden správně, tak se výrazně zvyšuje pravděpodobnost správného rozpoznání řečových segmentů a je možné vyřadit rozpoznávání později označených neřečových segmentů z dalšího zpracování. Tím je samozřejmě dosaženo podstatné úspory času. Na druhou stranu je třeba počítat s tím, že při této segmentaci vzniká určité procento chyb, které negativně ovlivňují rozpoznávání řeči. Je-li změna mluvčího vyznačena uprostřed věty nebo slova, je takto vzniklý segment příčinou zhoršeného rozpoznávacího skóre výsledného textového přepisu.

2.1.3. Nalezení řečových segmentů

Jak již bylo zmíněno v předchozím odstavci, je odstranění neřečových segmentů důležité nejen z hlediska úspěšnosti textového přepisu, ale má podstatný dopad i na rychlost celkového zpracování audionahrávky. Detekce řeč/neřeč je prováděna pravděpodobnostním klasifikátorem, který se pokouší zařadit daný segment do jedné z pěti širších kategorií. V případě řeči to jsou „čistá řeč", „řeč + hudba na pozadí" a „řeč + hluk na pozadí ". Na druhou stranu mohou být neřečové segmenty označeny jako „hudba" nebo jako „hluk". Podobné systémy ve světě používají některé techniky známé z oblasti rozpoznávání řeči a mluvčích, jako například [HAR01]. Jsou to především klasifikátory založené na metodách gaussovských modelů směsí, nejbližšího souseda (Nearest Neighborhood – NN), umělých neuronových sítí (Artificial Neural Networks – ANN) a skrytých markovských modelů. Systém prezentovaný v této práci používá poslední ze zmíněných variant.

2.1.4. Identifikace mluvčího

V tomto kroku je segment, označený jako řečový, zpracován modulem pro identifikaci mluvčího. Ten je postaven na databázi mluvčích, jež byla získána od nejfrekventovanějších osob z trénovacích nahrávek (dále již jen klíčoví mluvčí). Jedná se o

osoby, které se v těchto nahrávkách objevily alespoň po dobu 100 sekund, tudíž pro ně existuje dostatečný objem trénovacích dat. To je nutná podmínka jak pro rozpoznávání řečníka, tak i pro modul adaptující řečové modely na tyto mluvčí.

Cílem identifikace je určení identity mluvčího. Náš systém toho dosahuje výběrem nejpravděpodobnějšího mluvčího ze seznamu kandidátů použitím metod GMM [CAM97]. Ve fázi rozpoznávání je tedy pro vstupní segment spočítána pravděpodobnost, s jakou by byl tento segment generován daným mluvčím. Ten, kdo má se svým modelem nejvyšší pravděpodobnost, je prohlášen za mluvčího náležícího k dané promluvě [DUN00]. Kromě této informace je dalším blokům předávána informace o pohlaví tohoto kandidáta spolu se seznamem nebližších dalších mluvčích přicházejících v úvahu. Všechny tyto informace pak pomáhají zvolit optimální řečový model v modulu pro adaptaci (volbu) řečového modelu pro konkrétní segment.

2.1.5. Verifikace mluvčího

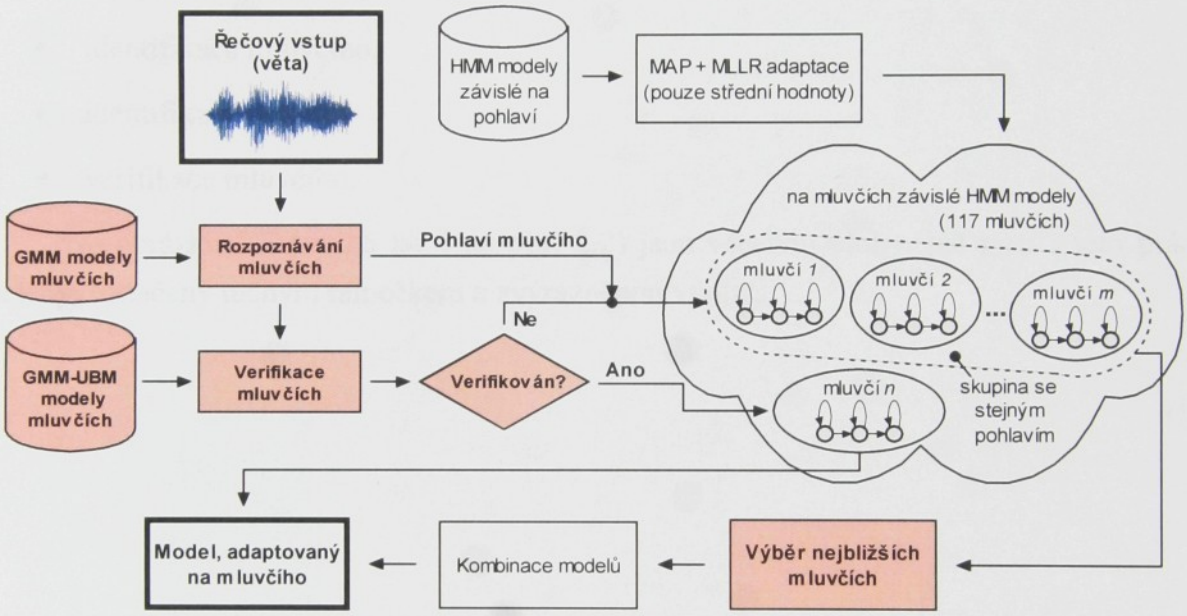
Na tomto místě je ověřováno, zda je navržený kandidát z procesu identifikace mluvčího opravdu tím mluvčím, kterému náleží ověřovaný segment. V podstatě se jedná o ověření dvou hypotéz: mluvčí je skutečně oním člověkem, který vyslovil testovanou promluvu, a nebo jím není. První hypotéza je reprezentována modelem mluvčího, k modelování druhé hypotézy slouží univerzální model (UBM) [REY00]. Ve fázi rozpoznávání zjišťujeme pravděpodobnost obou hypotéz a na základě přednastaveného prahu sloužícího k nastavení míry verifikační jistoty je potvrzena nebo zamítnuta identita mluvčího.

2.1.6. Volba optimálních řečových modelů

Volbu řečového modelu můžeme provést dvěma možnými způsoby. V dosud používaném postupu je buď zvolen dříve vytvořený model odpovídající verifikovanému mluvčímu z předchozího modulu, nebo (v případě zamítnutí ve verifikačním modulu) je zvolen na pohlaví závislý model. V případě úspěšné verifikace se druhý postup od prvního neliší. Je zvolen na mluvčího adaptovaný řečový model [LEG95]. Změna nastává v případě, že identita mluvčího je verifikačním procesem zamítnuta. Tady, na základě pohlaví nejpravděpodobnějšího mluvčího určeného při identifikaci a na základě skupiny dalších pravděpodobných kandidátů stejného pohlaví, je vytvořen adaptovaný model odpovídající této akusticky podobné skupině. Kroky vedoucí k volbě optimálních řečových modelů jsou znázorněny na obrázku 2.2.

2.1.7. Rozpoznávání spojité řeči

Při vývoji rozpoznávače spojité řeči pro český jazyk je nejdůležitějším úkolem umět efektivně pracovat s velkým slovníkem. Současný systém umožňuje používat až 400 tisíc slovníkových položek společně s dalšími 100 tisíci výslovnostními variantami. Základními jednotkami jsou fonémy, jichž je celkem 48, z toho 7 z nich je použito na modelování různých hluků. Tyto fonémy jsou modelovány třístavovými levo-pravými HMM s vícemixturovou reprezentací rozložení (až do 100 mixtur na jeden stav). Dekodér pracuje na principu časově synchronního, jednoprůchodového Viterbiho prohledávání [NOU05a].



Obr. 2.2: *Blokové schéma znázorňující postup výběru optimálního řečového modelu. Zvýrazněné bloky jsou tématem této práce.*

Úspěšnost rozpoznávání se v plně automatickém režimu pohybuje kolem 80 % správně rozpoznaných slov (v závislosti na typu pořadu). Rozpoznávání je podporováno složitým akustickým a jazykovým modelem.

2.1.8. Zpracování textového výstupu

Výstupem modulu pro rozpoznávání spojité řeči je text psaný malými písmeny (s určitými výjimkami). Na tomto místě je rozpoznaný text použitím jednoduchého schématu převeden do formy s velkými počátečními písmeny ve větách. To samé schéma je použito u potenciálních vlastních jmen. Nakonec jsou na místech, kde je to nejvíce pravděpodobné, věty doplněny o systém rozdělovacích znamének (čárky a tečky ve větách).

2.2. Bloky řešené v této práci

Na tomto místě je důležité upozornit, že se tato disertační práce bude zabývat pouze některými bloky uvedenými v předchozích kapitolách. Jak již bylo řečeno v úvodu, úloha přepisu zpravodajství je natolik komplexní, že si vyžaduje tým výzkumníků specializujících se na konkrétní témata.

Disertační práce bude zaměřena na následující úlohy:

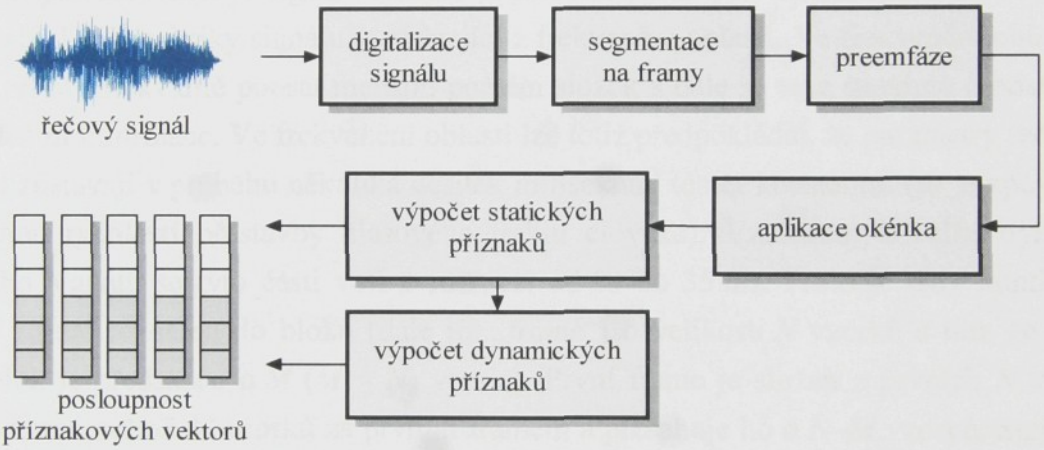
- parametrizace signálu,
- nalezení řečových segmentů,
- identifikace mluvčího,
- identifikace pohlaví,
- verifikace mluvčího.

Na obrázku 2.1 (stejně tak i na obr. 2.2) jsou všechny úlohy, jež jsou v této práci řešeny, označeny tučným rámečkem a zvýrazněnou výplní.

3. Předzpracování signálu

3.1. Cíl předzpracování

Hlavním úkolem parametrizace řečového signálu je snaha o odstranění nadbytečné (redundantní) informace, která je v řečovém signálu v poměrně hojné míře zastoupena. Parametrizaci provádíme tak, abychom získali příznaky, které budou co nejvěrněji reprezentovat originální signál. Po příznacích na druhou stranu požadujeme, aby parametrizovaný signál byl pokud možno co nejmenší. Samotná parametrizace signálu je prvním krokem v celém řetězci kroků, jež je nutné učinit pro realizaci rozpoznávání. Operace vedoucí k parametrizaci promluvy jsou schématicky znázorněny na obrázku 3.1 a v následujícím textu budou ještě podrobněji vysvětleny.



Obr. 3.1: *Základní schéma postupu parametrizace řečového signálu*

3.2. Digitalizace řečového signálu

Abychom byli schopni řečový signál zpracovávat v počítači, je nejprve nutné převést jej na jeho číslicovou reprezentaci, tj. provést digitalizaci. Kvalitu digitalizovaného signálu nejvíce ovlivňují dva základní parametry. Těmito parametry jsou *kvantizační krok* a *vzorkovací frekvence*. Abychom určili tyto parametry, je třeba nejprve analyzovat lidskou řeč. Obecně platí, že čím vyšší je použitá vzorkovací frekvence a čím větší je rozlišení převodu (menší kvantizační krok => menší kvantizační chyba), tím jsou výsledky rozpoznávání lepší. V systémech pro identifikaci nebo verifikaci mluvčích jsme ale většinou limitováni parametry přenosové cesty (rozpoznávaný signál je většinou přenášen po telefonní lince), a proto nemá smysl používat vzorkovací frekvenci 44 kHz, když přicházející signál je degradován přenosem po telefonní lince a má vzorkovací frekvenci pouze 8 kHz, nebo u televizních zpráv 16 kHz.

Frekvence lidského hlasu závisí na konfiguraci hlasového traktu (jakou hlásku mluvčí právě vyslovuje) a na konkrétním mluvčím (věk, pohlaví, psychické rozpoložení atd.).Ve většině případů frekvence lidského hlasu nepřesahuje 3 kHz. Na základě experimentů s lidským hlasem bylo dokázáno, že naprostá většina informací v lidské řeči je obsažena ve frekvencích řádově stovek Hz u znělých hlásek a maximálně několika kHz u hlásek neznělých. Proto je pro účely rozpoznávání naprosto postačující vzorkovací frekvence 16 kHz, ale je možné úspěšně pracovat i s frekvencí 8 kHz. Abychom byli schopni pokrýt poměrně velkou dynamiku lidského hlasu (hodnota kolem 60 dB), měli bychom použít minimálně 12 bitové rozlišení. Z praktických důvodů se používá rozlišení 16 bitů, v případě telefonního signálu pak 8 bitů s nelineárním kvantizačním krokem.

3.2.1. Segmentace signálu

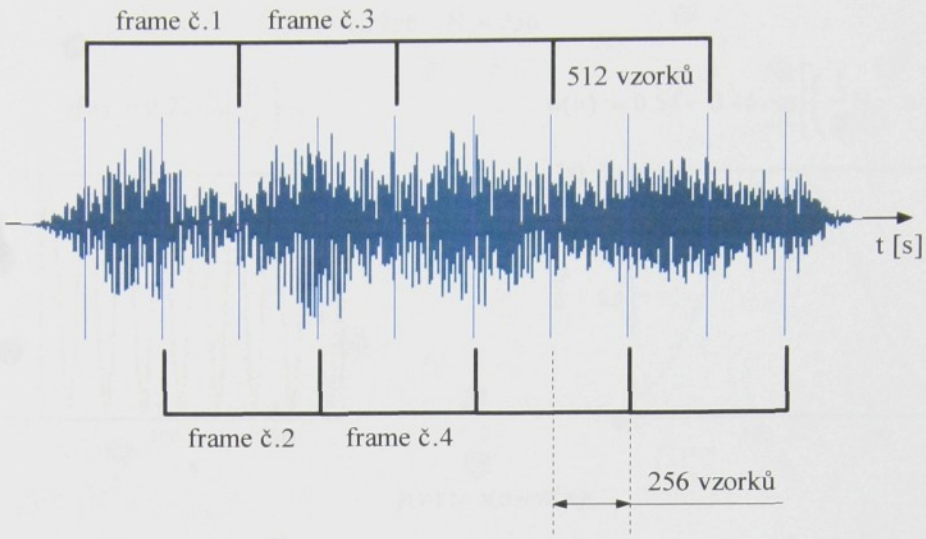
Při parametrizaci je signál většinou popisován v jiné než časové oblasti. Obvykle se využívají charakteristiky signálu vycházející z frekvenční oblasti. Ve frekvenční oblasti lze signál při stejné kvalitě popsat menším počtem složek a dále se také snažíme o odstranění redundantní informace. Ve frekvenční oblasti lze totiž předpokládat, že parametry řečového signálu zůstávají v průběhu několika desítek milisekund téměř konstantní (to je způsobeno omezenou rychlostí přestavby hlasového traktu člověka). Vzhledem k velké dynamice řečového signálu se tyto části volí v rozmezí od 10 do 35 ms. Proto je tedy kontinuální řečový signál rozdělen do bloků (dále již „*frame*“) o velikosti N vzorků s tím, že každý další blok je předsazen o M ($M < N$) vzorků. První frame je složen z prvních N vzorků. Druhý frame začíná M vzorků za prvním framem a přesahuje ho o $N–M$ vzorků. Stejně tak třetí frame začíná $2M$ vzorků za prvním (nebo M vzorků za druhým) a přesahuje jej o N-$2M$ vzorků, viz obr. 3.2. Typicky je velikost N rovna 512 (to je 32 ms při 16 kHz vzorkovací frekvenci a navíc 512 je mocninou čísla 2 a s framem je tedy možné provádět rychlou variantu DFT, tedy FFT) a M je rovna 256. Tyto hodnoty byly použity i v systémech vyvíjených na TU v Liberci.

3.2.2. Preemfáze

Preemfáze je metoda sloužící ke zvýraznění vyšších kmitočtů v signálu, protože mají obvykle nižší úroveň. Většinou se realizuje pomocí jednoduchého číslicového filtru:

$$y(n) = x(n) - a \cdot x(n-1), \tag{3.1}$$

kde $y(n)$ je n-tý vzorek signálu po preemfázi a $x(n)$ je n-tý vzorek původního signálu. Konstanta a se běžně volí v rozsahu 0,95–0,98. Použití preemfáze není nezbytné.



Obr. 3.2: *Segmentace řečového signálu*

3.2.3. Aplikace okénka

Rozdělíme-li signál na framy, vzniká na jejich okrajích náhlý přechod (uříznutí), který je nežádoucí při dalším zpracování framů (frekvenční analýza). To, že odstraníme nespojitosti na začátku a konci framu, má za důsledek vyhlazení průběhu spektra. Aplikací okénkovací funkce konkrétně provádíme potlačení váhy vzorků na začátku a konci framu a to tak, že jednotlivé vzorky framu násobíme *váhovou* nebo též *okénkovací funkcí* [PRO96]. Okénkovacích funkcí je celá řada, ale v oblasti zpracování řeči dominuje jediná a tou je Hammingova okénkovací funkce. Pokud si označíme funkci pro Hammingovo okénko jako $w(n)$, $0 \leq n \leq N-1$, kde N je opět počet vzorků v každém framu, tak funkce pro vážení jednotlivých framů bude mít tvar:

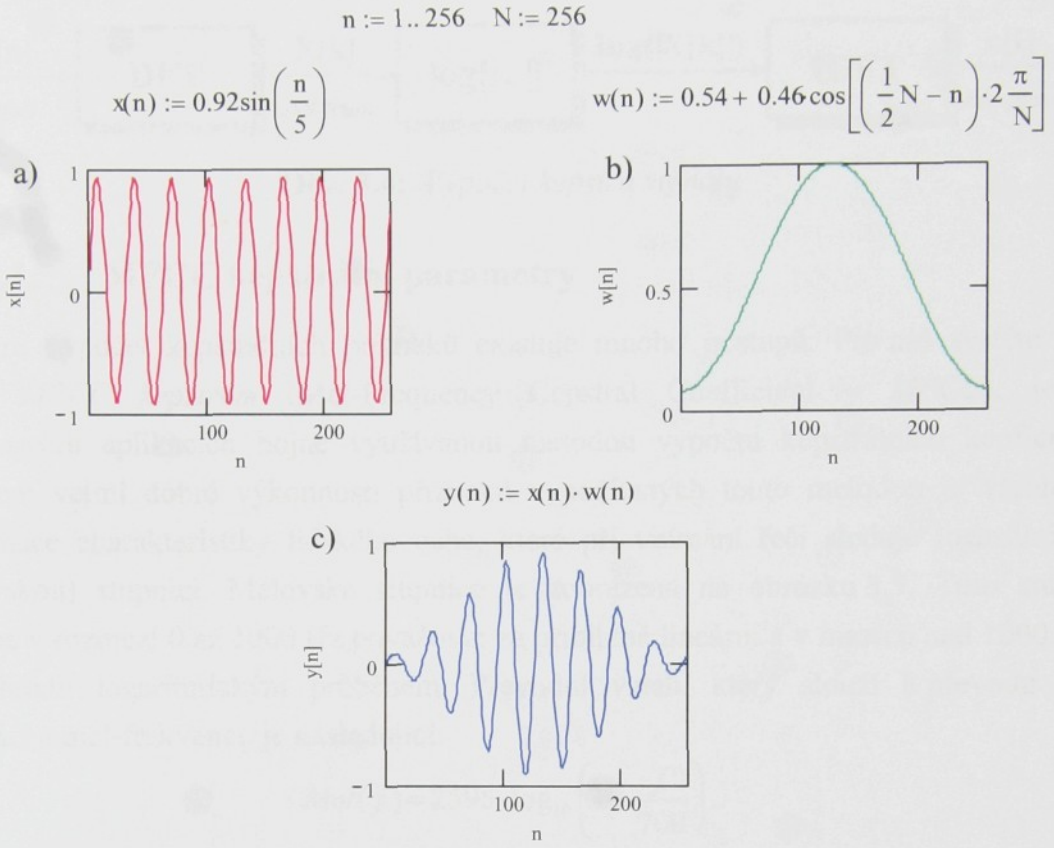
$$w(n) = 0,54 + 0,46 \cdot \cos\left[\left(\frac{1}{2} \cdot N - n\right)\frac{2 \cdot \pi}{N}\right]. \tag{3.2}$$

Použijeme-li tedy okénkovací funkci uvedenou ve vztahu (3.2), n-tý vzorek signálu vypočteme takto:

$$y(n) = x(n) \cdot w(n), \tag{3.3}$$

kde $x(n)$ je původní a $y(n)$ nově získaný vzorek signálu.

Na obrázku 3.3a je vykreslen jeden frame signálu před aplikací Hammingova okénka, na obr. 3.3b je samotná váhová funkce a na obr. 3.3c je frame již po vynásobení váhovou funkcí.



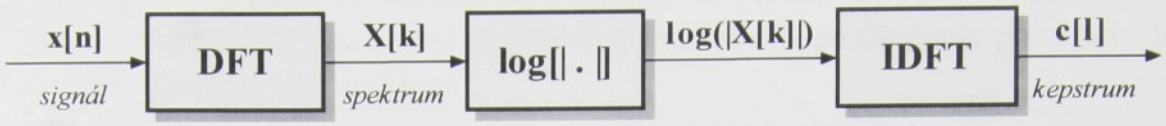
Obr. 3.3: *Signál jednoho framu před a po aplikaci Hammingova okénka*

3.3. Kepstrální parametry

V mnoha studiích bylo prokázáno, že nejvhodnějšími nízkoúrovňovými příznaky pro aplikace automatického rozpoznávání mluvčího jsou kepstrální koeficienty [REY94b]. Dále bylo prokázáno, že kepstrální příznaky jsou velmi dobře použitelné v textově závislé i nezávislé identifikaci, pokud byly nahrávky pořízeny za relativně vhodných podmínek (dobrý odstup signálu od šumu, stacionární podmínky při nahrávání).

Kepstrální analýzu používáme k oddělování složek signálu, který vznikl konvolucí několika složek (řečový signál je výsledkem konvoluce buzení hlasového traktu s impulzní odezvou hlasového ústrojí). Obecně se při výpočtu kepstra postupuje podle vztahu (3.4), kde kepstrum $c[l]$ vypočítané ze vstupní posloupnosti vzorků signálu $x[n]$ je definováno jako inverzní Fourierova transformace logaritmu absolutní hodnoty spektra signálu. Tento postup je naznačen i v blokové verzi, viz obrázek 3.4.

$$c[l] = DFT^{-1}\left\{\log\left(\left|DFT\left\{x[n]\right\}\right|\right)\right\} \tag{3.4}$$

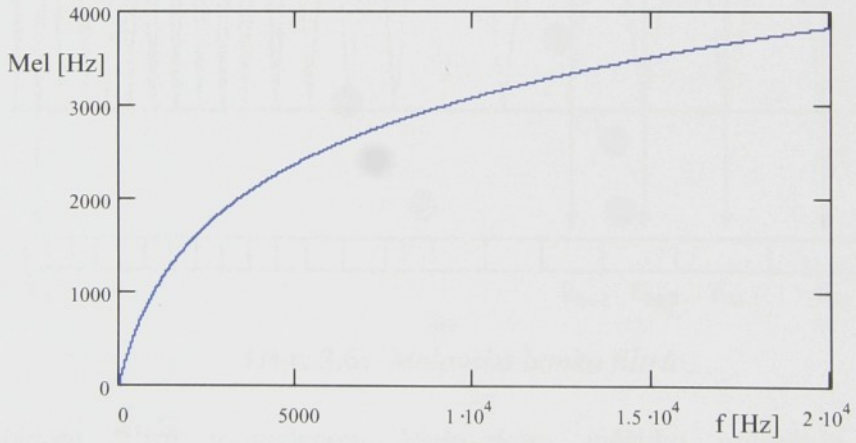


Obr. 3.4: *Výpočet kepstra signálu*

3.3.1. MFCC kepstrální parametry

Pro výpočet kepstrálních příznaků existuje mnoho postupů. Pro náš systém jsme zvolili *MFCC kepstrum* (Mel-Frequency Cepstral Coefficient – *MFCC*), jež je v současných aplikacích hojně využívanou metodou výpočtu kepstrálních koeficientů. Důvodem velmi dobré výkonnosti příznaků vypočítaných touto metodou je věrohodná aproximace charakteristiky lidského ucha, které při vnímání řeči sleduje logaritmickou (Melovskou) stupnici. Melovská stupnice je zobrazena na obrázku 3.5. Tuto stupnici můžeme v rozmezí 0 až 1000 Hz považovat za přibližně lineární a v mezích nad 1000 Hz ji lze nahradit logaritmickým průběhem. Převodní vztah, který slouží k převodu mezi frekvencí a mel-frekvencí, je následující:

$$Mel(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right). \tag{3.5}$$



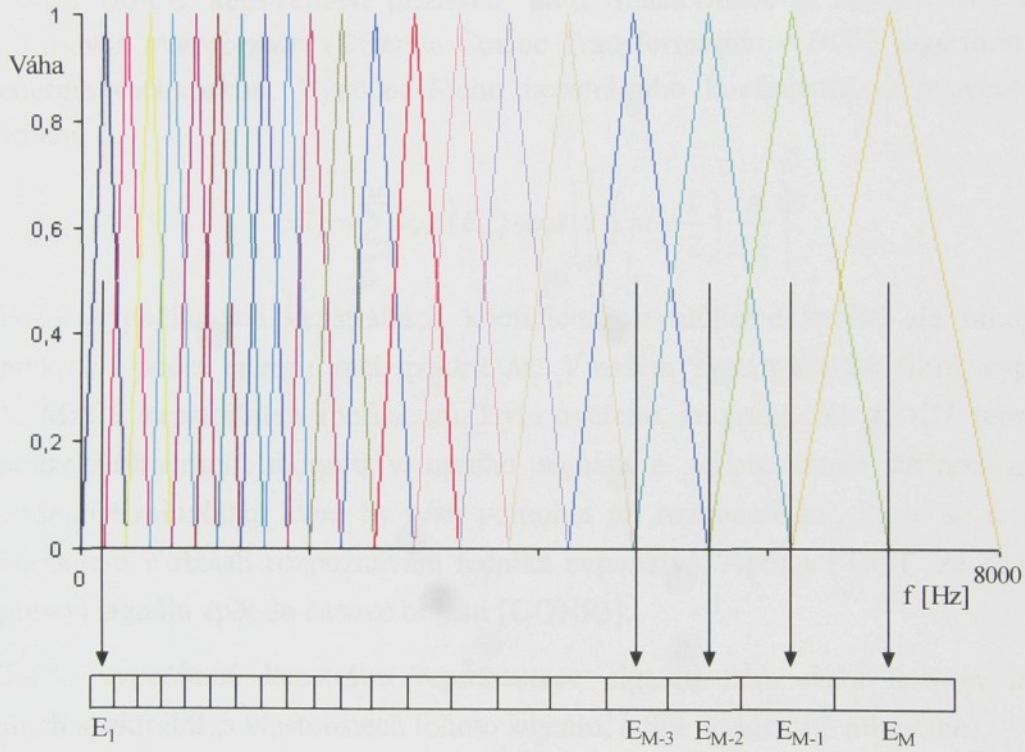
Obr. 3.5: *Graf převodní funkce mezi mel-frekvencí a frekvencí*

V praxi provádíme výpočet MFCC kepstrálních příznaků následujícím způsobem:

- Máme-li již digitalizovaný signál (viz předchozí kapitoly), převedeme jej vhodným způsobem do frekvenční oblasti. Nejvhodnějším způsobem výpočtu amplitudového spektra signálu je algoritmus *rychlé diskrétní Fourierovy transformace* (Discrete Fourier Transformation – *DFT*). Jak již bylo dříve uvedeno, spektrum signálu počítáme z jednotlivých framů. Nyní můžeme využít toho, že velikost framu je 256 (nebo 512) vzorků (mocnina dvou) a použít rychlou (klasickou) variantu FFT algoritmu. Více informací o algoritmu FFT je možné nalézt v [PSU95]. Výstupem

FFT je stejný počet vzorků jako na vstupu, tj. 256 (512), ale využitím souměrnosti spektra okolo $F_s/2$ bylo možné po provedení normalizace použít pouze první polovinu vzorků (jednostranné spektrum).

- Dalším krokem ve výpočtu kepstra je převod amplitudového spektra na mel-spektrum [SCH00]. Počet koeficientů mel-spektra se obvykle volí okolo 20. Jeden koeficient mel-spektra obdržíme tak, že všechny hodnoty spektra filtrujeme (vážíme) bankou M trojúhelníkových filtrů (průběhy hodnot jednotlivých filtrů jsou uvedeny na obr. 3.6). Zjednodušeně jde o 20 pásmových propustí. Pro jednotlivé filtry všechny vážené vzorky sečteme a dostaneme tak energie signálu v každém pásmu filtru.



Obr. 3.6: *Melovská banka filtrů*

Uspořádání filtrů v uvedeném Melovském měřítku ovlivňuje způsob jejich rozestavení na frekvenční ose. Rozestupy mezi *centrálními frekvencemi* Δ_m a šířky pásem filtrů jsou do frekvence 1 kHz konstantní, počet filtrů v této lineární oblasti volíme poloviční proti počtu všech filtrů v celém frekvenčním pásmu (v našem případě je to 10 filtrů). Směrem k vyšším frekvencím se rozestupy mezi jednotlivými filtry zvětšují, stejně tak se zvětšují i šířky pásem (toto je opět ilustrováno na obr. 3.6). Rozestupy se zvětšují tak, že $\Delta_m = 1{,}2 \cdot \Delta_{m-1}$ a centrální frekvence $b_m = b_{m-1} + \Delta_m$. Výpočet energie spektra v m-tém pásmu je tedy

$$E_m = \sum_{k=b_m-\Delta_m}^{b_m+\Delta_m} X(k) \cdot U_{\Delta m}(k+b_m), \tag{3.6}$$

kde $X(k)$ je k-tý vzorek spektra signálu framu získaný pomocí FFT a $U_{\Delta m}(k)$ je trojúhelníková váhová funkce získaná výpočtem ze vztahu

$$U_{\Delta m}(k) = \begin{cases} |k| < \Delta_m \rightarrow 1 - \dfrac{|k|}{\Delta_m} \\ |k| \geq \Delta_m \rightarrow 0 \end{cases}. \tag{3.7}$$

- Výsledek ve formě vektoru energie $\mathbf{E} = (E_1, E_2, E_3, \ldots, E_M)$ je dále transformován ve vektor MFCC kepstrálních příznaků. Tato transformace je uskutečněna *diskrétní kosinovou transformací* (Discrete Cosine Transformation – *DCT*) logaritmu vektoru energie mel-spektra. Výpočet l-tého kepstrálního koeficientu je proveden podle vztahu

$$c(l) = \sum_{m=1}^{M} \log(E_m) \cdot \cos\left[l \cdot \left(m - \frac{1}{2} \right) \cdot \frac{\pi}{M} \right]. \tag{3.8}$$

Počet vypočítaných kepstrálních koeficientů si můžeme zvolit, ale nikdy nesmí překročit počet energií mel-spektra M. V našem systému z 24 filtrů vypočítáme 13 MFCC kepstrálních koeficientů. Bylo ověřeno, že první člen z DCT reprezentuje pouze průměrnou hodnotu vstupního signálu a nenese téměř žádnou informaci o identitě mluvčího, která by nám pomohla při rozpoznávání. Proto se tento první koeficient v oblasti rozpoznávání řečníka nepoužívá. Aplikaci DCT lze chápat jako převod signálu zpět do časové oblasti [GON93].

Takto vypočítaná kepstrální reprezentace signálu nám dává dobrou informaci o lokálních spektrálních vlastnostech tohoto signálu, a tím i o identitě mluvčího.

3.3.2. Výpočet dynamických příznaků

Dynamickými příznaky rozumíme derivace kepstrálních příznaků[1] (nazvěme tyto příznaky statické) v čase, tzn. že dynamické příznaky vyjadřují změnu původních příznaků v čase. V reálných systémech pro rozpoznávání mluvčího jsou společně se statickými příznaky používány dynamické příznaky prvního (tzv. *delta příznaky*) nebo též i druhého řádu (tzv. *delta-delta příznaky* nebo také *akcelerační příznaky*). V systémech pro rozpoznávání řeči platí, že některé dynamické příznaky mohou mít vyšší vypovídající schopnost než jejich statické varianty, viz např. [NOU95].

[1] Obecně to samozřejmě mohou být i jakékoliv jiné příznaky.

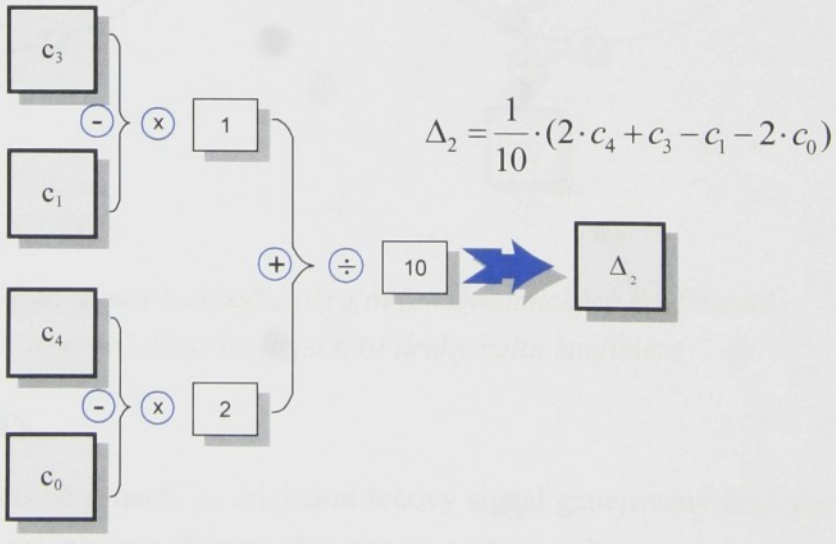
Dynamické příznaky počítáme pomocí numerické derivace průběhu statických příznaků v čase. Jelikož se pro výpočet dynamických příznaků používají různé konfigurace vstupních statických příznaků (od prosté diference až po složité vícebodové metody), bude v této kapitole uvedena obecná metoda, jak tyto dynamické příznaky vypočítat. V následujícím textu má c_n význam vstupu, kde $n=0, 1, ..., N-1$ a Δ_n je výstupem, tj. derivací tohoto vstupu

$$\Delta_n = \frac{\sum_{i=1}^{W} i \cdot [c_{n+i} - c_{n-i}]}{2 \cdot \sum_{i=1}^{W} i^2}. \tag{3.9}$$

Ze vztahu (3.9) je možné vypočítat libovolné derivace vstupu určením požadované délky okna W (udáme jeho poloměr). Poloměrem okna rozumíme počet koeficientů na každou stranu od pozice aktuálně vypočítávaného příznaku. Položíme-li W rovno jedné, dostaneme vztah pro jednoduchou nekauzální diferenci

$$\Delta_n = \frac{c_{n+1} - c_{n-1}}{2}. \tag{3.10}$$

Tento vztah lze také použít, ale pokud chceme obdržet vyhlazenější průběh derivace, je lepší za W zvolit číslo větší než jedna. Vztah (3.9) je graficky znázorněn na obr. 3.7. Pokud chceme vypočítat akcelerační příznaky, bude postup naprosto totožný, pouze s tím rozdílem, že jako vstup budeme brát delta příznaky a výstupem budou delta-delta příznaky.



Obr. 3.7: *Výpočet koeficientu Δ_2 dvoubodovou metodou*

Při praktické realizaci algoritmu pro výpočet dynamických příznaků je třeba dát pozor na okrajové podmínky (především počítáme-li diference z více bodů), které nastávají

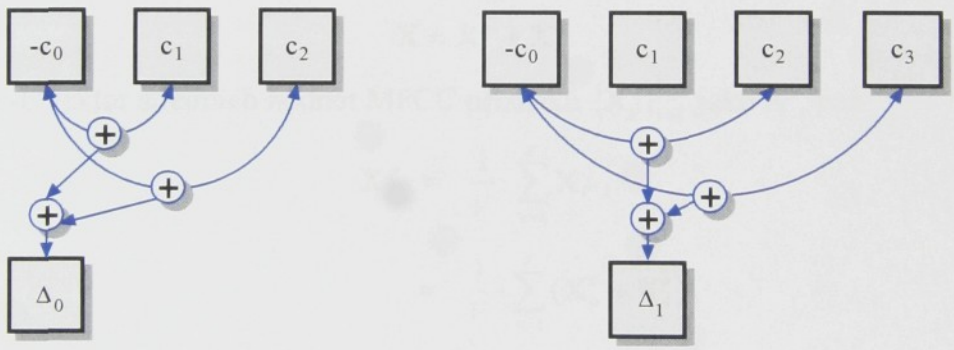
na okrajích zpracovávaného signálu. Tyto okrajové podmínky je nutné rozumným způsobem ošetřit. Jeden ze způsobů, kterým můžeme okrajové podmínky ošetřit, je pro začátek signálu uveden vztahem

$$\Delta_n = \frac{\sum_{i=1}^{W} i \cdot [c_{n+i} - c_0]}{2 \cdot \sum_{i=1}^{W} i^2}, \qquad n < W . \tag{3.11}$$

Pro opačnou stranu signálu platí vztah (3.12), který je velmi podobný předchozímu vztahu (3.11), jen s tím rozdílem, že je zaměněno pořadí okrajových podmínek. N je počet vstupních vektorů a W je délka okna

$$\Delta_n = \frac{\sum_{i=1}^{W} i \cdot [c_{N-1} - c_{n-i}]}{2 \cdot \sum_{i=1}^{W} i^2}, \qquad n \geq N - W . \tag{3.12}$$

Pracujeme-li s velikostí okna $W=2$ (tak jako v našich systémech), příklad výpočtu prvního a druhého příznaku je na obr. 3.8. Na obrázku je naznačen pouze způsob, jakým jsou k sobě vstupní příznaky přidruženy.



Obr. 3.8: *Způsob výpočtu okrajových dynamických koeficientů a) první delta koeficient, b) druhý delta koeficient*

3.3.3. CMS

V naprosté většině případů je originální řečový signál generovaný lidským hlasovým ústrojím nějakým způsobem ovlivněn přenosovou cestou, což se negativně projevuje na úspěšnosti rozpoznávání. Předpokládejme, že řečový signál *a* je filtrován kanálem[1] *b*:

[1] Příkladem takové filtrace může být například telefonní kanál, nebo „jen" vzdálenost mezi ústy řečníka a mikrofonem, případně různý nahrávací hardware.

$$z = a * b, \tag{3.13}$$

kde symbol * označuje operaci konvoluce. Ekvivalentně k tomu je ve spektrální oblasti možné provést zápis tohoto výrazu

$$Z = A \cdot B, \tag{3.14}$$

kde Z, A a B jsou spektra z, a a b. Zlogaritmováním výrazu (3.14) získáme vztah

$$\log(Z) = \log(A) + \log(B). \tag{3.15}$$

V kepstrální oblasti je tudíž kanál se signálem superponován. Protože vektor energie $\boldsymbol{E}$ uvedený ve vztahu (3.6) reprezentuje (Melovskou bankou filtrů) vážené spektrum, ekvivalentně k (3.14) můžeme zapsat

$$\mathbf{E} = [E_m]_{m=1}^{M} = [E_m^a \cdot E_m^b]_{m=1}^{M}, \tag{3.16}$$

kde vektory $\mathbf{E}^a$ a $\mathbf{E}^b$ reprezentují energie vážených spekter a a b, M je počet Melovských bank filtrů. Zlogaritmováním vztahu získáme

$$\mathbf{E}_{\log} = [\log(E_m)]_{m=1}^{M} = [\log(E_m^a) + \log(E_m^b)]_{m=1}^{M}. \tag{3.17}$$

Aplikací DCT na $\mathbf{E}_{log}$ bylo odvozeno [FUR81], že efektem přenosového kanálu je přidání aditivní složky k vektoru příznaků MFCC

$$\mathbf{X} = \mathbf{X}^a + \mathbf{X}^b. \tag{3.18}$$

Označíme-li vektor středních hodnot MFCC příznaků $\{\mathbf{X}_v\}_{v=1}^{V}$ jako $\mathbf{X}^\mu$, pak

$$\mathbf{X}^\mu = \frac{1}{V} \cdot \sum_{v=1}^{V} \mathbf{X}_V, \tag{3.19}$$

$$= \frac{1}{V} \cdot \sum_{v=1}^{V} (\mathbf{X}_v^a + \mathbf{X}_v^b), \tag{3.20}$$

$$= \frac{1}{V} \cdot \sum_{v=1}^{V} \mathbf{X}_v^a + \frac{1}{V} \cdot \sum_{v=1}^{V} \mathbf{X}_v^b. \tag{3.21}$$

Budeme-li dále předpokládat, že charakteristiky přenosového kanálu jsou časově invariantní, můžeme vztah zapsat jako

$$\mathbf{X}^\mu = \frac{1}{V} \cdot \sum_{v=1}^{V} \mathbf{X}_v^a + X^b. \tag{3.22}$$

Pokud budeme dále předpokládat, že spektrální energie je po dobu trvání promluvy rovnoměrně rozložena přes celé spektrum, potom se bude vztah $1/V \cdot \sum \mathbf{X}_v^a$ blížit konstantě a $\mathbf{X}^b$ bude nalezeno pomocí formule (3.19) a my získáme vztah pro kompenzaci vlivu přenosového kanálu

$$\left\{\mathbf{X}_v^{COMP}\right\}_{v=1}^{V} = \left\{\mathbf{X}_v - \mathbf{X}^{\mu}\right\}_{v=1}^{V}. \tag{3.23}$$

Tato normalizace je v literatuře známa pod označením odečítání kepstrálního průměru (CMS) [FUR81], [BAL99]. Ze vztahu (3.22) vyplývá, že vektor středních hodnot současně reprezentuje i střední hodnotu řečového spektra. Ve většině řečových aplikací, kde je CMS používáno, však není k dispozici řečová reprezentace o dostatečné délce. Proto při odstranění střední hodnoty z MFCC dochází i k určitému poklesu rozpoznávacího skóre u kvalitních promluv. Na druhou stranu získáme cennou odolnost vůči rušivým podmínkám přenosové cesty.

4. Metody rozpoznávání mluvčího

4.1. Úvod

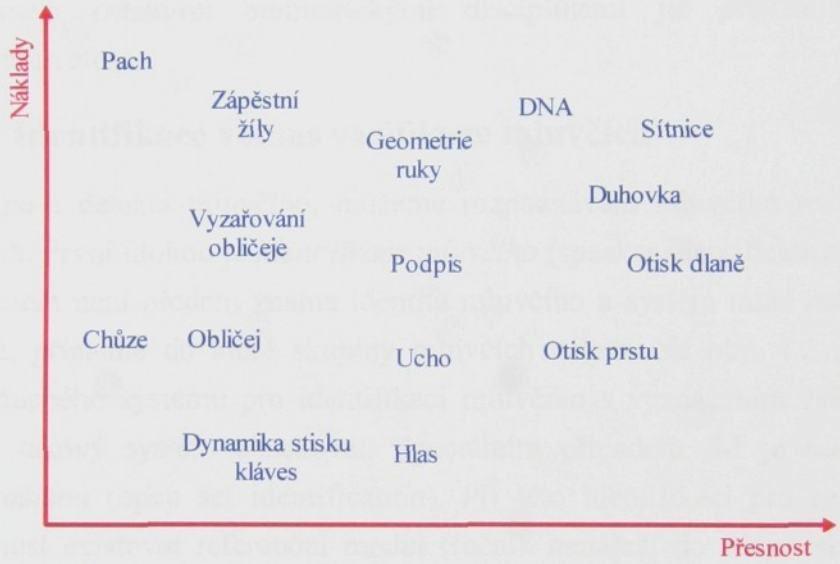
4.1.1. Biometrie

Biometrie ve své podstatě znamená „měření biologických faktorů člověka". Věda byla řadu let velmi zaneprázdněna výzkumem prostředků pro rozpoznávání osob pomocí měření a záznamu lidských fyziologických charakteristik. V dnešní době se k tomuto účelu používají i osobnostní rysy člověka. Přiřazení identity určitému člověku je obecně nazýváno rozpoznáváním osob. Problém určení identity osoby může být rozdělen na dvě kategorie, verifikace (nebo také autorizace) a identifikace, které jsou ve své podstatě elementárně odlišné. Verifikace je postup, kdy autentifikace osoby probíhá na základě jejího dříve zaznamenaného vzorku. V druhém případě se jedná o biometrický systém, který identifikuje osobu z celé evidované populace tím, že vyhledá odpovídající záznam v databázi.

Biometrická metoda	Univerzálnost	Osobitnost	Neměnnost	Měřitelnost	Výkonnost	Akceptovatelnost	Odolnost proti napodobení
DNA	V	V	V	N	V	N	N
Ucho	S	S	V	S	S	V	S
Obličej	V	N	S	V	N	V	V
Vyzařování obličeje	V	V	N	V	S	V	N
Otisk prstu	S	V	V	S	V	S	N
Chůze	S	N	N	V	N	V	S
Geometrie ruky	S	S	S	V	S	S	S
Zápěstní žíly	S	S	S	S	S	S	N
Duhovka	V	V	V	S	V	N	V
Dynamika stisku kláves	N	N	N	S	N	S	S
Pach	V	V	V	N	N	S	N
Otisk dlaně	S	V	V	S	V	S	S
Sítnice	V	V	S	N	V	N	V
Podpis	N	N	N	V	N	V	V
Hlas	S	N	N	S	N	V	V

Tab. 4.1: *Srovnání biometrických znaků, kdy „V" je zkratka pro vysokou vypovídací schopnost, „S" pro střední a „N" pro nízkou. Tabulka převzata z [JAI04].*

Důmyslné senzory jsou v dnešní době schopné identifikovat člověka podle tvaru ruky, chodidla nebo tvaru hlavy. Jsou schopny změřit a identifikovat lidské oko a topografii otisků našich prstů. Je zde rovněž typ naší krve, náš jedinečný kód DNA a spousta dalších více či méně použitelných biometrických charakteristik. Využití biometrie poskytuje některé jednoznačné výhody. Pouze biometrická autentifikace zakládá identifikaci na vlastní podstatě člověka jako živého tvora. Magnetické karty, klíče atd. se mohou ztratit, mohou být ukradeny, je možné vyrobit jejich duplikáty nebo je zapomenout doma. Hesla se často zapomínají.



Obr. 4.1: *Srovnání biometrických metod z hlediska ceny a přesnosti*

I když všechny biometrické systémy mají své vlastní výhody a nevýhody, existuje několik obecných vlastností, které je potřeba zajistit, aby bylo možné tyto systémy použít.

Které znaky se dají použít při biometrické identifikaci? Může to být jakákoliv fyziologická vlastnost (fyziologická biometrie) či v průběhu života získané osobnostní rysy chování (biometrie chování), jež splňují následující podmínky:

- univerzálnost – každý by tuto charakteristiku měl mít,
- jedinečnost – žádné dvě osoby by tento znak neměly mít společný,
- neměnnost – charakteristika by se neměla měnit v čase,
- měřitelnost – charakteristiku musíme být schopni kvantitativně změřit.

V praxi jsou ovšem nároky na biometrické znaky daleko vyšší. Tyto znaky by například měly splňovat podmínku dostatečné výkonnosti, s čímž přímo souvisí úspěšnost biometrického systému, systémová náročnost a také odolnost proti vnějším podmínkám, které také ovlivňují rozpoznávací úspěšnost systému. V neposlední řadě je to odolnost

biometrických znaků proti napodobení, osobitnost a akceptovatelnost. V současné době je například velmi těžké představit si situaci, kdy si někdo na letišti nechává dobrovolně odebrat tělní tekutiny k rozboru DNA pro potvrzení své totožnosti. Proto je třeba, aby sběr biometrických vzorků byl akceptován širokou veřejností.

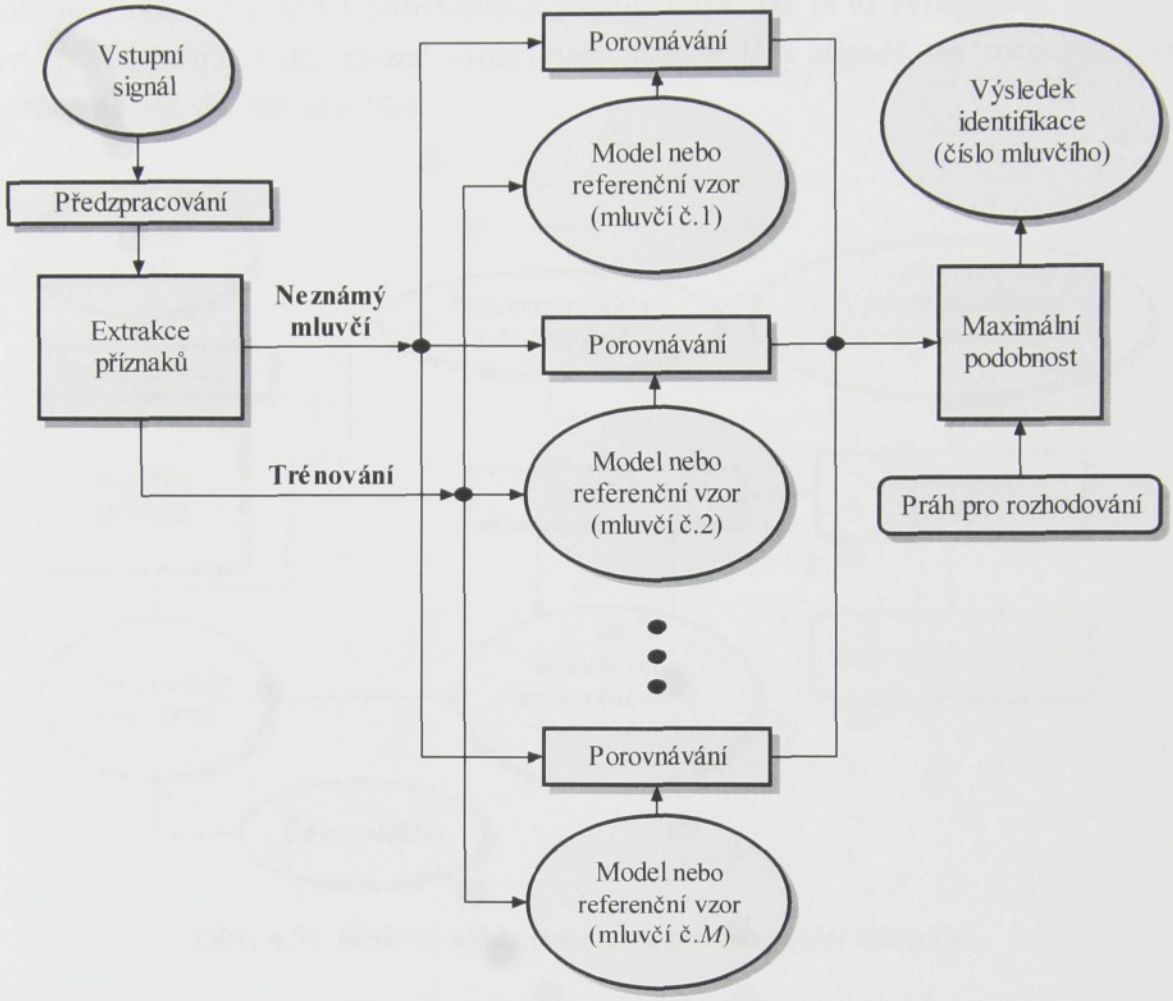
V tabulce 4.1 jsou souhrnně vypsány znaky používané při biometrické identifikaci, přičemž u každé metody jsou vypsány výše zmíněné vlastnosti ohodnocené písmeny „V" pro vysokou, „S" pro střední a „N" pro nízkou výkonnost biometrického znaku v dané oblasti. Po tomto krátkém úvodu upřesňujícím pozici metod identifikujících člověka podle jeho hlasu mezi ostatními biometrickými disciplínami již přejděme k vlastnímu rozpoznávání mluvčích.

4.1.2. Identifikace versus verifikace mluvčích

Pomineme-li detekci mluvčího, můžeme rozpoznávání mluvčího rozdělit do dvou základních úloh. První úlohou je *identifikace mluvčího* (speaker identification – IM), což je proces, při kterém není předem známa identita mluvčího a systém musí rozhodnout, kdo daná osoba je, případně do které skupiny mluvčích náleží. Na obr. 4.2 můžeme vidět příklad jednoduchého systému pro identifikaci mluvčího s vyznačením základních částí, které by měl takový systém obsahovat. Speciálním případem IM je *identifikace nad otevřenou množinou* (open set identification). Při této identifikaci pro rozpoznávaného mluvčího nemusí existovat referenční model (řečník nenáleží do žádné skupiny). To je většinou případ „soudních" aplikací. Zde by bylo vhodné, aby systém uměl rozpoznat neexistenci reference a vypsal, že testovaná promluva nepřísluší k žádnému referenčnímu modelu. Je tedy třeba nejdříve provést identifikaci a najít pro hlas neznámého člověka nejpravděpodobnějšího referenčního řečníka a pak rozhodnout, je-li zvolený referenční řečník skutečným autorem promluvy. Identifikaci v otevřené množině lze tedy chápat jako kombinaci identifikace a verifikace nad uzavřenou množinou. Druhou úlohou je *verifikace mluvčího* (speaker verification – VM).

Proces VM je používaný pro poskytnutí či zamítnutí požadavků pro přístup k určitým zdrojům na základě mluveného slova. Většina aplikací, ve kterých je použit hlas jako klíč pro ověření totožnosti mluvčího, je klasifikována jako VM. Na rozdíl od IM je v případě VM identita mluvčího známa (na základě jeho tvrzení, případně nějakého kódu atd.) a systém musí ověřit, zda mluvčí je skutečně tím, kým tvrdí, že je. Jednoduché blokové schéma systému pro VM je zobrazeno na obr. 4.3. Kritický problém, který se při aplikaci VM může vyskytnout, je pokus neoprávněné osoby o neautorizovaný vstup (např. přehrání pásku s hlasem registrovaného mluvčího říkajícího klíčová slova nebo věty). K překlenutí tohoto problému existují řešení, která pro každého mluvčího generují náhodnou sadu

klíčových slov (mohou to být například číslice jdoucí v určitém pořadí za sebou nebo posloupnost slov).

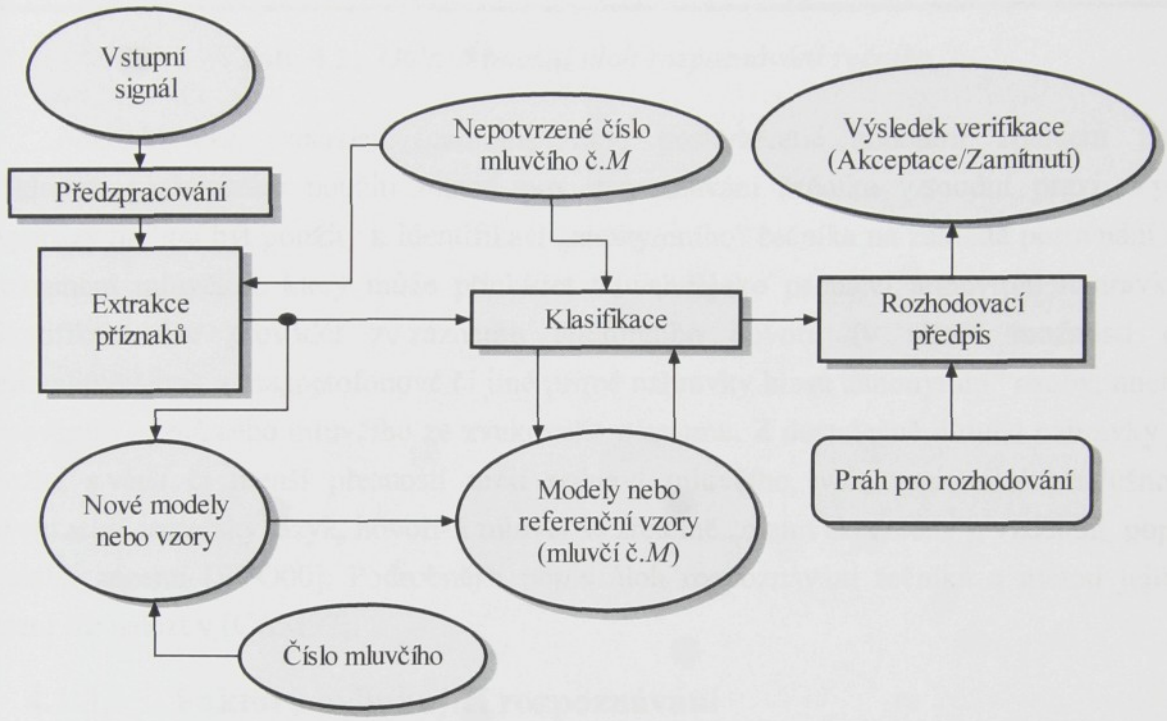


Obr. 4.2: *Blokové schéma systému pro identifikaci mluvčího*

Metody rozpoznávání řečníka lze dále rozdělit do dvou skupin podle toho, jaké typy promluv se při rozpoznávání používají. V případě rozpoznávání řečníka z daného textu jde o *textově závislé rozpoznávání*. Zde je vyžadováno, aby neznámý řečník vyslovil předem definovaný text. Je nutné, aby tento text byl identický při trénování i při rozpoznávání. Mohou to být například stejná klíčová slova, jejich různé kombinace a nebo tajný kód příslušející mluvčímu. Tento kód může být kupříkladu sestaven z číslic oddělených pauzami. Při tomto způsobu rozpoznávání mluvčího se samozřejmě předpokládá, že řečník si přeje být rozpoznán a je ochotný vyžadovanou promluvu vyslovit.

Naproti tomu u rozpoznávání řečníka z libovolného textu (*textově nezávislé rozpoznávání*) není na text kladeno naprosto žádné omezení a neznámý řečník může vyslovit v podstatě to, co ho právě napadne. V tomto případě je rozhodovací proces postaven pouze na charakteristikách řečového traktu mluvčího a na charakteristikách

buzení hlasu. Tento způsob rozpoznávání nachází využití hlavně v soudních vědách a kriminalistice (souhrnně *forensní aplikace*), kdy si řečník většinou nepřeje být rozpoznán, a proto není ochoten vyslovit požadovanou promluvu tak, jak je to vyžadováno u textově závislých systémů. Další možné využití této metody je v případě, že rozpoznání má proběhnout bez vědomí mluvčího.



Obr. 4.3: *Blokové schéma systému pro verifikaci mluvčího*

Různé oblasti použití jednotlivých úloh rozpoznávání řečníka jsou uvedeny v tabulce 4.2. V bezpečnostních systémech, které zajišťují např. fyzický vstup osob do objektů (hlasové zámky), přístup k databázím s tajnými nebo důvěrnými informacemi nebo provádění bankovních operací po telefonu, se obvykle využívá VM. U těchto systémů se předpokládá, že řečník si přeje být rozpoznán, a proto je ochoten prokázat svou totožnost. V soudních vědách a kriminalistice může najít své uplatnění jak verifikace, tak i identifikace řečníka. Verifikaci lze využít např. tehdy, existuje-li záznam hlasu pachatele z místa činu a je-li třeba ověřit, zda hlas na záznamu je shodný s hlasem určité podezřelé osoby. Identifikaci je pak možné využít v případě, že podezřelých je více a je přitom třeba určit konkrétního pachatele.

	IDENTIFIKACE		VERIFIKACE	
	v uzavřené množině	v otevřené množině	nezávislá na textu	závislá na textu
BEZPEČNOSTNÍ SYSTÉMY				
fyzický vstup				ANO
přístup k databázím				ANO
telefonní transakce			ANO	ANO
SOUDNÍ VĚDY	ANO	ANO	ANO	ANO
KRIMINALISTIKA	ANO	ANO	ANO	ANO

Tab. 4.2: *Oblasti použití úloh rozpoznávání řečníka*

Fonoskopické expertizy (*audioexpertizy*) poskytované soudním znalcem jsou příkladem praktického použití metod pro rozpoznávání řečníka v soudní praxi. Tyto expertizy mohou být použity k identifikaci „anonymního" řečníka na základě porovnání se záznamem mluvčího, který může přicházet v úvahu jako pachatel anonymní nahrávky. Identifikace lze provádět ze záznamu telefonního hovoru (v rámci možností co nejkvalitnějšího), z magnetofonové či jiné přímé nahrávky hlasu „anonymní" osoby, anebo typováním neznámého mluvčího ze zvukového záznamu. Z dostatečně dlouhé nahrávky je možné s větší či menší přesností určit pohlaví mluvčího, věk, regionální příslušnost (eventuelně mateřský jazyk, hovoří-li mluvčí se zřetelně „cizím akcentem"), vzdělání, popř. sociální zázemí [SVO00]. Podrobnější popis úloh rozpoznávání řečníka a metod jejich řešení lze nalézt v [CAM97].

4.1.3. Faktory ovlivňující rozpoznávání

Při provozu systému pro IM nebo VM je nutné počítat s faktory, které jsou mimo působnost vnitřních algoritmů programu a mohou v konečném důsledku způsobit chyby při rozpoznávání. Těmito faktory mohou být:

- nesprávně vyslovené či přečtené slovo,
- extrémní emoční stres,
- různé umístění mikrofonu, různé mikrofony pro testování a učení,
- špatná akustika v místnosti (vícečetný zvuk, hluk vznikající uvnitř i vně místnosti),
- velký časový odstup od namluvení referenčních a testovacích vzorků,
- dočasné změny řečového traktu (nachlazení).

Pokud bychom chtěli zkonstruovat robustní systém, je nutné vzít v úvahu všechny tyto faktory, protože právě ty přímo ovlivňují výkonnost rozpoznávacího algoritmu.

4.1.4. Obecné části systému pro rozpoznávání mluvčího

Jak vyplývá ze schémat na obr. 4.2 a obr. 4.3, systém pro IM nebo VM by měl obsahovat níže uvedené části. Prvním úkolem je převod akustické vlny na signál zpracovatelný v počítači. O to se stará blok pro *digitalizaci řečového signálu*. Velkou výhodou úlohy rozpoznávání mluvčího na počítači je její finanční nenáročnost (viz obr. 4.1). Pro nahrání řečového signálu do počítače je potřeba pouze mikrofon a zvuková karta, což je v současné době standardní výbava každého osobního počítače. Potřebný software pro nahrávání je dodáván s operačním systémem nebo také se zvukovou kartou. O náležité převedení analogového signálu na digitální se postará hardware zvukové karty. Dalším krokem je *extrakce příznaků* (podrobněji popsáno v kapitole 3), jež spočívá v rozčlenění signálu na intervaly o délce 10–30 ms. Každý tento interval je reprezentován vícerozměrným vektorem příznaků.

Pokud již máme signál vyjádřený pomocí vhodných příznaků, můžeme přejít k samotnému hledání správné referenční promluvy, která je také vyjádřena příznaky. Tato fáze se nazývá *rozpoznávání* (*klasifikace*). Pro jednoduchost zde budeme uvažovat systém pracující s referenčními vzory slov. U těchto systémů je sekvence příznakových vektorů porovnávána s referenčními modely (ty musí být reprezentovány stejnými příznaky). K porovnání použijeme metody pro rozpoznávání obrazů. Klasifikace pro danou promluvu a referenci vyústí ve *výsledné identifikační skóre*. Toto skóre nám pak poskytuje informaci o podobnosti testované a referenční promluvy. Protože je nutné do systému průběžně ukládat nové referenční promluvy, musíme do něj zařadit smyčku, která umožní *učení a ukládání nových referenčních obrazů*.

Jako poslední přichází na řadu *rozhodovací člen* (v nejjednodušším případě nějaké pravidlo), jímž vyhodnotíme provedená porovnání. U základní úlohy IM (pro příklad uvažujme uzavřenou sadu, textově nezávislý systém, vektorovou kvantizaci) je to minimální kumulovaná vzdálenost. U úlohy VM se jedná o rozhodnutí, zda mluvčí je opravdu tím, kým tvrdí, že je. Tato úloha je o něco obtížnější, ale v podstatě se jedná o ověření dvou hypotéz: mluvčí je skutečně oním člověkem, za kterého se prohlašuje, a nebo jím není.

Snahu o tvorbu systému pro rozpoznávání mluvčího lze tedy charakterizovat takto: získat rozpoznávací systém vykazující *vysokou rozlišovací schopnost*, *vysokou adaptabilitu* na různé mluvčí a *nízkou variabilitu* v rámci jednoho mluvčího při *přijatelných výpočetních nárocích* (pokud možno systém pracující v reálném čase). V neposlední řadě budeme po takovém systému požadovat, aby byl co *nejrobustnější* (co nejméně závislý na rušivých vlivech okolí).

4.2. Přehled metod používaných při rozpoznávání mluvčích

Pro oblast rozpoznávání mluvčích můžeme v principu použít všechny metody navržené pro rozpoznávání řeči s tím, že se v nich budeme snažit o zvýraznění těch rysů, které od sebe odliší jednotlivé mluvčí. Jde tedy o pravý opak cíle, kterého chceme dosáhnout při návrhu systému pro rozpoznávání řeči, u něhož se většinou vyžaduje, aby byl nezávislý na hovořící osobě. U rozpoznávání mluvčího je kladen hlavní důraz na zachycení a modelování řečových charakteristik jednotlivých řečníků, naopak potlačeny mohou být (a v některých případech dokonce musí, viz textově nezávislé rozpoznávání) informace o vlastním obsahu sdělení i o jeho časovém průběhu. Výběr konkrétní metody rozpoznávání mluvčích závisí také na druhu a kvalitě řečových dat, očekávané úspěšnosti rozpoznávání, snadnosti trénování, rozšiřování databáze mluvčích a v neposlední řadě i na výpočetních nárocích a systémových požadavcích. V následujícím přehledu jsou uvedeny možnosti využití základních technik používaných v úloze rozpoznávání mluvčího.

4.2.1. Metody pro textově závislé rozpoznávání mluvčího

Textově závislé rozpoznávání většinou bývá základem úlohy verifikace totožnosti řečníka, ale může být i součástí obecného systému rozpoznávání řečníka, tj. osoby, která má být „hlasově prověřena", má říci „heslo", kterým může být slovo či celá věta. Systém zanalyzuje signál, porovná ho se vzory uloženými v paměti a na základě zjištěné míry podobnosti rozhodne o identitě osoby. K řešení takto definované úlohy lze v zásadě použít nejjednodušší metodu pro porovnávání dvou řečových signálů, tj. *dynamické borcení času* (Dynamic Time Warping – DTW). Její princip je detailně vysvětlen v [NOU01].

Metoda DTW

DTW je nejpopulárnější metoda pro kompenzaci časové variability řečového signálu v systémech používajících referenční modely. Předpokládejme, že promluva řečníka, kterého chceme identifikovat, byla zparametrizována a je popsána časovou posloupností vektorů příznaků $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ... \mathbf{x}_t ... \mathbf{x}_T)$. V paměti systému se nacházejí referenční vzory odpovídající stejným způsobem zparametrizovaným záznamům hesel, které předtím nahrály všechny osoby s přístupem do systému. Předpokládejme, že počet těchto osob je M a referenčním vzorem osoby s indexem m je sekvence $\mathbf{R}^m = (\mathbf{r}_1^m, \mathbf{r}_2^m ... \mathbf{r}_t^m ... \mathbf{r}_{Tm}^m)$. Míru podobnosti mezi posloupnostmi $\mathbf{X}$ a $\mathbf{R}^m$ určíme jako vzdálenost $D(\mathbf{X}, \mathbf{R}^m)$ měřenou pomocí algoritmu DTW:

$$D(\mathbf{X}, \mathbf{R}^m) = \underset{f}{Min} \sum_{t=1}^{T} d(\mathbf{x}_t, \mathbf{r}_{f(t)}^m), \tag{4.1}$$

kde *f(t)* je transformační funkce splňující podmínky metody DTW. Při výpočtu se míra podobnosti průběžně optimalizuje vzhledem k nelineárnímu zobrazení *f* zavedenému mezi prvky posloupností **X** a $\mathbf{R}^m$ tak, aby celková vzdálenost byla minimální. Oba signály (v nejjednodušších případech například energie signálů) by si měli, po provedení zásahů do časové osy, co nejvíce odpovídat. Nejdůležitějším prvkem celého algoritmu je transformační funkce *f*, která musí splňovat řadu podmínek vycházejících ze struktury řečových obrazů v časové ose. Tyto podmínky je potom třeba respektovat při hledání optimální cesty [NOU01]. Některé z těchto podmínek však lze modifikovat a vytvořit tak různé varianty algoritmu DTW.

Máme-li tedy určit, které z možných osob nejvíce odpovídá daná promluva, musíme najít nejmenší z výše uvedených vzdáleností, tedy:

$$m^* = \underset{m}{ArgMin}\, D(\mathbf{X}, \mathbf{R}^m) = \underset{m}{ArgMin}\left[\underset{f}{Min} \sum_{t=1}^{T} d(\mathbf{x}_t, \mathbf{r}^m_{f(t)}) \right]. \qquad (4.2)$$

Vztahem (4.2) lze určit index osoby *m** , která je nejpravděpodobnějším kandidátem z daného okruhu osob.

Úspěšnost a tudíž i použitelnost základní metody můžeme zvýšit dalšími faktory

a) Každá osoba má jiné přístupové heslo, které je známo pouze jí a systému. Rozpoznávání se tedy opírá především o fonetickou podobu hesla a jen částečně o hlasovou charakteristiku řečníka. Heslem nemusí být jen jediné slovo, ale může to být i celá věta, což dále zvyšuje možnost lepšího odlišení různých osob.

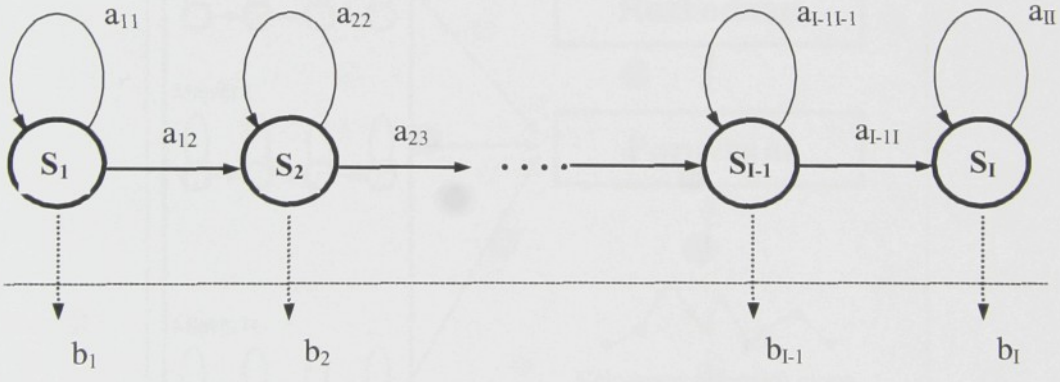
b) V přípravné (trénovací) fázi každá osoba vysloví přístupové heslo vícekrát (ideálně s časovým odstupem mezi jednotlivými promluvami). Podle vztahu (4.1) se z nahrávek určí vzor nejlépe reprezentující danou osobu a zároveň se změří míra odchylky mezi jednotlivými promluvami téže osoby. Tuto míru pak lze využít pro nastavení prahu pro odmítnutí.

c) V přípravné fázi je několik dalších osob požádáno, aby též vyslovily stejná hesla jako oprávnění mluvčí. Reprezentace těchto promluv mohou být zařazeny do databáze vzorů, s nimiž se provádí klasifikace podle vztahu (4.2). Tyto reprezentace mohou být zároveň použity pro upřesnění hodnoty prahu podle bodu b). Systém je pak schopen lépe odolávat pokusům o neoprávněný přístup.

Tato metoda je relativně výpočetně nenáročná (ať už se jedná o režim trénování či testování) a snad i proto je stále ještě používána v některých komerčně dostupných systémech. Přesto je nutné poukázat na poměrně vysokou citlivost vůči externím vlivům a z toho vyplývající nesnadnost nastavit spolehlivý verifikační práh.

Metoda HMM

Princip metody modelování řeči *skrytými markovskými modely* (Hidden Markov Models – HMM) vychází z představy o vytváření řeči. Při generování řeči člověkem si lze představit, že hlasové ústrojí je během krátkého časového intervalu (např. framu) v jednom z konečného počtu stavů artikulačních konfigurací (např. generuje určitý foném). V uvažovaném krátkém časovém období je pak hlasovým ústrojím produkován krátký signál, který závisí na stavu artikulačního ústrojí a může být popsán vícerozměrným vektorem příznaků. Každý stav markovského modelu pak reprezentuje časovou posloupnost příznakových vektorů modelované řečové jednotky (například již zmíněného fonému), v níž se jednotlivé parametry příznakových vektorů mění jen málo. Jinak řečeno, stavy markovských modelů v sobě nesou statistickou informaci o typických hodnotách příznaků v těchto vektorech. Tato informace je pro každý stav dána parametry (středními hodnotami a rozptyly) výstupní pravděpodobnostní funkce b_i s normálním (Gaussovým) rozložením. Kromě parametrů této funkce se dá také ve fázi trénování statisticky vyhodnotit, kolik framů signálu řeči daný stav představuje a na základě toho určit pravděpodobnosti setrvání v jednotlivých stavech a také pravděpodobnosti přechodů mezi jednotlivými stavy uvnitř modelu.



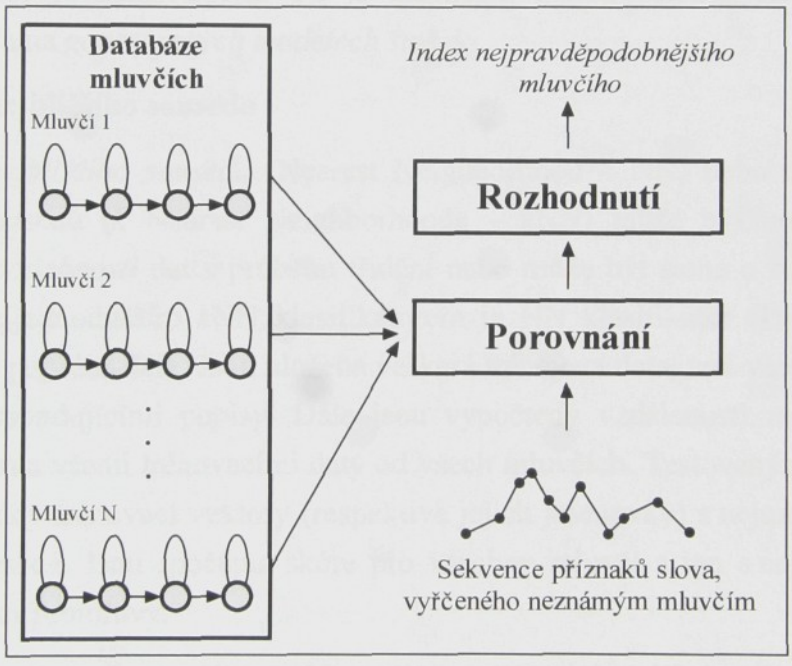
Obr. 4.4: *HMM – struktura modelu, která je obvyklá pro reprezentaci promluvy*

Z hlediska struktury se v úloze textově závislého rozpoznávání mluvčího, kde kromě charakteristik řečníka chceme zachytit i lingvistický obsah promluvy, prioritně využívají tzv. levo-pravé markovské modely. Používáme je tam, kde chceme zachytit nějaký vývoj s postupujícím časem. Z obr. 4.4 vyplývá, že v takto navržené lineární struktuře je přechod mezi stavy možný pouze mezi dvěma sousedními stavy a jen zleva doprava. Smyčky u každého stavu na obr. 4.4 pak znázorňují situaci, kdy model setrvává v daném stavu. Pravděpodobnost přechodu do sousedního stavu je dána hodnotou a_{ii+1}, pravděpodobnost setrvání v daném stavu hodnotou a_{ii}. Protože oba jevy jsou komplementární, platí

$a_{ii} + a_{ii+1} = 1$. První frame řečového signálu přitom vždy musí být přiřazen prvnímu stavu modelu a poslední frame signálu poslednímu stavu.

Metodu skrytých markovských modelů lze použít prakticky stejným způsobem jako v předešlém případě metodu DTW. Každá osoba má svůj model natrénovaný na základě několika nahrávek svého přístupového hesla. Při rozpoznávání se pak hledá osoba, jejíž model dosahuje nejvyšší pravděpodobnosti (viz obr. 4.5). Vzhledem k faktu, že k určení parametrů modelu je potřeba více nahrávek téhož hesla, má v sobě model již schopnost lépe určovat míru věrohodnosti, že osoba s nejvyšším skóre je ta pravá. I zde je však dobré využít další faktory zvyšující pravděpodobnost úspěšného rozpoznání, které byly naznačené v předchozí kapitole.

Rozdíl oproti metodě DTW tkví především v tom, že databáze mluvčích není složena z referenčních vzorů promluv, ale z modelů mluvčích. Další podrobnosti o metodě HMM lze nalézt v [NOU01] nebo [RAD04].



Obr. 4.5: *Klasifikátor založený na metodě HMM. Mluvčí ve slovníku jsou reprezentováni modely stejných slov.*

4.2.2. Metody pro textově nezávislé rozpoznávání mluvčího

K rozpoznávání mluvčího, které má být nezávislé na obsahu promluvy, již nelze použít metodu DTW a ani výše uvedenou metodu HMM. Teoreticky by bylo možné použít klasický levo-pravý hláskově orientovaný rozpoznávač postavený na metodě HMM. Každý mluvčí by měl v paměti systému sadu modelů všech hlásek a systém by pro každou takovou sadu vyhodnotil nejpravděpodobnější sekvenci hlásek. K vyhodnocení bychom

využili Viterbiho dekodér, který je známý z úlohy rozpoznávání řeči (viz kapitola 5.2). Mluvčí, jehož sekvence by získala nejvyšší skóre, by byl prohlášen za hlavního kandidáta při rozpoznávání. Tento způsob by však byl velmi náročný na množství řeči, jež by každá osoba musela dodat pro natrénování modelů. Prakticky použitelný by byl pouze v případech, kdy jsou k dispozici rozsáhlé záznamy promluv konkrétní osoby a kdy se spíše jedná o ověření hypotézy, že jiné hlasové záznamy patří téže osobě. Dále by pak bylo možné pro identifikaci a verifikaci mluvčích využít ergodických HMM. V tomto případě jsou plně propojené markovské modely trénovány (Viterbiho algoritmus nebo „Forward-Backward" algoritmus) řečovými daty jednotlivých uživatelů. Propojenost modelu pak umožňuje sledovat časový vývoj promluvy a dovoluje tak nasazení HMM i pro textově nezávislou IM a VM.

Metody DTW a HMM využívají informaci o časovém vývoji parametrů signálu řeči. Při rozpoznávání obsahu řeči je to nutné, v úloze identifikace mluvčího se však bez této časové informace můžeme obejít. To je základem metod založených na *vektorovém kvantování* řeči a na *gaussovských modelech směsí*.

Metoda nejbližšího souseda

Metoda *nejbližšího souseda* (Nearest Neighborhood – NN) nebo obecněji metoda *k* nejbližších sousedů (*k* Nearest Neighborhoods – *k*NN) může být použita k výpočtu *hustoty pravděpodobnosti* dat v průběhu třídění nebo může být sama o sobě použita jako klasifikátor. Nejjednodušším *k*NN klasifikátorem je NN klasifikátor (*k*NN, kde $k = 1$), jehož činnost je popsána dále. Jsou uložena veškerá trénovací data, tzn. všechny příznakové vektory s korespondujícími popisy. Dále jsou vypočteny vzdálenosti mezi testovanými vektory příznaků a všemi trénovacími daty od všech mluvčích. Testovaným vektorům jsou následně přiřazeny trénovací vektory (respektive jejich jmenovky) s nejmenší vzdáleností, tj. nejbližší sousedé. Jsou spočtena skóre pro všechny mluvčí a ten s nejlepším skóre je vybrán jako autor promluvy.

Mezirámcová matice vzdáleností je spočtena měřením vzdálenosti mezi testovanými framy (vstup) a žadatelovými referenčními framy, které byly uloženy v režimu trénování. NN vzdálenost je minimální vzdálenost mezi testovanými a zaznamenanými framy. Pro vypočtení konečného výsledku je třeba všechny NN vzdálenosti testovaných framů zprůměrovat. Testované framy jsou ještě porovnány s uloženými referencemi od dalších mluvčích. Výsledky jsou pak zkombinovány a je z nich vypočten výsledný pravděpodobnostní koeficient z.

Klasifikátor založený na metodě nejbližšího souseda uchovává všechna trénovací data a pracuje tak, že hledá nejbližšího souseda mezi všemi daty. Tento postup je ovšem velmi

náročný jak na operační paměť, tak i na výpočetní výkon počítače, proto se většinou při výpočtu používá techniky prořezávání (pruningu) příznakových vektorů, která tento handicap snižuje.

Vektorová kvantizace

Vektorové kvantování (Vector Quantization – VQ) je metoda, která spojité vektorové hodnoty transformuje na konečný počet diskrétních vektorových hodnot. Představuje zobecnění klasického skalárního kvantování, dobře známého z procesu digitalizace, kdy jsou spojitým hodnotám analogového signálu přiřazována čísla z omezeného souboru. Např. v osmibitovém AD převodníku jsou všechna napětí převedena do škály 256 hodnot.

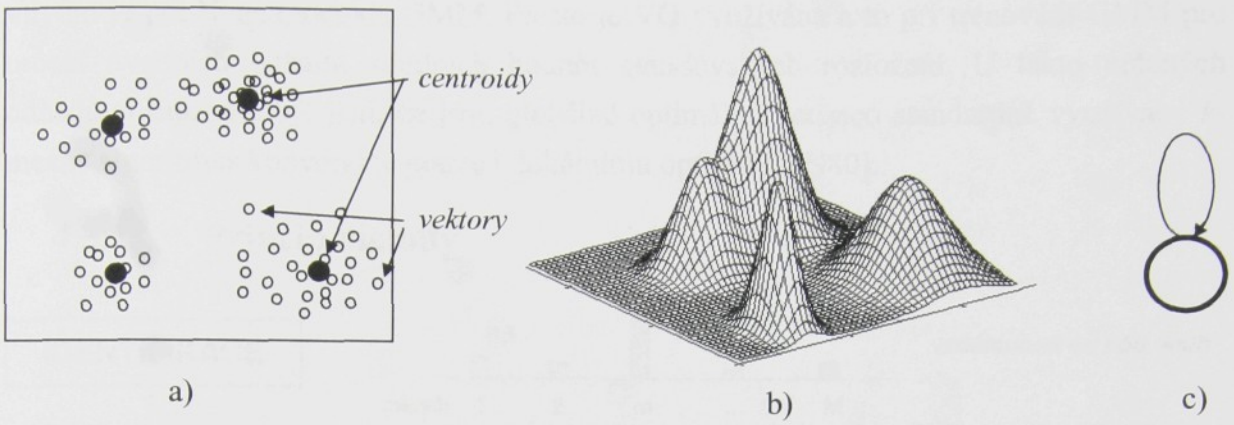
Jelikož byla tato metoda použita v našem systému pro identifikaci mluvčích, bude podrobně popsána v samostatné kapitole 4.3.

Gaussovské modely směsí

Pokud bychom chtěli co nejjednodušeji vysvětlit princip metody GMM v systémech rozpoznávajících řečníka, je možná jistá analogie s VQ. Obě metody mají společné to, že pracují s jistým druhem kódové knihy.

Při vektorové kvantizaci jsou shluky příznakových vektorů reprezentovány pouze centroidy, tedy jakýmisi těžišti těchto shluků – viz obr. 4.6a. Mnohem úplnější informaci o poloze a tvaru shluků získáme, jestliže se pokusíme stanovit hustotu pravděpodobnosti vektorů v prostoru. Vzhledem k existenci shluků nelze aplikovat jednomodální gaussovské rozložení, ale je potřeba použít lineární kombinaci více gaussovských rozložení. Tento způsob modelování nazýváme *gaussovské modely směsí*. Vektor příznaků modelujeme jako vektor náhodných veličin, jejichž sdružené rozložení aproximujeme součtem gaussovských hustot pravděpodobnosti. Sekvence příznakových vektorů je poté chápána jako posloupnost nezávislých realizací zmíněného vektoru. Zkombinováním více gaussovských rozložení můžeme aproximovat jakékoliv rozložení hustoty pravděpodobnosti. A právě to je velmi vhodné pro aplikace textově nezávislého rozpoznávání mluvčích.

Ukázka výběrové hustoty pravděpodobnosti složené ze 4 mixtur je na obr. 4.6b a odpovídá přibližně rozložení vektorů znázorněných na obr. 4.6a. Hlavní rozdíl mezi VQ a GMM je tedy v tom, že jednotlivé gaussovské funkce jsou popsány hustotou pravděpodobnosti rozložení *všech* příznaků (každý příznak nějakým způsobem ovlivní gaussovskou funkci), kdežto jednotlivé příznaky v případě vektorové kvantizace jsou přiřazeny pouze jednomu centroidu.



Obr. 4.6: *GMM – gaussovský model směsí*
a) znázornění prostorového rozložení vektorů a centroidů
b) odpovídající 4mixturový model, c) GMM jako zvláštní případ jednostavového HMM

Samozřejmě, že kromě uvedených existuje ještě celá řada dalších metod používaných v úlohách IM a VM jako například umělé neuronové sítě nebo „Support Vector Machines" (SVM) [SAN03]. Svou podstatou jsou však již příliš vzdálené metodám aplikovaným v této práci, a proto zde nebudou diskutovány.

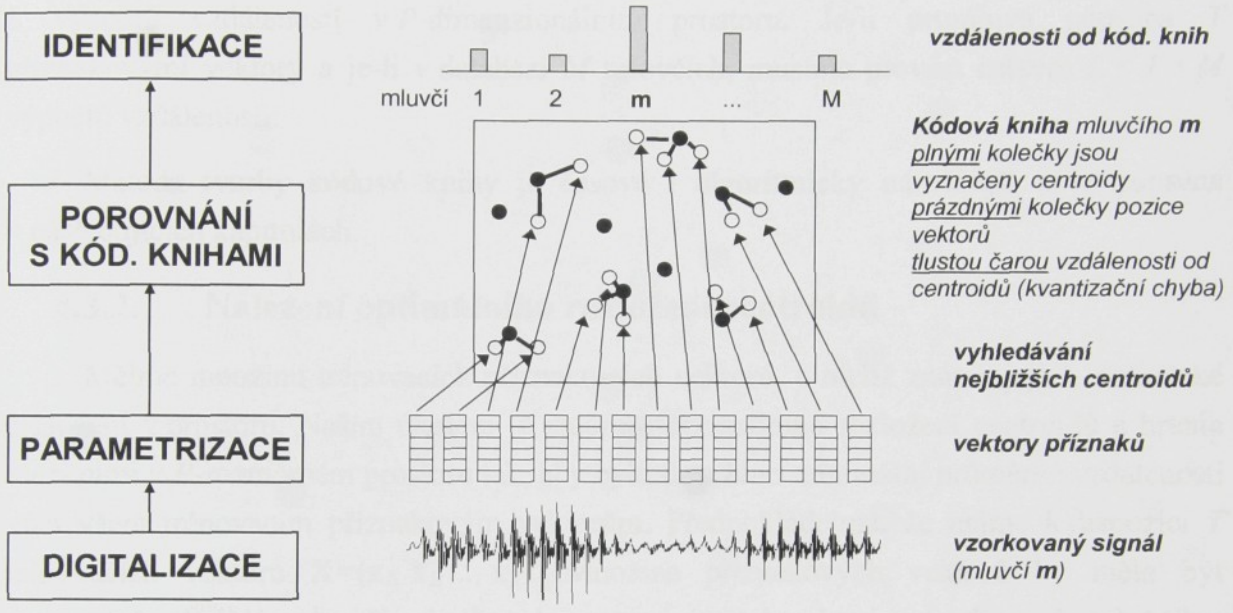
4.3. Vektorová kvantizace

Princip vektorového kvantování spočívá v tom, že prostor, v němž jsou definovány P-dimenzionální vektory, je rozdělen do L oblastí a každá oblast je reprezentována jediným vektorem, který se nazývá *centroid*. Rozdělení prostoru většinou není rovnoměrné, ale řídí se skutečným prostorovým rozmístěním vektorů v dané úloze. Tyto vektory často tvoří nepravidelné shluky a právě tyto shluky jsou jádry zmíněných oblastí, z nichž se metodami založenými na průměrování stanovují centroidy. Centroidy lze proto považovat za jakási těžiště shluků a nejlepší reprezentanty vektorů v daném prostoru. Metoda vektorového kvantování se často používá pro úsporné kódování a přenos vektorových signálů. Z této skutečnosti vyplývá i terminologie VQ. Centroidy se běžně označují jako *kódová slova*, která tvoří *kódovou knihu* o velikosti L. Princip úsporného kódování spočívá v tom, že místo, aby se přenosovým kanálem přenášely původní vektory, vysílá se pouze index nejbližšího kódového slova. Redukce datového toku může být značná. Např. k přenosu 10 složkového vektoru složeného z čísel typu *float* by bylo třeba $10 \times 4 = 40$ bytů. Při úsporném přenosu s využitím kódové knihy o 256 slovech by stačil pouze jediný byte; došlo by tedy ke čtyřicetinásobné úspoře.

V našem současném systému pro rozpoznávání mluvčích a pohlaví není metoda VQ použita přímo pro IM a VM (i když taková konfigurace již byla implementována a úspěšně otestována viz [DAV01a]). Z důvodu lepších výsledků a vyšší odolnosti proti rušivým

vlivům je používána metoda GMM. Přesto je VQ využívána a to při trénování GMM pro určení úvodních odhadů středních hodnot gaussovských rozložení. U takto určených odhadů si můžeme být jisti, že jsou globálně optimální, zatímco standardně využívaný *k*-means algoritmus konverguje pouze k lokálnímu optimu [LIN80].

4.3.1. Princip metody



Obr. 4.7: *Proces identifikace mluvčího metodou vektorové kvantizace*

Aplikaci VQ v úloze rozpoznávání mluvčího lze popsat následujícím způsobem. Každý mluvčí je požádán, aby vyslovil několik trénovacích promluv. Jejich počet nemusí být velký, ale vychází z požadavku, aby věty pokryly většinu hlásek vyskytujících se v daném jazyce. Signál těchto nahrávek je zparametrizován s použitím P příznaků. Pro každého mluvčího je vytvořena vlastní kódová kniha, nejlépe reprezentující všechny takto získané vektory P-dimenzionálního prostoru. Kódovou knihu mluvčího s indexem m tvoří L vektorů $\{\mathbf{V}_1^m ....\mathbf{V}_L^m\}$. V procesu identifikace promluvy neznámého mluvčího se vypočítá vzdálenost sekvence příznakových vektorů $\mathbf{X}$ od kódových knih jednotlivých osob:

$$D(\mathbf{X},\mathbf{V}^m)=\sum_{t=1}^{T} Min_{i=1..L}(\mathbf{x}_t,\mathbf{V}_i^m)\,, \tag{4.3}$$

kde $\mathbf{V}_i^m$ je kódové slovo nejbližší vektoru $\mathbf{x}$ v čase t. Tato vzdálenost se nazývá *celková kvantizační chyba* (nebo také *kvantizační zkreslení*). Mluvčí, jehož kódová kniha má nejnižší celkovou vzdálenost stanovenou podle vztahu (4.3), se stává vítězem klasifikace – viz obr. 4.7. Na tomto obrázku je ukázáno, jak by mohlo vypadat rozdělení ve

dvojdimenzionálním vektorovém prostoru. Každý shluk příznaků je reprezentován jedním centroidem.

Vztah (4.3) se podobá vztahu (4.1). Významný rozdíl však spočívá ve skutečnosti, že vektory kódové knihy **V** nejsou časově uspořádány tak, jako vektory referencí **R**. Nehledá se zde proto časové, nýbrž prosté vzdálenostní přiřazení. Náročnost výpočtu rozpoznávání lze odhadnout z následující úvahy. Pro nalezení nejbližšího kódového slova v čase t je třeba L výpočtů vzdáleností v P-dimenzionálním prostoru. Je-li promluva popsána T příznakovými vektory a je-li v databázi M mluvčích, musíme provést celkem $L \times T \times M$ výpočtů vzdáleností.

Metoda tvorby kódové knihy je časově i algoritmicky náročnější a je popsána v následujících kapitolách.

4.3.2. Nalezení optimálního rozložení centroidů

Mějme množinu trénovacích příznakových vektorů, u nichž známe jejich statistické rozložení v prostoru. Naším úkolem je nalezení optimálního rozložení centroidů a hranic mezi nimi v P-rozměrném prostoru tak, aby bylo dosaženo minimální průměrné vzdálenosti vůči všem trénovacím příznakovým vektorům. Předpokládejme, že máme k dispozici T trénovacích vektorů $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$. Množina příznakových vektorů by měla být dostatečně obsáhlá, aby **X** obsahovala co nejvíce charakteristických rysů mluvčího. Připomeňme ještě, že sekvence příznakových vektorů nacházejících se v prostoru S je P-rozměrná, tj.

$$\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tP}), \quad t = 1, 2, \ldots, T. \tag{4.4}$$

Dále mějme určen i počet kódových slov L v kódové knize $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L\}$. Protože kódová slova jsou stejně jako příznakové vektory P-rozměrná, platí

$$\mathbf{v}_l = (v_{l1}, v_{l2}, \ldots, v_{lP}), \quad l = 1, 2, \ldots, L. \tag{4.5}$$

Prostor S poté rozdělme na regiony B_l asociované s jednotlivými kódovými slovy, takže můžeme psát $S = \{B_1, B_2, \ldots, B_L\}$. Pokud příznakový vektor $\mathbf{x}_t$ leží v buňce B_l, potom jeho aproximací je právě $\mathbf{v}_l$:

$$Q(\mathbf{x}_t) = \mathbf{v}_l, \quad \text{pokud } \mathbf{x}_t \in B_l. \tag{4.6}$$

Pro vyjádření míry podobnosti (vzdálenosti) mezi příznakovým vektorem a kódovým slovem se nejčastěji používá euklidovská vzdálenost (squared-error). Průměrnou vzdálenost tedy vypočteme podle vztahu:

$$D = \frac{1}{T \cdot P} \cdot \sum_{l=1}^{L} \left\| \mathbf{x}_t - Q(\mathbf{x}_t) \right\|^2, \qquad \|e\|^2 = e_1^{\,2} + e_2^{\,2} + \ldots + e_P^{\,2}. \tag{4.7}$$

Úloha vytvoření kódové knihy by tedy mohla být stručně vyjádřena takto. Na základě známých veličin **X** a L najděte **V** a S takové, aby vzdálenost D byla minimální. Za předpokladu, že **V** a S jsou řešením naší úlohy, musí splňovat následující dvě podmínky:

- Podmínka nejbližšího souseda

$$B_l = \{\mathbf{x}: \| \mathbf{x} - \mathbf{v}_l \|^2 \leq \| \mathbf{x} - \mathbf{v}_{l'} \|^2, \quad \forall l' = 1, 2, \ldots, L\}. \tag{4.8}$$

Tato podmínka říká, že kódovaná oblast B_l by měla obsahovat všechny vektory, které jsou ke kódovému slovu $\mathbf{v}_l$ bližší než ke kterémukoliv jinému slovu z dané kódové knihy.

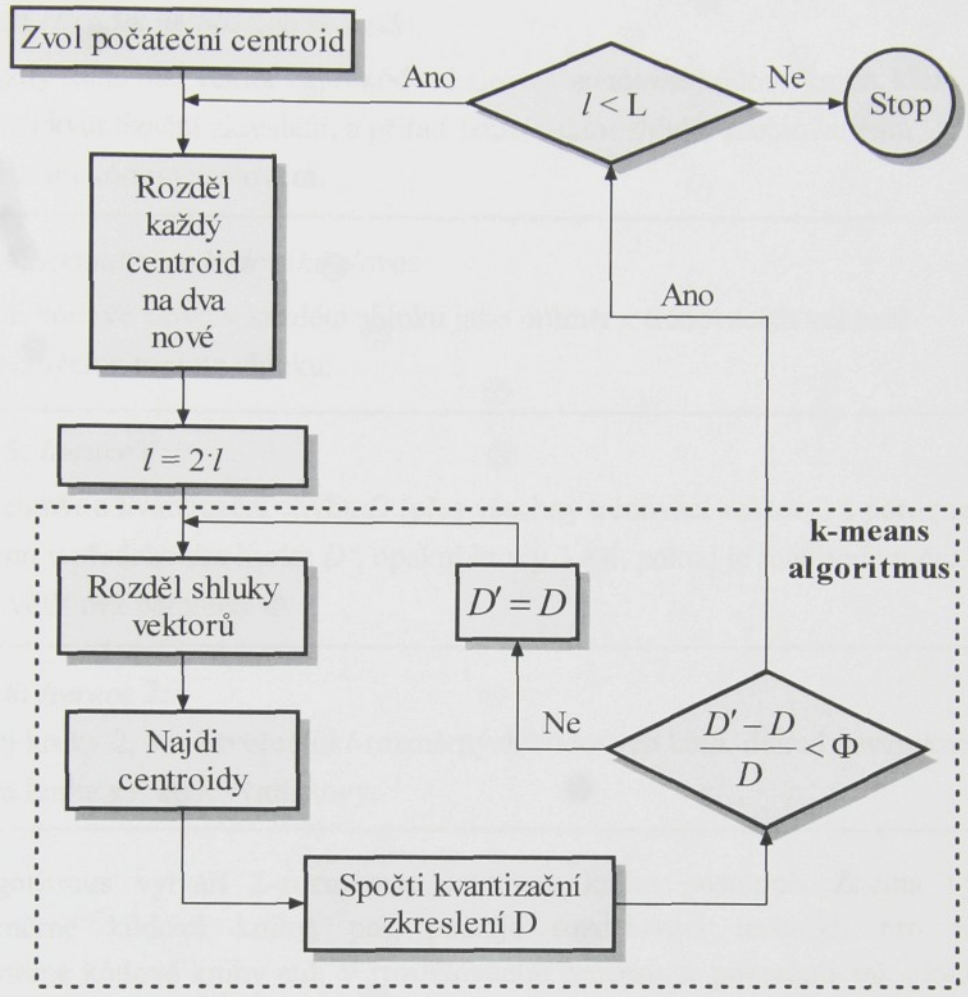
- Pro centroid musí platit:

$$c_l = \frac{\sum_{x_m \in S_l} \mathbf{x}_m}{\sum_{x_m \in S_l} l}, \qquad l = 1, 2, \ldots, L. \tag{4.9}$$

Tento vztah říká, že kódové slovo by mělo být průměrným vektorem všech trénovacích příznakových vektorů spadajících do oblasti B_l. Při implementaci je třeba dát pozor na to, aby se nikdy nevyskytlo kódové slovo, kterému neodpovídá žádný z trénovacích příznaků. V takovém případě by jmenovatel výrazu mohl být roven nule.

4.3.3. Implementace

Na rozdíl od textově závislého rozpoznávače založeného na metodě DTW je nutné u VQ ještě před procesem testování projít trénovacím režimem, tedy fází, v níž se pro každého mluvčího vytvoří kódové knihy. Jelikož inicializační kódová kniha (základní odhad, ze kterého vycházejí další iterace) má bezprostřední vliv na kvalitu výsledné kódové knihy, došlo časem k vylepšení *k-means algoritmu* [HUA01], který byl do té doby používán pro tvorbu kódových knih. Bylo dokázáno [GRA82], že přístup, ve kterém je L-rozměrná kódová kniha konstruována po částech, je globálně optimální (k-means algoritmus může konvergovat pouze k lokálnímu optimu). Tento vylepšený k-means algoritmus dostal jméno po svých tvůrcích a nazývá se *LBG* (Linde, Buzo a Gray [LIN80]). Algoritmus je implementován jako rekurzivní procedura znázorněná vývojovým diagramem na obr. 4.8. Velikost výsledné kódové knihy vypočítané tímto algoritmem je mocninou čísla 2. Tato skutečnost vyplývá z již zmíněného principu práce LBG algoritmu, který výslednou kódovou knihu vytváří postupným zdvojováním předchozích kódových knih. Tyto knihy mohou mít různé velikosti a právě na počtu elementů do značné míry závisí výsledek i rychlost rozpoznávání.



Obr. 4.8: *Vývojový diagram LBG algoritmu pro tvorbu kódové knihy*

LBG algoritmus pracuje následujícím způsobem :
Krok 1: *Vytvoření kódové knihy pouze s jedním kódovým slovem* To je centroid z úplné sady trénovacích vektorů (zde není nutná žádná iterace, pouze se nalezne vektor, který nejlépe reprezentuje všechny vektory – střední hodnota).
Krok 2: *Zdvojnásobení kódové knihy rozdělením současných shluků a vytvořením nových kódových slov* (*centroidů*) podle pravidla: $y_n^+ = y_n \cdot (1 + \Omega)$, $y_n^- = y_n \cdot (1 - \Omega)$, kde y_n je n-tý centroid z kódové knihy vzniklé v předešlé iteraci, y_n^+ a y_n^- jsou nově určené centroidy $(n = 1, \ldots l)$, Ω je vektor iteračních parametrů (obvykle 0,01–0,001).

Krok 3: *Hledání nejbližšího souseda:* Pro každý trénovací vektor najdi kódové slovo v současné kódové knize, které má nejmenší kvantizační zkreslení, a přiřaď tento vektor shluku asociovanému s nejbližším kódovým slovem.
Krok 4: *Aktualizace kódového slova:* Vypočti kódové slovo v každém shluku jako průměr z trénovacích vektorů asociovaných k tomuto shluku.
Krok 5: *Iterace 1:* Urči celkovou kvantizační chybu D (přes všechny trénovací vektory) a porovnej ji s chybou v předchozím kroku D', opakuj kroky 3 a 4, pokud je relativní změna chyby větší než parametr Φ.
Krok 6: *Iterace 2:* Opakuj kroky 2, 3 a 4 tvořením l-rozměrných kódových knih, dokud nevznikne kódová kniha s L kódovými slovy.

LBG algoritmus vytváří L-rozměrnou kódovou knihu postupně. Začíná vytvořením jednorozměrné kódové knihy, pak použije rozdělovací techniku pro inicializaci dvourozměrné kódové knihy atd. V rozdělovacím procesu se pokračuje tak dlouho, dokud není vytvořena požadovaná L-rozměrná kódová kniha. V každém iteračním kroku se určuje celková kvantizační chyba a sleduje se její konvergence. Kroky 3, 4 a 5 tvoří dohromady k-means algoritmus, který zajišťuje optimální rozdělení P-dimenzionálního prostoru do L buněk. Toto optimum je bráno jako minimální celková distorze všech vstupních vektorů.

4.4. Gaussovské modely směsí

4.4.1. Rozpoznávání mluvčích metodou GMM

Rozložení náhodné proměnné X bývá v mnoha reálných aplikacích modelováno gaussovským rozložením, pokud hustota pravděpodobnosti rozložení proměnné X o střední hodnotě $\bar{x}$ a rozptylu σ^2 je dána vztahem:

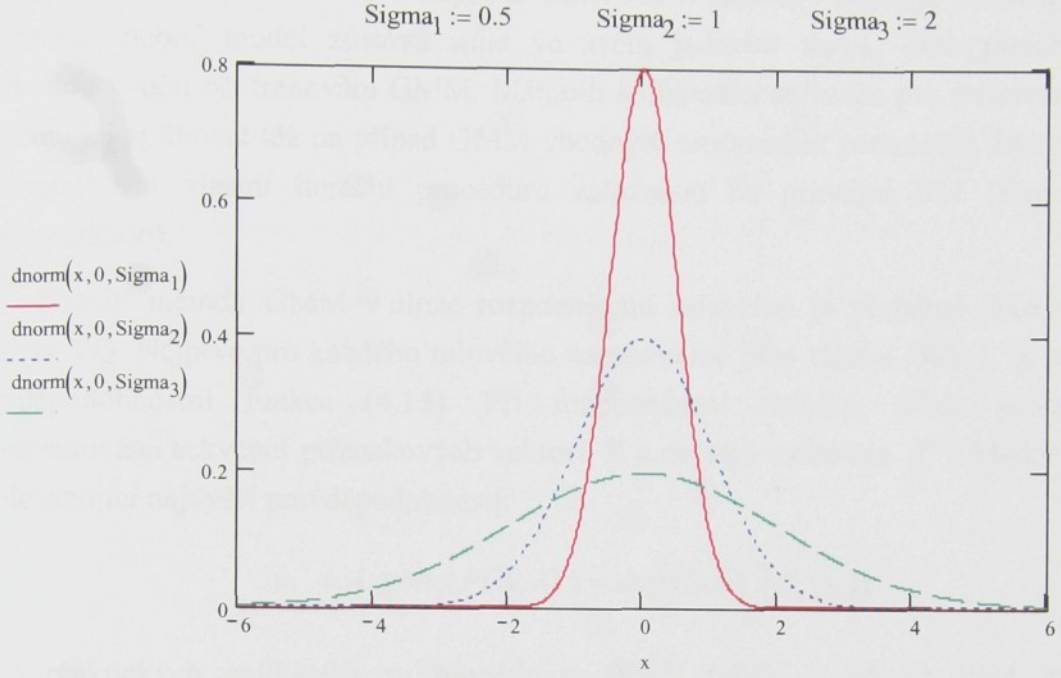
$$f(x) = N(\bar{x}, \sigma^2) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \cdot \exp\left[-\frac{(x-\bar{x})^2}{2\sigma^2} \right]. \tag{4.10}$$

Parametry $\bar{x}$ a σ^2 plně popisují normální náhodnou veličinu a platí pro ně

$$\bar{x} = E(X), \tag{4.11}$$

$$\sigma^2 = E(X - E(X))^2, \tag{4.12}$$

kde $E(.)$ má význam matematického operátoru očekávání.



$p_1 := \text{dnorm}(1,0,\text{Sigma}_1)$ $p_1 = 0.108$ $p_4 := \text{pnorm}(1,0,\text{Sigma}_1)$ $p_4 = 0.977$

$p_2 := \text{dnorm}(1,0,\text{Sigma}_2)$ $p_2 = 0.242$ $p_5 := \text{pnorm}(1,0,\text{Sigma}_2)$ $p_5 = 0.841$

$p_3 := \text{dnorm}(1,0,\text{Sigma}_3)$ $p_3 = 0.176$ $p_6 := \text{pnorm}(1,0,\text{Sigma}_3)$ $p_6 = 0.691$

Obr. 4.9: *Tři gaussovská rozdělení se stejnou hodnotou $\bar{x}$, ale každé s jinou směrodatnou odchylkou (Sigma$_1$ až Sigma$_3$). Rozdělení s nejmenší směrodatnou odchylkou má nejstrmější průběh.*

Na obr. 4.9 je vidět, že gaussovské rozdělení je symetrické kolem $\bar{x}$ a největší hodnotu má také v $\bar{x}$. Funkce *dnorm* počítá hustotu pravděpodobnosti, kde prvním parametrem této funkce je bod, pro který hodnotu počítáme. Druhým a třetím parametrem jsou již zmíněné hodnoty $\bar{x}$ *a* σ^2. Hodnoty p_1, p_2 a p_3 mají význam hustoty pravděpodobnosti spočtené v bodě jedna. V odpovídajícím řádku jsou pak p_4, p_5 a p_6 hodnotami distribuční funkce pro dané gaussovské rozložení, opět spočtené pro bod jedna.

GMM je definován jako vážená směs gaussovských (normálních) rozložení:

$$b(\mathbf{x}) = \sum_{m=1}^{M} c_m N(\bar{\mathbf{x}}_m, \Sigma_m) = \sum_{m=1}^{M} c_m \frac{1}{\sqrt{(2\pi)^P \det \Sigma_m}} . \exp[-\tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_m)^T \Sigma_m^{-1} (\mathbf{x}-\bar{\mathbf{x}}_m)], \qquad (4.13)$$

kde M je počet mixtur, c_m je váhový koeficient m-té mixtury, $\bar{\mathbf{x}}_m$ a Σ_m jsou střední hodnoty a kovarianční matice jednotlivých mixtur. Pro c_m platí, že $\sum_{m=1}^{M} c_m = 1$, $0 \le c_m \le 1$. Vztah (4.13) je ekvivalentem vztahu pro výstupní funkci jednoho stavu u vícesložkových HMM

používaných běžně při rozpoznávání řeči. GMM lze skutečně považovat za zvláštní případ HMM, jak naznačuje obr. 4.6c v kapitole 4.2.2. Jde o HMM, v němž je potlačena časová informace, neboť model zůstává stále ve svém jediném stavu. Tuto podobnost lze s výhodou využít při trénování GMM. Máme-li k dispozici software pro trénování HMM, můžeme jej aplikovat též na případ GMM vhodným nastavením parametrů. Jinak je nutné naprogramovat vlastní iterační proceduru založenou na principu EM (Expectation – Maximization).

Použití metody GMM v úloze rozpoznávání mluvčího je podobné jako v případě metody VQ. Nejprve pro každého mluvčího natrénujeme jeho vlastní GMM, tj. parametry pravděpodobnostní funkce (4.13). Při rozpoznávání řečníka, jehož promluva je reprezentována sekvencí příznakových vektorů $\mathbf{X}$ a on sám modelem λ^m [1], hledáme model m^* dosahující nejvyšší pravděpodobnosti:

$$m^* = \underset{m}{ArgMax}\, P(\mathbf{X}, \lambda^m) = \underset{m}{ArgMax}[\prod_{t=1}^{T} b_m(x_t)]. \qquad (4.14)$$

V praktických aplikacích se obvykle používají GMM s 16 až 1024 mixturami v závislosti na množství trénovacího materiálu. Bližší informace o metodě GMM lze nalézt v [REY92]. Výsledky nasazení bývají o něco málo lepší v porovnání s metodou VQ, a to díky věrnějšímu popisu prostorového rozložení příznakových vektorů jednotlivých osob.

4.4.2. Stanovení parametrů GMM metodou maximální věrohodnosti

Následující odstavce jsou věnovány způsobu reprezentace dat v prostoru při použití GMM. VQ má bohužel tu vlastnost, že rozděluje data v prostoru do určitých regionů podle měření určité vzdálenosti, ale nerespektuje již pravděpodobnostní rozložení vstupních dat. To může zapříčinit chybné vymezení hranic a následně i poškození originální struktury dat. Alternativní cestou jak nahradit toto „tvrdé" přiřazení jsou gaussovská pravděpodobnostní rozložení. Pro tyto funkce již není problém překročit hranice rozdělující prostor a mohou tak lépe reprezentovat strukturu dat. Gaussovská hustota pravděpodobnosti (PDF) je funkcí dvou parametrů. Je to vektor středních hodnot a matice kovariancí.

Při tvorbě modelu mluvčího se potýkáme s jedním zásadním problémem. Chceme zjistit optimální parametry modelu, ale nemáme k dispozici kompletní data potřebná k určení jeho parametrů. Pro hledání parametrů modelu se s úspěchem používá *kritéria*

[1] V dalším textu bude mluvčí označován i značkou S z anglického slova „speaker" a to zejména v případech, kdy nechceme předjímat, že pro reprezentaci mluvčího je použit model λ.

maximální věrohodnosti (ML) [DUD01], [SAN03], [RED84]. Toto kritérium nám pomáhá nalézt parametry modelu

$$\Theta = \{ c_m, N(\bar{x}_m, \Sigma_m) \}_{m=1}^{M} = \{ c_m, \theta_m \}_{m=1}^{M}, \tag{4.15}$$

pokud máme k dispozici pouze omezenou množinu trénovacích dat $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. Cílem tohoto kritéria je metodou *odhadu maximální věrohodnosti* (MLE) nalézt takové parametry $\Theta^* \in \Omega$, jež maximalizují věrohodnost $\ell(\Theta)$. Tato věrohodnost je díky předpokladu nezávislých dat $\mathbf{X}$ a v souladu se vztahem (4.14) definována jako

$$\ell(\Theta) = P(\mathbf{X}|\Theta) = \prod_{t=1}^{T} P(\mathbf{x}_t|\Theta). \tag{4.16}$$

Řešíme tedy následující optimalizační problém

$$\Theta^* = \underset{\Theta\in\Omega}{ArgMax}\ \ell(\mathbf{X}|\Theta) = \underset{\Theta\in\Omega}{ArgMax} \prod_{t=1}^{T} P(\mathbf{x}_t|\Theta). \tag{4.17}$$

Z důvodu výpočetní náročnosti je vhodné maximalizovat logaritmus věrohodnosti $L(\Theta) = \log(\ell(\Theta))$. To je možné díky tomu, že logaritmus je monotónně rostoucí funkce a náš optimalizační problém se tedy dá zapsat jako

$$\Theta^* = \underset{\Theta\in\Omega}{ArgMax}\ L(\mathbf{X}|\Theta) = \underset{\Theta\in\Omega}{ArgMax} \sum_{t=1}^{T} \log P(\mathbf{x}_t|\Theta). \tag{4.18}$$

Problém (4.17) nemá analytické řešení, a právě proto je třeba použít iteračních optimalizačních technik. Jednou takovou technikou je i EM algoritmus, který se v podobných úlohách využívá díky své poměrné jednoduchosti a zaručené konvergenci k lokálnímu optimu.

4.4.3. EM algoritmus

EM algoritmus je iterativní postup, který nám převádí úlohu (4.17) na sérii jednodušších, analyticky řešitelných optimalizačních problémů [DEM77]. Pro aplikaci EM algoritmu předpokládejme existenci neznámých dat $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, kde hodnoty $\mathbf{y}_t$ označují složku gaussovského rozložení, která „generovala" vektor $\mathbf{x}_t$. Tudíž $\mathbf{y}_t \in [1, M]\ \forall\, i$ a $\mathbf{y}_t = m$, pokud příznakový vektor $\mathbf{x}_t$ v čase t byl generován m-tou složkou gaussovského rozložení. Pokud bychom znali hodnoty $\mathbf{Y}$, mohl by být na základě vztahů (4.15) a (4.13) přeformulován vztah (4.16) na

$$P(\mathbf{X}, \mathbf{Y}|\Theta) = \prod_{t=1}^{T} c_{y_t} P(\mathbf{x}_t|\theta_{y_t}). \tag{4.19}$$

Jak již sám název EM algoritmu napovídá, výpočet parametrů modelu sestává ze dvou kroků. Prvním z nich je „Expectation" krok, který vypočítává očekávanou střední hodnotu logaritmu věrohodnostní funkce $\log P(\mathbf{X},\mathbf{Y}|\Theta)$ vzhledem k datům $\mathbf{Y}$, daným trénovacím datům X a současně odhadnutým parametrům $\Theta^{[k]}$, kde k je index iteračního kroku:

$$Q(\Theta|\Theta^{[k]}) = E\left(\log P(\mathbf{X},\mathbf{Y}|\Theta) \mid \mathbf{X},\Theta^{[k]}\right). \tag{4.20}$$

Protože $\mathbf{Y}$ je náhodná proměnná s rozložením $P(\mathbf{y}|\mathbf{X},\Theta^{[k]})$, můžeme vztah (4.20) zapsat následujícím způsobem:

$$Q(\Theta|\Theta^{[k]}) = \int_{y\in\psi} \log P(\mathbf{X},\mathbf{y}|\Theta)\log P(\mathbf{y},\mathbf{X}|\Theta^{[k]})\ d\mathbf{y}, \tag{4.21}$$

kde $\mathbf{y}$ je instance chybějících dat a ψ je prostor hodnot, kterých může $\mathbf{y}$ nabývat. Hodnotu $P(\mathbf{y},\mathbf{X}|\Theta^{[k]})$ potřebnou pro vztah (4.21) vypočteme jako:

$$P(\mathbf{y},\mathbf{X}|\Theta^{[k]}) = \prod_{t=1}^{T} P(\mathbf{y}_t|\mathbf{x}_t,\Theta^{[k]}). \tag{4.22}$$

Na základě známých parametrů $\Theta^{[k]}$ [1] dokážeme vypočítat $P(\mathbf{x}_t|\theta_m^{[k]})$. Mimo to můžeme interpretovat váhy jednotlivých mixtur c_m jako jejich apriorní pravděpodobnosti $c_m = P(m|\theta^{[k]})$, takže můžeme použít Bayesův vzorec pro výpočet

$$P(\mathbf{y}_t|\mathbf{x}_t,\Theta^{[k]}) = \frac{P(\mathbf{x}_t|\theta_{y_t}^{[k]})P(\mathbf{y}_t|\Theta^{[k]})}{P(\mathbf{x}_t|\Theta^{[k]})} = \frac{P(\mathbf{x}_t|\theta_{y_t}^{[k]})P(\mathbf{y}_t|\Theta^{[k]})}{\sum_{m=1}^{M} P(\mathbf{x}_t|\theta_m^{[k]})P(m|\Theta^{[k]})}. \tag{4.23}$$

V „Maximization" kroku se hodnoty získané v „Expectation" kroku maximalizují tak, aby platilo:

$$\Theta^{[k+1]} = \underset{\Theta}{ArgMax}\, Q(\Theta|\Theta^{[k]}). \tag{4.24}$$

Pro jednotlivé mixtury lze odvodit finální vztahy pro výpočty vah, středních hodnot a kovariančních matic:

$$c_m^{[k+1]} = \frac{1}{N}\sum_{t=1}^{T} P(m|x_t,\Theta^{[k]}), \tag{4.25}$$

$$\bar{x}_m^{[k+1]} = \frac{\sum_{t=1}^{T} \mathbf{x}_t P(m|x_t,\Theta^{[k]})}{\sum_{t=1}^{T} P(m|x_t,\Theta^{[k]})}, \tag{4.26}$$

[1] V nultém iteračním kroku (k=0) bereme za počáteční odhady střední hodnoty získané z VQ, kovarianční matice a váhy mixtur dopočítáme vzhledem k rozložení sekvence trénovacích příznakových vektorů.

$$\Sigma_m^{[k+1]} = \frac{\sum_{t=1}^{T} (\mathbf{x}_t - \bar{x}_m^{[k+1]})(\mathbf{x}_t - \bar{x}_m^{[k+1]})^T P(m|x_t, \Theta^{[k]})}{\sum_{t=1}^{T} P(m|x_t, \Theta^{[k]})}, \quad (4.27)$$

kde

$$P(m|x_t, \Theta^{[k]}) = \frac{P(x_t|\theta_m^{[k]}) P(m|\Theta^{[k]})}{\sum_{n=1}^{M} P(x_t|\theta_n^{[k]}) P(n|\Theta^{[k]})}. \quad (4.28)$$

EM algoritmus ukončíme až ve chvíli, kdy hodnota věrohodnosti přestane monotónně růst, nebo když dosáhneme předem stanovené konvergenční konstanty ε:

$$STOP \quad if \ (L(\mathbf{X}|\Theta^{[k-1]}) \geq L(\mathbf{X}|\Theta^{[k]}) \quad || \quad L(\mathbf{X}|\Theta^{[k]}) - L(\mathbf{X}|\Theta^{[k-1]}) < \varepsilon \ ). \quad (4.29)$$

4.4.4. Implementace

Mějme populační hustotu $P(\mathbf{X}|\Theta)$, kde Θ jsou parametry, které určují rozdělení. Naší snahou je modelování této hustoty určením parametrů Θ. To znamená, že p může být množina gaussovských funkcí a Θ budou jejich váhy, střední hodnoty a kovariance. Vzorek dat má velikost T, ℓ je věrohodnost parametrů daných daty. Předpokládejme vektory dat s rozdělením p. Chceme nalézt Θ takové, aby maximalizovalo ℓ, respektive $L = \log(\ell)$. A právě toto iterativně vykonává EM algoritmus. Pokud si shrneme poznatky o EM algoritmu, tak je to metoda umožňující nalezení nejpravděpodobnějšího odhadu parametrů Θ z dostupných, ale neúplných vzorků dat.

EM algoritmus začíná odhadem parametrů Θ_i, poté se iteračním způsobem hledají optimální parametry Θ. Pro iterace k a $k+1$ platí: $L(\mathbf{X}|\Theta^{[k+1]}) > L(\mathbf{X}|\Theta^{[k]})$. Pro konvergenci parametrů obecně postačuje pět až deset iterací, pak bývá splněna podmínka (4.29).

EM algoritmus pracuje následujícím způsobem :

Krok 1: *Inicializace*

Zvol počáteční odhad Θ. V naší implementaci EM algoritmu jsou k volbě počátečního odhadu použita kódová slova z vektorové kvantizace, to znamená LBG algoritmus. Ten dosahuje, podle našich poznatků, o něco lepší výsledky než běžně používaný *k*-means algoritmus.

Krok 2: „*Expectation*"

Vypočítej pomocnou Q-funkci $Q(\Theta \mid \Theta^{[k]})$, což je vlastně také „expectation" logaritmické věrohodnosti získané z kompletních dat. Tato funkce je závislá na parametrech Θ.

Krok 3: „*Maximalization*"

Vypočítej $\hat{\Theta} = ArgMax_{\Theta^{[k]}} Q(\Theta \mid \Theta^{[k]})$, tímto způsobem bude maximalizována pomocná funkce Q.

Krok 4: *Iterace*

$\Theta = \hat{\Theta}$, opakuj od **kroku 2** do konvergence.

Konvergence EM algoritmu je zaručena pouze k lokálnímu maximu věrohodnosti. To znamená, že nalezené parametry modelu závisí na počátečním odhadu parametrů. Prakticky by se dal tento problém řešit spuštěním EM algoritmu z více náhodně generovaných nebo jiným způsobem vybraných bodů a výběrem řešení s konečnou největší věrohodností. Velmi často se provádí normalizace této věrohodnosti vydělením počtem příznakových vektorů T [HUA01].

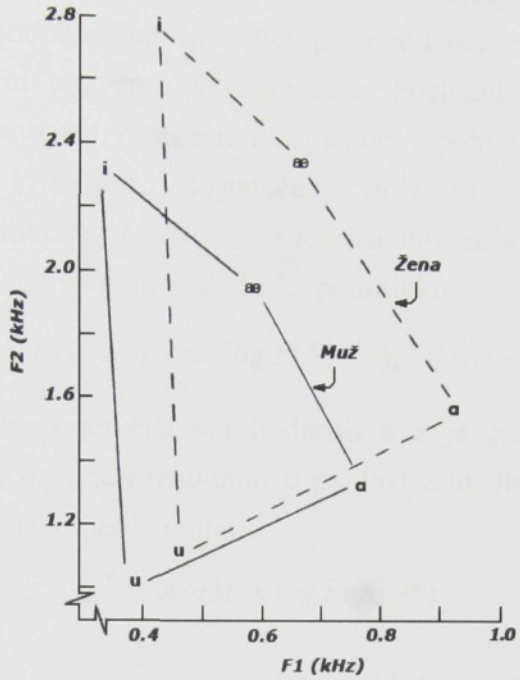
4.4.5. Identifikace pohlaví

Identifikace pohlaví (IP) se svou podstatou velmi podobá klasické identifikaci mluvčích nad uzavřenou množinou. Na rozdíl od IM, kde můžeme problém identifikace zapsat jako volbu 1:M, kde M je počet mluvčích v databázi, je problém identifikace pohlaví zredukován na 1:2. Proto také u IP budeme očekávat výrazně lepší výsledky než u IM.

V oblasti IP existují dvě hlavní cesty jak co nejlépe rozlišit muže od žen. V prvním případě se výzkumníci snaží o nalezení příznakových vektorů závislých na pohlaví, které mají pokud možno co nejlepší rozlišovací schopnost. Druhou možností je využití klasických přístupů z oblasti rozpoznávání řeči spojených například s MFCC kepstrálními příznaky. Jako příklad můžeme uvést neuronové sítě s klasickými příznaky [HAR01], speciální příznaky založené na spektrálních parametrech akustických vektorů [HAR03], HMM spolu s akustickými příznaky a hlasivkovou periodou [PAR96], nebo systém využívající lineární klasifikátor a fůze více znalostních systémů [SLO97].

Námi implementovaný způsob IP je úzce spjat se systémem pro IM, tj. pracuje na principu GMM. Přesněji řečeno, IP je provedena po IM pouhým zanalyzováním výsledků identifikace. Tento postup má pro nás jednu podstatnou výhodu a tou je časová nenáročnost takto provedené IP. Není tudíž třeba vypočítávat nová skóre nebo znovu parametrizovat

promluvu a vytvářet tak nové příznaky. IM je téměř vždy nutné provést na celé množině osob uložených v databázi (výjimku tvoří např. stromově prohledávané seznamy). Tím je provedeno M porovnání s modely mluvčích a na základě jednotlivých věrohodností je možné učinit rozhodnutí o pohlaví mluvčího.



Obr. 4.10: *Ukázka rozdílného rozložení formantů F1 a F2 u mužů a žen*

Nejjednodušší cestou, jak rozhodnout o pohlaví mluvčího, je přímo přes identifikovanou osobu (respektive její pohlaví). Pouze prohlásíme, že výsledkem identifikace pohlaví je pohlaví vítězného řečníka, tj.

$$g^* = \underset{m=1,\ldots,M}{ArgMax} \sum_{t=1}^{T} \log P(\mathbf{x}_t | \lambda^m), \tag{4.30}$$

kde g^* značí pohlaví vítězného mluvčího m, M je celkový počet mluvčích v databázi, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, kde $\mathbf{x}_t$ je vektor příznaků v diskrétním čase $t \in [1, 2, \ldots, T]$ a λ^m je model m-tého mluvčího. Tento postup nazvěme *metodou nejlepšího kandidáta*.

Dalším způsobem jak určovat pohlaví mluvčího je *metoda skupiny*, kde pohlaví testovaného řečníka reprezentujeme kohortou (skupinou řečníků), tj. vytváříme dvě množiny reprezentativních mluvčích. Výpočet obou věrohodností pohlaví můžeme zapsat následujícím vztahem

$$L(g) = \log P(\mathbf{X} | \lambda^g) = \log \left\{ \frac{1}{C(g)} \sum_{m \in K(g)} \prod_{t=1}^{T} P(\mathbf{x}_t | \lambda^m) \right\}, \tag{4.31}$$

kde $C(g)$ je počet mluvčích v kohortě $K(g)$ náležící k pohlaví g a λ^g reprezentuje model mužské či ženské kohorty, jejíž věrohodnost počítáme. Požadovanou kohortu je možné vytvořit podle různých pravidel. Je například možné roztřídit do kohort všechny identifikované mluvčí podle jejich pohlaví, tj. provést *tvrdé rozřazení všech mluvčích* v databázi do obou kohort. Další cesta, jak vytvořit kohortu, může vést přes stanovení pevného počtu nejpravděpodobnějších mluvčích pro obě kategorie, tj. $C(g) = C$ pro obě g. Takto vytvořenou kohortu nazvěme *n nejlepších*. Posledním v této práci uváděným způsobem jak vytvořit kohortu reprezentující jedno z pohlaví, je výběr n nejlepších mluvčích splňujících kritérium dostatečné blízkosti k identifikačnímu skóre nejpravděpodobnějšího mluvčího. Mluvčího m reprezentovaného modelem λ^m zařadíme do kohorty *n% nejlepších*, pokud splňuje následující podmínku:

$$\log P(\mathbf{X}|\lambda^m) > \alpha \cdot \log P(\mathbf{X}|\lambda^{g^*}), \quad 0 < \alpha < 1, \tag{4.32}$$

kde λ^{g^*} je model mluvčího s největší věrohodností a α je práh rozhodující o zařazení mluvčího m do kohorty. Konečné rozhodnutí o pohlaví mluvčího (volíme jednu možnost z množiny G) je pak provedeno podle vtahu

$$g^* = \underset{g \in G}{ArgMax} \log P(\mathbf{X}|\lambda^g). \tag{4.33}$$

Identifikaci můžeme samozřejmě provést i výpočtem věrohodnosti pro speciálně natrénovaný mužský a ženský model, tj. IP *metodou mužského a ženského modelu*. Princip, jakým by měla být zvolena trénovací data pro tento model, je blíže upřesněn v kapitole 4.5.2. Alternativou by mohlo být i vytvoření více modelů podle akustických podmínek a kvality přenosové cesty, které by se buď mohly sloučit, nebo používat samostatně. Tím bychom ovšem ztráceli výhodu napojení na modul provádějící IM.

Kvalitní IP mluvčích je pro náš systém přepisující zpravodajské pořady důležitá zejména z pohledu adaptace řečových modelů pro automatické rozpoznávání spojité řeči (Large Vocabulary Continuous Speech Recognition – LVCSR). Mnoho mluvčích (především z reportáží) se v audiozáznamech vyskytuje pouze jednou, tudíž nemají svůj model uložen v databázi a neexistují pro ně ani adaptační data. IM nad otevřenou množinou je ve většině případů označí jako neznámé řečníky a v tu chvíli je jim třeba adaptovat modely pro rozpoznávání řeči dle jejich pohlaví (řečový model je adaptován dle modelů kohorty mluvčích se shodným pohlavím). Pokud však informace o pohlaví řečníka není pravdivá, dochází ke znatelnému snížení kvality výsledného textového přepisu.

4.4.5.1. Identifikace pohlaví založená na LVCSR s GD modely

Odlišným způsobem je IP prováděna v modulu LVCSR (bližší informace o tomto modulu je možno nalézt v kapitole 2.1.7, případně v [NOU04], [NOU05a] a [NOU05b]).

Rozpoznávačem spojité řeči s dvěma různými nastaveními jsou testované promluvy přepsány do textové podoby. Jednou je použit mužský na pohlaví závislý (Gender Dependent – GD) model, podruhé ženský. V případě, kdy používáme LVCSR pro IP, nás nezajímá rozpoznávací skóre řečového přepisu, ale pouze výsledná pravděpodobnost vygenerování testované promluvy GD modelem.

Na myšlenku využít LVCSR s GD modely pro úlohu IP nás přivedly výsledky experimentů obdržených při rozpoznávání spojité řeči s adaptovanými modely. Jak již bylo předesláno v předchozích odstavcích, experimentálně jsme ověřili skutečnost, že pokud je při rozpoznávání spojité řeči použit nesprávný GD řečový model, dojde ve většině případů k signifikantnímu snížení rozpoznávacího skóre. To by však pro využití LVCSR pro IP nestačilo. V případě, že bychom chtěli postavit rozhodnutí o pohlaví řečníka na sníženém rozpoznávacím skóre, potřebovali bychom referenční textový přepis promluvy. Ten ovšem v reálných podmínkách chybí. Naše hypotéza byla postavena na předpokladu, že snížení rozpoznávacího skóre je přímo spjaté se snížením celkové vypočtené věrohodnosti, což prokázaly experimentální výsledky uvedené v kapitole 7.3. V případě věrohodnosti nepotřebujeme žádné referenční údaje, neboť referencí je nám samotná věrohodnost vypočtená pomocí druhého GD modelu.

Fůze výsledků s GMM IP

Pokud si výsledky IP z GMM a LVCSR navzájem odpovídají, mluvčímu je toto společné pohlaví přiřknuto. Pokud však nastane případ, že výsledky obou metod si navzájem protiřečí, je třeba výstupy obou sloučit a rozhodnout o pohlaví mluvčího pomocí nějakého kriteria. Pravděpodobnost, že promluva byla vyslovena mužem, pak je

$$P(m) = \beta_1 \cdot P(m|GMM) + \beta_2 \cdot P(m|ACSR), \tag{4.34}$$

kde $P(m|GMM)$ je pravděpodobnost vyřčení promluvy mužem vypočtená z výsledků IM, $P(m|ACSR)$ je pravděpodobnost vyřčení promluvy mužem vypočtená rozpoznávačem spojité řeči s GD modely a β_1, β_2 jsou váhové koeficienty stanovené v režimu trénování. Stejným postupem určíme i pravděpodobnost vyslovení promluvy ženou a oba výsledky porovnáme.

4.5. GMM-UBM

4.5.1. Stanovení verifikační míry v úloze VM

GMM je použito jako všeobecný pravděpodobnostní model pro vícerozměrné rozdělení hustoty pravděpodobnosti, které je schopné reprezentovat jakékoliv rozdělení hustoty pravděpodobnosti. Právě tato schopnost je velmi vhodná pro aplikace textově

nezávislé verifikace mluvčích. Jak již bylo uvedeno dříve, verifikace mluvčího je úloha, při níž ověřujeme, zda promluva reprezentovaná posloupností vektorů příznaků **X** byla opravdu vyslovena proklamovaným mluvčím S_v. Pokud chceme toto tvrzení ověřit, je třeba, abychom stanovili nějakou *verifikační míru* $V(\mathbf{X}, S_v)$, která nám dovolí porovnávat žadatele o ověření S_v reprezentovaného sekvencí příznakových vektorů **X** proti identitě reprezentované modelem λ_{hyp}. Konečné rozhodnutí učiníme na základě nějakého verifikačního prahu θ [RAD04]. Platí tedy:

$$\begin{aligned} &\text{mluvčí je akceptován jako } v, \text{ pokud} && V(\mathbf{X}, S_v) > \theta, \\ &\text{mluvčí není akceptován jako } v, \text{ pokud} && V(\mathbf{X}, S_v) < \theta, \\ &\text{není možné rozhodnout, pokud} && V(\mathbf{X}, S_v) = \theta. \end{aligned} \tag{4.35}$$

Jedním z možných způsobů, jak míru $V(\mathbf{X}, S_v)$ určit, je použití *normalizačního činitele odvozeného z aposteriorní pravděpodobnosti*. Stejně jako pro rozpoznávání řeči platí i pro rozpoznávání řečníka, že odhady parametrů rozložení jsou určovány před vlastním rozpoznáváním v režimu trénování (například metodou maximalizace věrohodnosti). Jakmile jsou tyto parametry jednou určeny, jmenovatel Bayesova vzorce

$$P(S_v|\mathbf{X}) = \frac{P(\mathbf{X}|S_v)P(S_v)}{\sum_{m=1}^{M} P(\mathbf{X}|S_m)P(S_m)} \tag{4.36}$$

je nezávislý na testovaných promluvách a může být v úloze identifikace mluvčího vyřazen. Ne však v úloze verifikace, tam je nutné tento člen vzít do úvahy, abychom podchytili i změny ve způsobu vyslovení promluvy, které mění jmenovatel výrazu (4.36). S_v ve vztahu reprezentuje testovaného (verifikovaného) mluvčího a $P(\mathbf{X}|S_m)$ je pravděpodobnost reprezentace **X** za předpokladu, že promluvu, ze které byla tato reprezentace získána, vyslovil mluvčí S_m. Je třeba pouze podotknout, že do množiny m spadá i mluvčí S_v. Míru $V(\mathbf{X}, S_v)$ je možno určit tak, že ji položíme rovnu $P(S_m|\mathbf{X})$. Bayesův vzorec se nám poté změní na

$$V(\mathbf{X}|S_v) = \frac{P(\mathbf{X}|S_v)P(S_v)}{\sum_{m=1}^{M} P(\mathbf{X}|S_m)P(S_m)}. \tag{4.37}$$

Za předpokladu, že rozložení apriorních pravděpodobností $P(S_m)$ bude rovnoměrné, zlogaritmováním vztahu (4.37) získáme

$$V(\mathbf{X}|S_v) = \log P(\mathbf{X}|S_v) - \log \sum_{m=1}^{M} P(\mathbf{X}|S_m). \tag{4.38}$$

Druhým způsobem, jak získat normalizační činitel a míru $V(\mathbf{X}, S_v)$, je přeformulování problému VM na testování dvou hypotéz:

$\boldsymbol{H_0}$: **X** je promluva od mluvčího S_v

a

$\boldsymbol{H_1}$: **X** není promluva od mluvčího S_v.

Optimální způsob, jak rozhodnout, která z daných hypotéz je správná, je tzv. *test poměru věrohodností* (Likelihood Ratio test – LR test). Zjišťování správnosti hypotéz je optimální pouze tehdy, pokud přesně známe pravděpodobnostní funkce. Bohužel v praxi je to téměř neřešitelný problém. LR test je definován takto:

$$\left.\begin{array}{l} \dfrac{P(\mathbf{X}|H_0)}{P(\mathbf{X}|H_1)} \\ \end{array}\right\} \begin{array}{ll} > \theta & \text{akceptování hypotézy } H_0, \\ \leq \theta & \text{zamítnutí hypotézy } H_0, \end{array} \tag{4.39}$$

kde $P(\mathbf{X}|H_i)$, $i = 0, 1$ je hustota pravděpodobnosti pro hypotézu H_i vypočtená pro sledovaný úsek řeči **X**. $P(A \mid B)$ bereme jako pravděpodobnost, pokud B je v této funkci nezávislá proměnná. Rozhodovací práh pro akceptování nebo zamítnutí H_0 je θ. Na obr. 4.12 jsou vyobrazeny dva páry věrohodnostních funkcí $P(\mathbf{X} \mid H_0)$ a $P(\mathbf{X} \mid H_1)$. Hlavním úkolem systému pro verifikaci mluvčího je stanovení technik pro přesné určení těchto dvou pravděpodobností, $P(\mathbf{X}|H_0)$ a $P(\mathbf{X}|H_1)$.

Z matematického hlediska je hypotéza H_0 reprezentována modelem označeným λ_{hyp}, který charakterizuje testovaného mluvčího S_v popsaného vektory příznaků **X**. Můžeme například říci, že gaussovské rozdělení nejlépe reprezentuje rozložení vektorů příznaků pro H_0. Pak λ_{hyp} je dána střední hodnotou a kovarianční maticí gaussovského rozdělení. Alternativní hypotéza H_1 je pak reprezentována modelem $\lambda_{\overline{hyp}}$.

Zatímco model pro reprezentaci H_0 je velmi dobře popsán a lze ho poměrně snadno získat z trénovacích promluv mluvčího m, model pro $\lambda_{\overline{hyp}}$ je vymezen daleko hůře, protože musí reprezentovat veškeré ostatní alternativy k danému mluvčímu. Pro modelování této druhé hypotézy však existuje několik přístupů.

Jak již bylo uvedeno, VM se od IM liší tím, že je určitou formou testování hypotézy. Při verifikaci budeme tedy testovat hypotézu, že mluvčí S je skutečně mluvčím S_v, pokud platí

$$P(S_v|\mathbf{X}) > P(\overline{S}_v|\mathbf{X}), \tag{4.40}$$

kde $\overline{S}_v$ reprezentuje množinu všech možných konkurenčních mluvčích a celá pravá část výrazu tak reprezentuje pravděpodobnost, že promluvu vyslovil jakýkoliv mluvčí kromě S_v.

Aby bylo možno ovlivnit, kdy mluvčího S akceptovat za mluvčího S_v, vstupuje do vztahu rozhodovací práh θ_v a tedy

$$\frac{P(S_v|\mathbf{X})}{P(\overline{S}_v|\mathbf{X})} > \theta_v, \tag{4.41}$$

kde θ_v>1. Chceme-li dosáhnout co možná nejlepších výsledků, měl by být rozhodovací práh pro každého mluvčího optimalizován v režimu trénování zvlášť. Dále můžeme zapsat

$$P(\overline{S}_v|\mathbf{X}) = P(S_1 \text{ nebo } S_2 \text{ nebo } \ldots \text{ nebo } S_{m\neq v}|\mathbf{X}) = \sum_{m\neq v} P(S_m|\mathbf{X}). \tag{4.42}$$

Pokud jsou případy S_m nezávislé (což je náš případ), dohromady úplné (což většinou splněno nebude, ale uvažovat to můžeme) a rozložení apriorních pravděpodobností $P(S_m)$ rovnoměrná, tak užitím vztahů (4.36) a (4.40) obdržíme

$$S = S_v \quad if \quad \frac{P(\mathbf{X}|S_v)}{P(\mathbf{X}|\overline{S}_v)} = \frac{P(\mathbf{X}|S_v)}{\sum_{m\neq v} P(\mathbf{X}|S_m)} > \theta_v, \tag{4.43}$$

což je vztah nazývaný *kritérium poměru věrohodností*. Suma přes m by měla obsahovat všechny možné mluvčí. Zlogaritmováním předchozí formule pak ve výsledku získáme

$$S = S_v \quad if \quad \log P(\mathbf{X}|S_v) - \log P(\mathbf{X}|\overline{S}_v) > \Delta_v, \tag{4.44}$$

kde $\Delta_v = \log\theta_v$.

Navažme na vztah (4.38), ve kterém je normalizační činitel roven $\log\sum P(\mathbf{X}|S_m)$ a kde sumu počítáme přes všechny mluvčí, kteří teoreticky mohou žádat o autorizaci včetně aktuálně verifikovaného mluvčího. Z (4.44) můžeme míru $V(\mathbf{X}, S_v)$ vyjádřit jako

$$V(\mathbf{X}|S_v) = \log P(\mathbf{X}|S_v) - \log P(\mathbf{X}|\overline{S}_v) \tag{4.45}$$

a *normalizační činitel odvozený z poměru věrohodností* je tak roven $\log P(X|\overline{S}_v)$. Je zřejmé, že správné určení normalizačního činitele spolu se správným určením verifikačního prahu mají velice výrazný vliv na celkovou úspěšnost verifikačního procesu. V případě poměru věrohodností je možno použít hned několik postupů. První z nich je založen na předpokladu, že reprezentativnost mluvčích, kteří jsou již zavedeni v databázi, je dostatečná k tomu, aby jejich řečové charakteristiky pokryly veškeré možné narušitele. Normalizační činitel tak může být vypočten následujícím způsobem

$$\log P(\mathbf{X}|\overline{S}_v) \approx \sum_{S_i\in R, i\neq v} \log P(\mathbf{X}|S_i), \tag{4.46}$$

kde R reprezentuje množinu mluvčích majících svůj model uložen v databázi. Pokud budeme předpokládat, že dominantní referenční mluvčí ze vztahu (4.46) má dostatečně

velkou pravděpodobnost vůči testované promluvě (v kapitole 4.6.4 bude zaveden pojem *kritická pravděpodobnost*, který této pravděpodobnosti odpovídá), lze napsat

$$\log P(\mathbf{X}|\overline{S}_v) \approx \max_{S_i \in R, i \neq v} \log P(\mathbf{X}|S_i). \tag{4.47}$$

Tyto postupy mají bohužel určité podstatné nevýhody:

- v obou případech bude nutné pro všechny referenční mluvčí vypočítat podmíněné pravděpodobnosti, což povede ke zvýšení výpočetní náročnosti;
- v případě vztahu (4.47) bude maximální podmíněná pravděpodobnost kolísat podle toho, jak blízko bude nejbližší referenční mluvčí od testovaného.

Alternativním řešením může být vhodná volba množiny blízkých referenčních mluvčích, ze kterých bude $P(\mathbf{X}|\overline{S}_v)$ vypočtena. Tato množina se v literatuře značí jako *kohorta* [RAD04]. Kohorta je většinou definována jako množina mluvčích, jejichž modely jsou dostatečně podobné modelu řečníka S_v. Tato skupina mluvčích by ale měla i dostatečně (pokud možno co nejlépe) pokrývat blízké okolí řečníka S_v [ROS92]. Pro každého mluvčího v databázi je v režimu trénování automaticky stanovena originální kohorta. Při každém přidání nového mluvčího do databáze by mělo následovat určení nové kohorty. Normalizační činitel je v případě kohorty možno zapsat jako

$$\log P(\mathbf{X}|\overline{S}_v) \approx \sum_{S_i \in R_v} \log P(\mathbf{X}|S_i), \tag{4.48}$$

kde R_v reprezentuje kohortu přiřazenou k mluvčímu S_v. Experimentálně bylo dokázáno, že tento způsob normalizace dokáže zvýšit separovatelnost mluvčích spolu se snížením citlivosti na určení verifikačního prahu. Mluvčí, jehož identita je ověřována, může být v některých případech taktéž zařazen v sumě[1] [MAT93]. Místo modelu ověřovaného mluvčího je také možné do kohorty zařadit náhodně vybrané mluvčí, jejichž hlasové charakteristiky se od řečníka S_v významně odlišují. V [REY94a] byl tímto způsobem prokázán stejný efekt, jako v případě zařazení S_v.

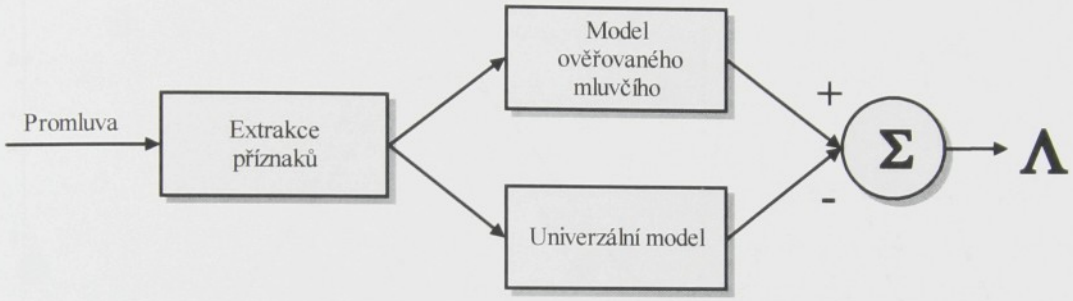
Dalším a v současnosti asi nejpoužívanějším přístupem, jak získat informace o $P(\mathbf{X}|H_1)$, je spojení promluv několika (co možná největšího a nejreprezentativnějšího počtu) mluvčích a provedení trénování jediného modelu. Tento model je pak v literatuře nazýván „*General Model*“, „*World Model*“ nebo „*Universal Background Model*“ (UBM) a můžeme ho označit λ_{UBM}. Je velmi důležité, aby počet promluv mluvčích byl co největší a aby tito mluvčí dobře reprezentovali osoby, se kterými se počítá při rozpoznávání. Velkou

[1] Někdy, pokud jsou řečové reprezentace ověřovaného řečníka a ostatních mluvčích od sebe velmi vzdáleny jak tomu bývá například u opačných pohlaví, to může vést k přesnějším odhadům a výsledkům.

výhodou tohoto přístupu je fakt, že po pečlivém natrénování univerzálního modelu je možno tento model použít pro všechny rozpoznávané mluvčí.

$$\log P(\mathbf{X}|\overline{S}_v) \approx \log P(\mathbf{X}|\lambda_{UBM}). \tag{4.49}$$

Na obr. 4.11 jsou zobrazeny základní komponenty systému pro verifikaci mluvčích, který pracuje s hodnotami věrohodnosti. Výstupem členu extrahujícího příznaky je sekvence vektorů příznaků, které reprezentují danou promluvu (mohou to být opět MFCC kepstrální příznaky) $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, kde $\mathbf{x}_t$ je vektor příznaků v diskrétním čase $t \in [1, 2, \ldots, T]$. Tyto vektory příznaků jsou použity pro výpočet pravděpodobností hypotéz H_0 a H_1 podle vztahů (4.45) a (4.49).



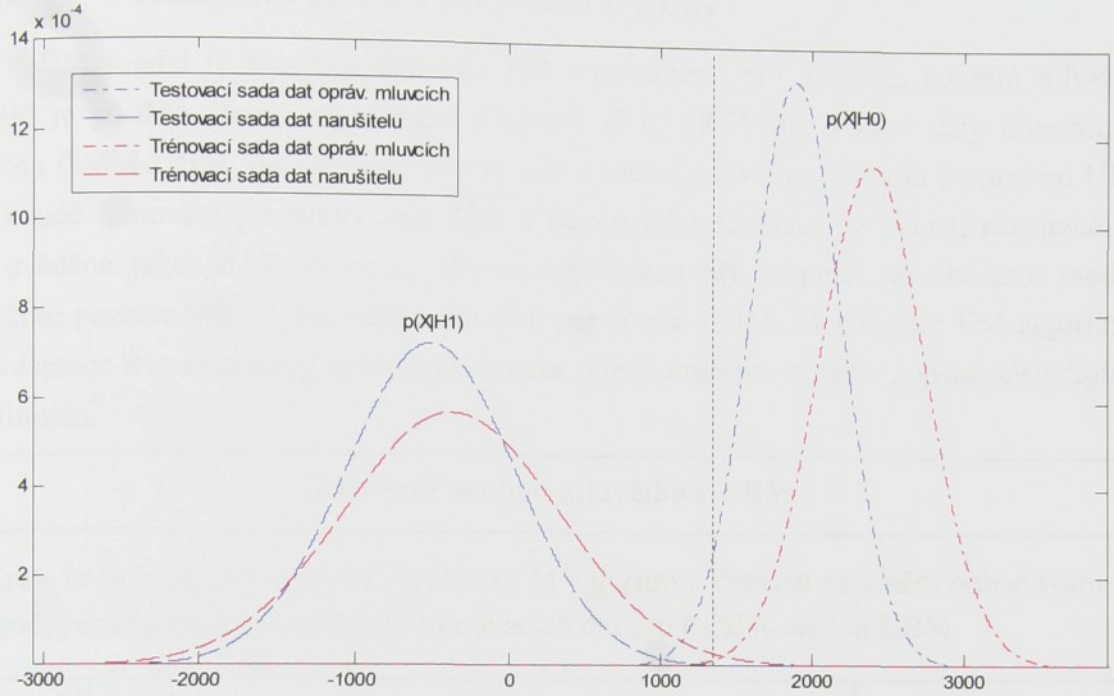
Obr. 4.11: *Ukázka, jak může vypadat GMM systém pro verifikaci mluvčích. Tento konkrétní příklad je založen na reprezentaci hypotézy H_1 modelem UBM.*

Reprezentace hypotézy H_1 založené na UBM je použita i v našem systému pro VM, a proto bude princip tohoto přístupu vysvětlen podrobněji v následujících odstavcích.

4.5.2. Tvorba UBM modelu

Z kapitoly 4.5.1 vyplývají požadavky, které UBM model musí splňovat, pokud má kvalitně reprezentovat populaci všech možných narušitelů. Tím je myšleno nejen rozložení řečových charakteristik narušitelů v prostoru kolem verifikovaného řečníka, ale i kvalita řeči jako taková. V některých úlohách je dostupná apriorní informace o kvalitě přenosové cesty, nahrávacím zařízení či pohlaví mluvčího. V takovém případě je logické vytvářet UBM model z kvalitativně příbuzných nahrávek, čímž se bezesporu sníží vliv těchto faktorů na úspěšnost verifikace [REY97], [CAR91]. Zmíněné atributy promluvy negativně ovlivňují především verifikační míru $V(\mathbf{X}, S_v)$, a tím i umístění verifikačního prahu θ_v. Příklad na obr. 4.12 dokládá, jak je důležité správně určit verifikační práh reprezentovaný svislou čerchovanou čarou. Tento práh byl stanoven v bodě blížícímu se *EER* (Equal Error Rate – viz kapitola 4.6.2) na základě rozložení trénovacích dat. Jak je z obrázku zřejmé, pozice a tvar gaussovských křivek se může v režimu testování změnit. Verifikační práh nyní zamítá velké množství oprávněných mluvčích. Tento příklad ilustruje potřebu

normalizace skóre nebo alespoň rozdělení dat na trénovací data, vývojová testovací data a evaluační testovací data. Přidáním vývojových testovacích dat je možné alespoň v hrubých rysech zjistit možnou změnu verifikačního prahu vůči trénovacím datům.



Obr. 4.12: *Příklad dvou párů věrohodnostních funkcí $P(\mathbf{X}|H_0)$ a $P(\mathbf{X}|H_1)$. Na horizontální ose je vynesena veličina* **X**, *na vertikální ose pak* $P(\mathbf{X})$.

Bohužel v případě IM nad otevřenou množinou z oblasti zpravodajských pořadů není jakákoliv z výše zmíněných apriorních informací dostupná. Proto je nutné pro tuto úlohu zvolit pro trénování UBM co možná nejobecnější data z celého spektra možností.

Vytvoření UBM modelu je možné hned několika způsoby. Pro nás bylo nejjednodušším řešením použití EM algoritmu vzhledem k jeho předešlé implementaci v úloze IM. Pokud máme připravena trénovací data, je postup naprosto shodný s vytvářením modelů mluvčích pro IM. Pokud hodláme natrénovat celý UBM model z jedné množiny nahrávek, je třeba ohlídat námi očekávané rozložení dat (například pohlaví) a zamezit tak dominanci, která by mohla být v konečném důsledku příčinou vzniku chyb. V nejhorším případě pak budeme verifikovat přenosový kanál, nahrávací hardware či pohlaví mluvčího a ne jeho identitu. Druhou možností jak vytvořit UBM je spojit několika separátně vytvořených modelů do jednoho. Tímto způsobem lze získat kontrolu nad přítomností všech skupin, které chceme mít v modelu obsaženy.

Náš systém pro VM používá druhou zde zmíněnou variantu a kombinuje separátně vytvořený mužský a ženský model podle vztahu

$$P(\mathbf{X}|\lambda_{UBM}) = \{P(\mathbf{X}|\lambda_{MALE}), P(\mathbf{X}|\lambda_{FEMALE})\}. \tag{4.50}$$

Více o postupech tvorby modelu UBM se lze dočíst například v [ISO99].

4.5.3. Adaptace modelu mluvčího z UBM

UBM model je klasicky trénován EM algoritmem, pro adaptaci modelů mluvčích z UBM je použita forma bayesovské adaptace (BA) [REY00]. Pokud tedy hovoříme o systému GMM-UBM, odvozujeme model testovaného mluvčího úpravou parametrů UBM na základě trénovací promluvy mluvčího a bayesovské adaptace (v oblasti rozpoznávání řeči uváděna jako MAP analýza). Hlavní myšlenkou při adaptaci je odvození modelu mluvčího pozměněním dobře natrénovaných parametrů UBM. Stejně jako EM algoritmus je i adaptace dvoustupňový analytický proces, který můžeme zhruba popsat následujícím algoritmem.

Odvození modelu mluvčího z UBM
Krok 1: Je identický s prvním krokem EM algoritmu. Provádí se v něm odhadování vhodných statistických přehledů trénovacích dat pro každou složku UBM.
Krok 2: Na rozdíl od EM algoritmu jsou nové statistické odhady zkombinovány se starými odhady získanými z UBM. K tomu jsou použity směšovací koeficienty, které jsou závislé na datech.

Postup při adaptaci je dle [REY00] graficky naznačen na obr. 4.13. Nejprve je určeno pravděpodobnostní uspořádání trénovacích vektorů příznaků vůči komponentám UBM (obr. 4.13a). S ohledem na toto uspořádání se určí váhové koeficienty c_m, střední hodnoty $\bar{\mathbf{x}}_m$ a kovarianční matice Σ_m všech mixtur. To je vlastně stejný postup jako u „Expectation" kroku EM algoritmu. Nakonec se tyto nové parametry použijí pro aktualizaci statistických údajů starého UBM (viz obr. 4.13b).

Adaptační proces vycházející z [BIM04] je podrobně specifikován v následujících odstavcích. Nejprve zjistíme pravděpodobnostní uspořádání trénovacích vektorů $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ vůči jednotlivým složkám UBM. Pro m-tou mixturu UBM pak platí

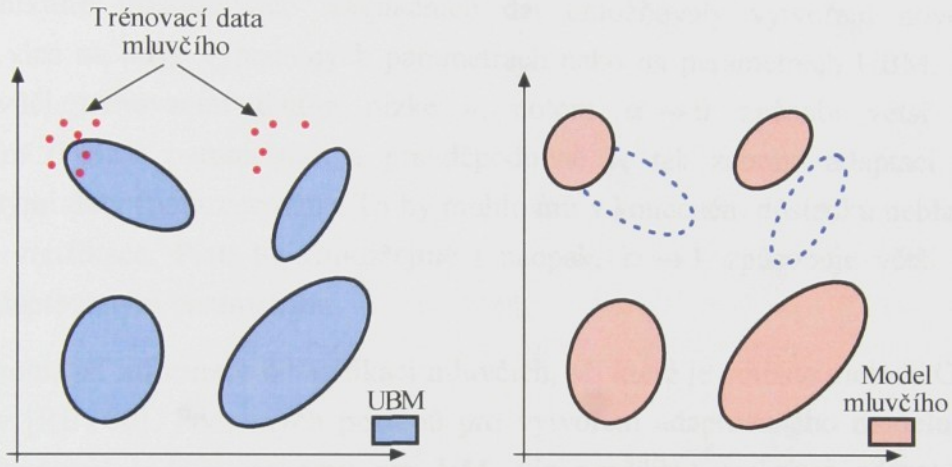
$$P(m|\mathbf{x}_t) = \frac{c_m p_m(\mathbf{x}_t)}{\sum_{i=1}^{M} c_i p_i(\mathbf{x}_t)}, \tag{4.51}$$

kde c_m je váha m-té mixtury a M je počet všech mixtur UBM. Použitím vztahu (4.51) určíme

$$n = \sum_{t=1}^{T} P(m \mid \mathbf{x}_t), \tag{4.52}$$

což je pravděpodobnostní počet všech trénovacích vektorů náležících k m-té mixtuře s prvotními parametry $(c_p, \bar{x}_p, \Sigma_p)$. Jelikož všechny následující adaptační výpočty budou prováděny s m-tou mixturou, můžeme si dovolit, z důvodu přehlednosti, index m v dalších vztazích neuvádět. Adaptační koeficient α pro tuto mixturu vyjádříme jako

$$\alpha = \frac{n}{n+r}, \tag{4.53}$$



Obr. 4.13: *Grafické znázornění dvou kroků, které je třeba učinit při transformaci UBM na model testovaného mluvčího*

kde r je fixní relevanční faktor (stanovuje se experimentálně a souvisí s počtem mixtur). Adaptovaná váha mixtury je pak rovna

$$c_\alpha = [\alpha c_s + (1-\alpha) c_p]\gamma, \tag{4.54}$$

kde $c_s = n/T$ je váha mixtury pro novou sekvenci příznakových vektorů a T je počet těchto trénovacích vektorů. Rozsahový faktor γ je vypočten přes všechny adaptované váhy mixtur, abychom zachovali podmínku $\sum_{m=1}^{M} c_{\alpha m} = 1$. Adaptovanou střední hodnotu vypočteme na základě vztahu

$$\bar{x}_\alpha = \alpha \bar{x}_s + (1-\alpha)\bar{x}_p, \tag{4.55}$$

kde

$$\bar{x}_s = \frac{1}{n}\sum_{t-1}^{T} P(m \mid \mathbf{x}_t)\mathbf{x}_t \tag{4.56}$$

je střední hodnota mixtury pro nové trénovací vektory. Nakonec je provedena i adaptace kovarianční matice

$$\Sigma_\alpha = \alpha E\{\mathbf{x}_t^2\} + (1-\alpha)(\Sigma_p + \bar{x}_p^2) - \bar{x}_\alpha^2, \tag{4.57}$$

kde

$$E\{\mathbf{x}_t^2\} = \frac{1}{n}\sum_{t=1}^{T} P(m|\mathbf{x}_t)\mathbf{x}_t^2 \tag{4.58}$$

je očekávaná druhá mocnina příznakových vektorů adaptovaného mluvčího a $\mathbf{x}_t^2 = diag(\mathbf{x}_t \cdot \mathbf{x}_t')$.

Na datech závislé adaptační koeficienty jsou navrženy takovým způsobem, aby dle počtu k mixtuře příslušejících adaptačních dat umožňovaly vytvoření nové mixtury závisející více na nově vypočtených parametrech nebo na parametrech UBM. Pokud má mixtura vůči trénovacím datům nízké n, potom $\alpha \to 0$ způsobí větší přiklonění k původním (UBM) parametrům a pravděpodobně se tak zabrání adaptaci statisticky neprůkaznými daty (podtrénování). To by mohlo mít v konečném důsledku neblahý vliv na úspěšnost verifikace. Platí to samozřejmě i naopak, $\alpha \to 1$ způsobuje větší přiklonění k nově zadaptovaným parametrům.

Podrobnější informace o verifikaci mluvčích, při které je použita metoda GMM, jsou uvedeny v [REY00]. Podobných postupů pro vytvoření adaptovaného modelu mluvčího z UBM využívají i další systémy pro VM, viz například real-time adaptace modelů mluvčích z UBM [PON04] nebo použití klastrově závislého modelu pozadí [TIN03].

Tento postup adaptace parametrů GMM modelu verifikovaného mluvčího z UBM má jednu významnou výhodu. Pokud si při adaptaci zapamatujeme, které složky modelu mluvčího byly adaptovány a které byly ponechány nezměněny (tj. stejné jako v UBM viz obr. 4.13b), můžeme při testování provádět výpočet pravděpodobností pouze z těchto upravených složek modelu. To může mít za následek podstatné zkrácení času potřebného pro výpočet verifikačního skóre. Výhodou použití metody MAP v porovnání s dalšími adaptačními postupy je také skutečnost, že díky svému principu s rostoucím množstvím adaptačních dat konverguje, i když ne příliš rychle, k teoreticky nejlepšímu adaptovanému modelu mluvčího.

4.5.4. Normalizace skóre

Hledání nejlepšího verifikačního prahu pro konkrétní systém VM je samo o sobě nesnadný úkol. Velikost skóre, jež je se stanoveným verifikačním prahem porovnávána, je navíc velmi citlivá na různé vnější vlivy. Především to může být různý zdroj nahrávek jednotlivých mluvčích. Další rozdíly mohou být způsobeny lišícím se audiopozadím nahrávek (hluk, podbarvení hudbou apod.), obsahem vyřčeného textu, způsobem promluvy a délkou dat ovlivňující trénování modelů mluvčích. Neméně podstatným problémem je

pak i rozdílnost trénovacích a testovacích dat v rámci promluv jednoho mluvčího. Tato citlivost způsobuje širokou variabilitu výsledného skóre a určení hodnoty prahu je tudíž velmi obtížné. Abychom tyto negativní vlivy co nejvíce eliminovali, jsou v úloze IM nad otevřenou množinou a v úloze VM používány normalizační techniky.

Použití normalizace v úlohách IM a VM vychází z [LIK88], kde autoři prokázali velké kolísání rozložení skóre jak u oprávněných mluvčích tak u narušitelů. Na základě těchto pozorování autoři navrhli řešení, které se nazývá *standardizace rozložení pravděpodobnosti skóre*. Principem metody je transformace *rozložení skóre narušitelů* $L_{S_v}(\mathbf{X})$ do standardní formy [BIM04]. Tím by se měl eliminovat vliv podmínek při pořízení testovací nahrávky. Normalizované skóre $\hat{L}_{S_v}(\mathbf{X})$ tedy získáme ze vztahu

$$\hat{L}_{S_v}(\mathbf{X}) = \frac{L_{S_v}(\mathbf{X}) - \mu_{S_v}}{\sigma_{S_v}}, \tag{4.59}$$

kde μ_{S_v} a σ_{S_v} jsou normalizované parametry mluvčího S_v. Volba normalizovat rozložení skóre narušitelů vychází z faktu, že realizací nahrávek narušitelů máme k dispozici mnohem více než realizací oprávněných mluvčí. To je dáno i tím, že narušitele můžeme snadno vytvořit tak, že provedeme verifikaci promluvy proti několika modelům.

Připomeňme, že použití UBM nebo kohorty mluvčích je také jistou formou normalizace. Dále tedy budeme předpokládat, že níže popsané normalizační techniky jsou aplikovány na skóre

$$\widetilde{L}_{S_v}(\mathbf{X}) = \frac{L_{S_v}(\mathbf{X})}{L_{\bar{S}_v}(\mathbf{X})} \tag{4.60}$$

a tedy $L_{S_v}(\mathbf{X}) = \widetilde{L}_{S_v}(\mathbf{X})$.

Z-normalizace (*Znorm*) je označována jako na mluvčích závislá metoda normalizace. Je to dáno tím, že střední hodnotu a směrodatnou odchylku potřebnou pro normalizaci získáme z rozložení skóre vzniklého testováním modelu mluvčího proti nahrávkám od skupiny narušitelů. Variantou Z-normalizace je i *Hnorm*, jež je určena k potlačování změn vyvolaných různými typy telefonních sluchátek. Normalizační parametry jsou stanovovány testováním modelů mluvčích vůči promluvám narušitelů pořízených různými druhy sluchátek. Tímto způsobem jsou mluvčímu stanoveny pro jednotlivá zařízení normalizační parametry, které se v režimu testování volí na základě informace přicházející společně s testovanou promluvou. Tuto normalizační metodu však lze použít pouze v případech, kdy jsou známy podmínky nahrávání testovací promluvy **X**. Naopak výhodou obou metod je stanovení normalizačních parametrů v režimu trénování, tj. tato normalizační metoda nezvyšuje výpočetní náročnost vlastní verifikace.

T-normalizace (*Tnorm*) se od Znorm normalizace liší použitím modelů narušitelů namísto jejich nahrávek. Tato metoda je proto někdy označována jako na testech závislá normalizační technika. Normalizační parametry jsou navíc oproti Znorm stanoveny až v režimu testování, což můžeme považovat za drobný nedostatek, neboť je třeba testovat promluvu vůči dalším modelům narušitelů pro zjištění rozložení skóre pro normalizaci. *HTnorm* je podskupinou T-normalizace a stejně jako Hnorm normalizuje skóre v případě telefonních nahrávek. Normalizační parametry pro daný přístroj stanovujeme testováním promluvy vůči skupinám na typu přístroje závislých narušitelů. Dále je možné využít i dalších normalizačních technik jako například *Cnorm*, *Dnorm* případně *WMAP*, jejichž shrnující popis lze nalézt v [BIM04].

4.6. Experimentální porovnání metod IM, IP a VM

V této kapitole jsou uvedeny některé zajímavé výsledky přípravných experimentů nesouvisejících s identifikací audiosegmentů jako takových, ale spíše s experimentálním ověřením funkčnosti implementovaných metod a se zjištěním a nastavením optimálních parametrů pro jednotlivé úlohy. V druhé části je krátce představen real-time systém pro identifikaci a verifikaci mluvčích, jenž vznikl v průběhu vývoje jako funkční aplikace ověřující chování rozpoznávacího softwaru v reálných podmínkách.

Před zveřejněním vlastních výsledků je nejprve třeba ujasnit si způsob, jakým se experimentální výsledky v oblasti IM, IP a VM vyhodnocují.

4.6.1. Vyhodnocování výsledků systému pro IM

Úspěšnost systému pro identifikaci mluvčích je třeba rozlišit dle typu úlohy na identifikaci nad otevřenou či uzavřenou množinou viz kapitola 4.1.2. Identifikace nad uzavřenou množinou je z hlediska vyhodnocování jednodušší a metodika je použitelná i pro vyhodnocování úspěšnosti identifikace pohlaví. Při IM nad uzavřenou množinou totiž existují pouze dvě možnosti identifikace – správná a chybná. V prvním případě je mluvčí identifikován správně a v druhém je totožnost mluvčího určena chybně. Rozdělíme-li si každou nahrávku na nepřekrývající se mikrosegmenty o určité velikosti (v našem případě byla délka mikrosegmentu stanovena na 100ms), můžeme pak počítat s mikrosegmenty jako s jednotlivými testy. Toto rozdělení na menší části je u BN nahrávek velmi důležité, neboť pracujeme s velmi proměnlivými délkami promluv a při vyhodnocování potřebujeme míry, které nám říkají skutečnou délku dobře či špatně rozpoznaných promluv. Kvalitu systému pro IM lze tedy ohodnotit *mírou neúspěšných identifikací*

$$R_{ER} = \frac{n_{err}}{n_{tests}} \times 100\%, \tag{4.61}$$

kde n_{err} značí počet chybně identifikovaných mikrosegmentů a n_{tests} je počet všech mikrosegmentů z nichž se skládá testovací množina. Druhým způsobem jak ohodnotit kvalitu identifikačního systému může být *míra úspěšných identifikací* reprezentovaná vztahem

$$R_{SUC} = \frac{n_{suc}}{n_{tests}} \times 100\%, \tag{4.62}$$

kde ekvivalentně k předchozímu vztahu značí n_{suc} počet správně identifikovaných mikrosegmentů. Je zřejmé, že pro vztahy R_{ER} a R_{SUC} platí

$$R_{ER} = 100\% - R_{SUC}. \tag{4.63}$$

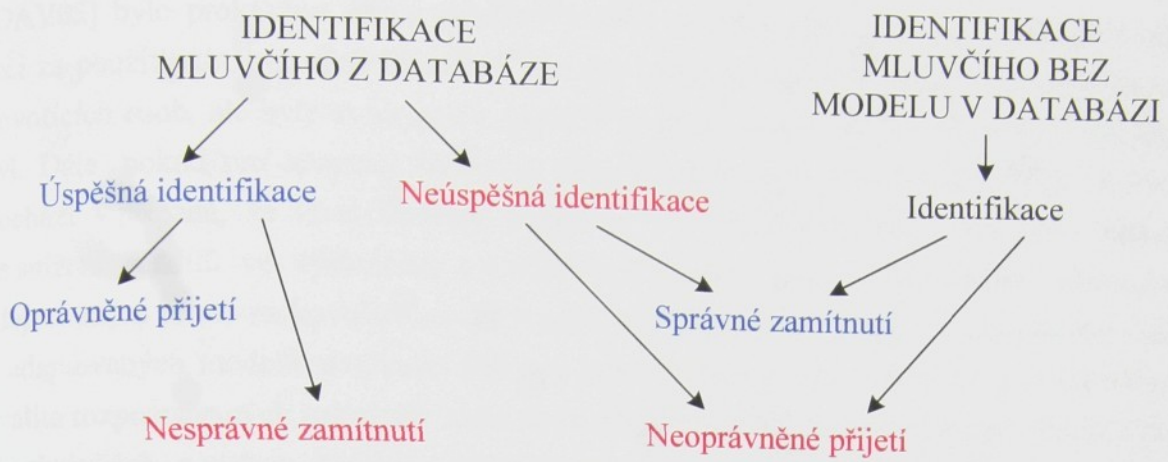
Vyhodnocení identifikace nad otevřenou množinou je v porovnání s předchozím případem o něco složitější. Důvodem je propojení metod IM nad uzavřenou množinou a VM. V praxi je postup identifikace následující: nejprve je provedena IM, na jejímž základě je neznámému řečníkovi přiřazen nejpravděpodobnější mluvčí z databáze (model hlasové reprezentace tohoto mluvčího), a v druhém kroku je provedena verifikace promluvy neznámého řečníka proti tomuto modelu. Až po provedené verifikaci je rozhodnuto o identitě mluvčího. Pokud je neznámý mluvčí akceptován, prohlásíme jej za mluvčího, proti jehož modelu byl verifikován. Pokud je při verifikaci tato hypotéza zamítnuta, je o neznámém mluvčím prohlášeno, že není nikým z databáze použité pro IM. Proces IM nad otevřenou množinou lze tedy zapsat vztahem

$$\max_{1\le m\le M}\{P(\mathbf{X}\mid\lambda_m)\} \begin{matrix} > \\ < \end{matrix} \theta \rightarrow \mathbf{X} \in \begin{Bmatrix} \lambda_i, i = \arg\max_{1\le m\le M}\{P(\mathbf{X}\mid\lambda_m)\} \\ \text{model neznámé osoby} \end{Bmatrix}, \tag{4.64}$$

kde $\mathbf{X}$ je promluva reprezentovaná sekvencí příznakových vektorů, M je počet mluvčích uložených v databázi, kteří jsou reprezentováni modely $\lambda_1, \lambda_2, \ldots, \lambda_M$ a θ je verifikační práh.

Úspěšnost systému pro IM nad otevřenou množinou lze vyjádřit hned několika způsoby. Graficky je to možné pomocí DET a ROC křivek (tyto dva pojmy budou vysvětleny v kapitole 4.6.2), numericky pak *poměrným počtem nesprávných zamítnutí*, *neoprávněných přijetí* a *mírou neúspěšnosti identifikace* pro daný verifikační práh [GIB97].

Z obrázku 4.14 vyplývá, že bezchybnou identifikací nad otevřenou množinou je pouze oprávněné přijetí úspěšně identifikovaného mluvčího (mluvčí má svou identitu zanesenou v databázi) a správné zamítnutí mluvčího, který nenáleží do množiny osob s uloženou hlasovou reprezentací v databázi.



Obr. 4.14: *Naznačení možných výsledků při identifikaci nad otevřenou množinou*

V našem systému pro přepis audiozáznamů však výsledek identifikace slouží nejen pro označení (zobrazení) identity mluvící osoby, ale výsledek identifikace je dále využit v LVCSR modulu, kde jsou na základě identifikované osoby, jejího pohlaví, případně pořadí dalších pravděpodobných mluvčích, voleny a vytvářeny adaptované řečové modely právě pro konkrétní promlouvající osobu. V takovém případě pro nás není důležité vyhodnocovat poměrný počet nesprávných zamítnutí či míru neúspěšnosti identifikace, spíše potřebujeme znát, v kolika případech systém provedl chybnou identifikaci. Proto zavádíme míru *poměrný počet chyb identifikace nad otevřenou množinou*

$$R_{OI} = \frac{\sum_{i=1}^{n_{tests}} err_i}{n_{tests}} \times 100\%, \tag{4.65}$$

kde n_{tests} je počet všech mikrosegmentů s nimiž byly prováděny testy a err_i je počet mikrosegmentů při i-tém identifikačním pokusu, jenž může nabývat hodnoty počtu mikrosegmentů nebo 0 na základě formule

$$if\ (\ FI \parallel FR \parallel FA_1 \parallel FA_2\ ) \begin{cases} true & err_i = \text{počet mikrosegmentů} \\ false & err_i = 0 \end{cases} \tag{4.66}$$

kde FI značí neúspěšnou identifikaci, FR nesprávné zamítnutí správně identifikovaného mluvčího, FA_1 neoprávněné přijetí chybně identifikovaného mluvčího majícího model v databázi a FA_2, jež reprezentuje neoprávněné přijetí mluvčího, který nepatří do referenční skupiny osob s modelem uloženým v databázi. Protože chyby FA_1 je možné dosáhnout pouze v případě, že je dosaženo i chyby FR, můžeme podmínku *if* zjednodušit na $if\ (\ FI \parallel FR \parallel FA_2\ )$.

S ohledem na uspořádání našeho systému pro přepis zpravodajských pořadů bylo třeba přidat ještě jednu míru, která zohledňuje specifické chování našeho systému. V článku

[DAV05] bylo prokázáno určité zlepšení rozpoznávacího skóre při rozpoznávání spojité řeči za použití adaptovaných řečových modelů, které sice neodpovídaly skutečné identitě hovořících osob, ale byly zvoleny dle identity nejpravděpodobnějšího mluvčího z procesu IM. Dále, pokud pro adaptaci řečového modelu využíváme skupinu mluvčích (k tomu dochází v případě, že identifikovaná osoba je ve verifikační fázi zamítnuta), dochází ke snížení rozdílů ve výsledném rozpoznávacím skóre proti rozpoznávání s modelem adaptovaným přímo na hovořící osobu. V mnoha případech je výsledné rozpoznávací skóre u adaptovaných modelů skupinou dokonce lepší. Příčinou tohoto jevu je pravděpodobně kvalita rozpoznávaných nahrávek, která je ve většině případů nesprávného zamítnutí nízká. V nahrávkách s nízkou kvalitou bývají řečové charakteristiky mluvčího zkresleny a v takovém případě může být skupina mluvčích pro adaptaci řečového modelu výhodnější. Pokud chceme tuto skutečnost zohlednit, je třeba zavést druhou míru k R_{OI}. Tuto veličinu nazvěme *poměrný počet chyb identifikace neoprávněným přijetím*

$$R_{OIFA} = \frac{\sum_{i=1}^{n_{tests}} FA_1 + \sum_{i=1}^{n_{tests}} FA_2}{n_{tests}} \times 100\%. \tag{4.67}$$

4.6.2. Vyhodnocování výsledků systému pro VM

Chybovost systému pro verifikaci řečníka hodnotíme podle počtu nesprávně vpuštěných narušitelů, respektive podle počtu oprávněných uživatelů, kteří jsou systémem neoprávněně zamítnuti jako narušitelé. V případě *poměrného počtu chyb neoprávněného přijetí* je pro výpočet využit vztah

$$R_{FA}(\theta) = \frac{n_{FA}(\theta)}{n_{impostors}} \times 100\%, \tag{4.68}$$

kde $n_{FA}(\theta)$ je počet mikrosegmentů neoprávněně akceptovaných narušitelů vpuštěných systémem na základě prahu θ a $n_{impostors}$ je počet mikrosegmentů všech pokusů o neoprávněnou autorizaci. Ekvivalentně s tímto výrazem je vypočten i *poměrný počet nesprávných zamítnutí mluvčích*, kteří měli být systémem označeni jako oprávnění uživatelé

$$R_{FR}(\theta) = \frac{n_{FR}(\theta)}{n_{authorized}} \times 100\%, \tag{4.69}$$

kde $n_{FR}(\theta)$ je počet mikrosegmentů neoprávněně zamítnutých oprávněných mluvčích a $n_{authorized}$ je počet všech pokusů o autorizaci oprávněnými mluvčími. Další mírou, kterou systém ověřující totožnost můžeme popsat, je *EER* udávající poměrný počet chyb systému s rozhodovacím prahem θ_{EER}, který je nastaven tak, aby bylo dosaženo rovnosti $R_{FA}(\theta)$ a

$R_{FR}(\theta)$. To je samozřejmě vhodnější pro automatické vyhodnocování než kombinace dvou hodnot měnících se s nastavením verifikačního prahu. Právě z těchto důvodů je asi nejnázornějším ohodnocením úspěšnosti verifikace grafická závislost $R_{FA}(\theta)$ na $R_{FR}(\theta)$. Spíše než *ROC* (Receiver Operating Characteristic) křivky jsou pro jejich přehlednost používány *DET* (Detection Error Tradeoff) křivky. Problémem ROC je totiž jejich lineární uspořádání. Pokud bychom vynášeli $R_{FA}(\theta)$ na $R_{FR}(\theta)$ do ROC grafu, okrajové části budou velice nahuštěné a rozumné rozlišení dostaneme pouze v okolí levého dolního rohu grafu, kde leží optimální bod verifikačního systému (čím více se tomuto bodu přiblížíme, tím bude verifikační systém dosahovat lepších výsledků). Protože do DET grafů vynášíme hodnoty kvantilů normovaného gaussovského rozložení (ty odpovídají poměrnému počtu chyb), jsou měřítka os vyjadřující $R_{FA}(\theta)$ a $R_{FR}(\theta)$ nelineární. Pokud má vynášená křivka přibližný tvar přímky, máme kontrolu, že rozložení verifikační míry je normální. Pokud propojíme levý spodní roh a pravý horní roh grafu pomyslnou přímkou, pak průsečíky této přímky s vynesenými DET křivkami znázorňují hodnotu *EER*.

4.6.3. Výsledky vývojových experimentů

Vývojová databáze

Pro tuto sadu experimentů sloužících k nalezení optimálního nastavení modulů pro rozpoznávání mluvčího byla použita speciální vývojová databáze čítající 4070 promluv. Tato data jsou podmnožinou databáze sloužící k experimentům s rozpoznáváním řeči v češtině. Proto také všichni mluvčí vyslovovali stejné věty. Jednotlivé promluvy byly navíc vybrány tak, aby byly pokud možno co nejbohatší na množství obsažených fonémů. Databáze se skládala z celkového počtu 55 mluvčích – 29 mužů a 26 žen z různých věkových skupin. Každý z mluvčích pronesl celkem 74 vět. Poměr testovacích vůči trénovacím datům se v průběhu experimentů bude lišit, proto budou tyto hodnoty uváděny vždy přímo pro daný test.

Mluvčí byli nahráváni v průběhu jednoho či dvou sezení, ale doba mezi sezeními nebyla nikdy delší než jeden měsíc. Nahrávky byly opět čisté, pořízené v laboratorních podmínkách bez vnějšího akustického rušení. Promluvy byly uloženy v souboru formátu „wav" s šestnáctibitovým kódováním, vzorkovací frekvence 8 kHz a jeden mono kanál jsou v této oblasti jedním ze standardů.

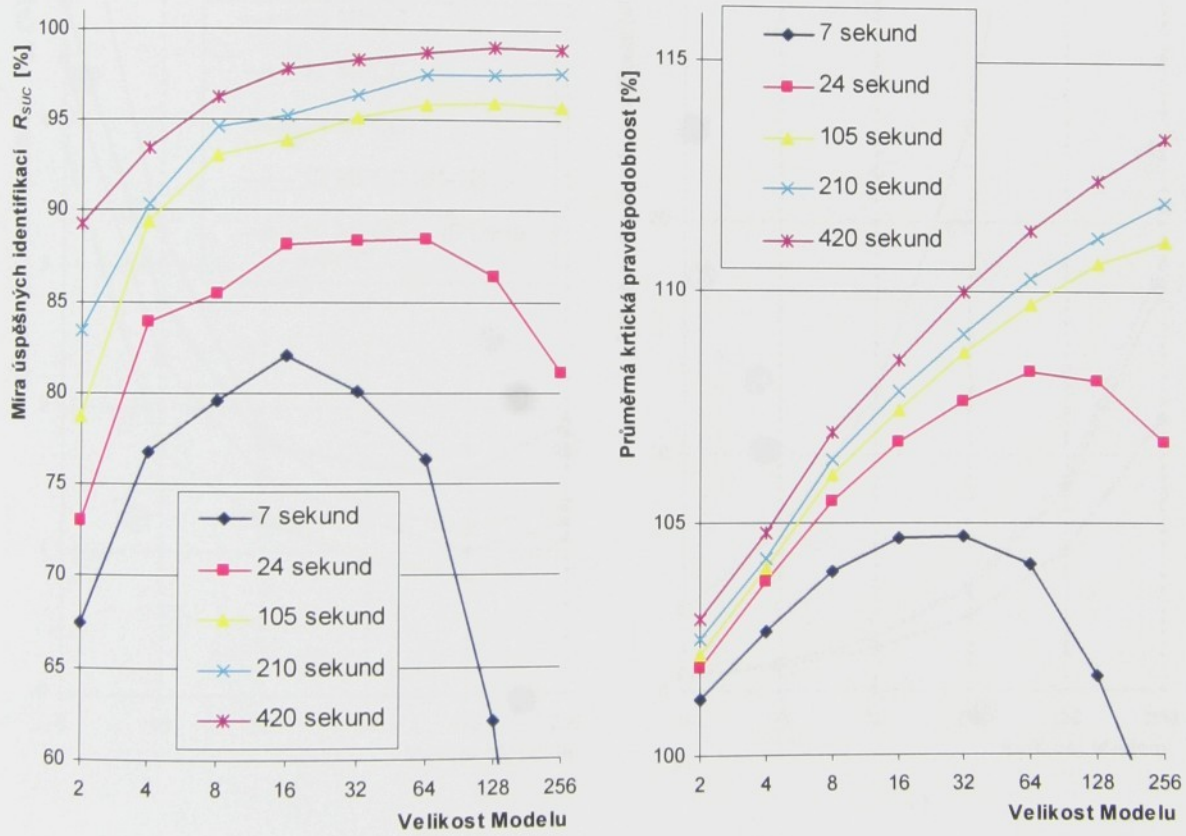
Míra ovlivnění úspěšnosti identifikace trénovacími daty

Kvalitní a vhodně zvolená data použitá v režimu trénování jsou velmi důležitou podmínkou úspěšné identifikace a verifikace mluvčího. Pro ověření míry ovlivnění úspěšnosti identifikace množstvím trénovacích dat bylo k dispozici 49 trénovacích nahrávek od každého mluvčího, pro testy tedy zbylo 25 nahrávek (v souhrnu 1375 testů, tj.

jedna správná identifikace nám zvýší míru úspěšných identifikací R_{SUC} o 0,073 %). Tyto nahrávky byly rozděleny následujícím způsobem:

Trénovací soubory	73–73	–> 7,18 sekundy	=> 0,12 minuty
	71–73	–> 23,6 sekundy	=> 0,39 minuty
	68–73	–> 48,5 sekundy	=> 0,81 minuty
	62–73	–> 104 sekund	=> 1,73 minuty
	52–73	–> 210 sekund	=> 3,5 minuty
	25–73	–> 419 sekund	=> 7 minut

GMM systém s diagonální kovarianční maticí jehož modely byly trénovány uvedenými délkami pomluv je zobrazen na obrázku 4.15. U promluvy s délkou trvání 7 sekund je nejmarkantnější velmi rychlý pokles rozpoznávacího skóre (obr. 4.15a). Pokud se pohybujeme v oblasti s větším počtem mixtur, nestačí tak krátká promluva pro statisticky věrohodné natrénování modelu. Naproti tomu R_{SUC} u modelů trénovaných 420 sekundami řeči má stoupající tendenci i při 256 mixturách.



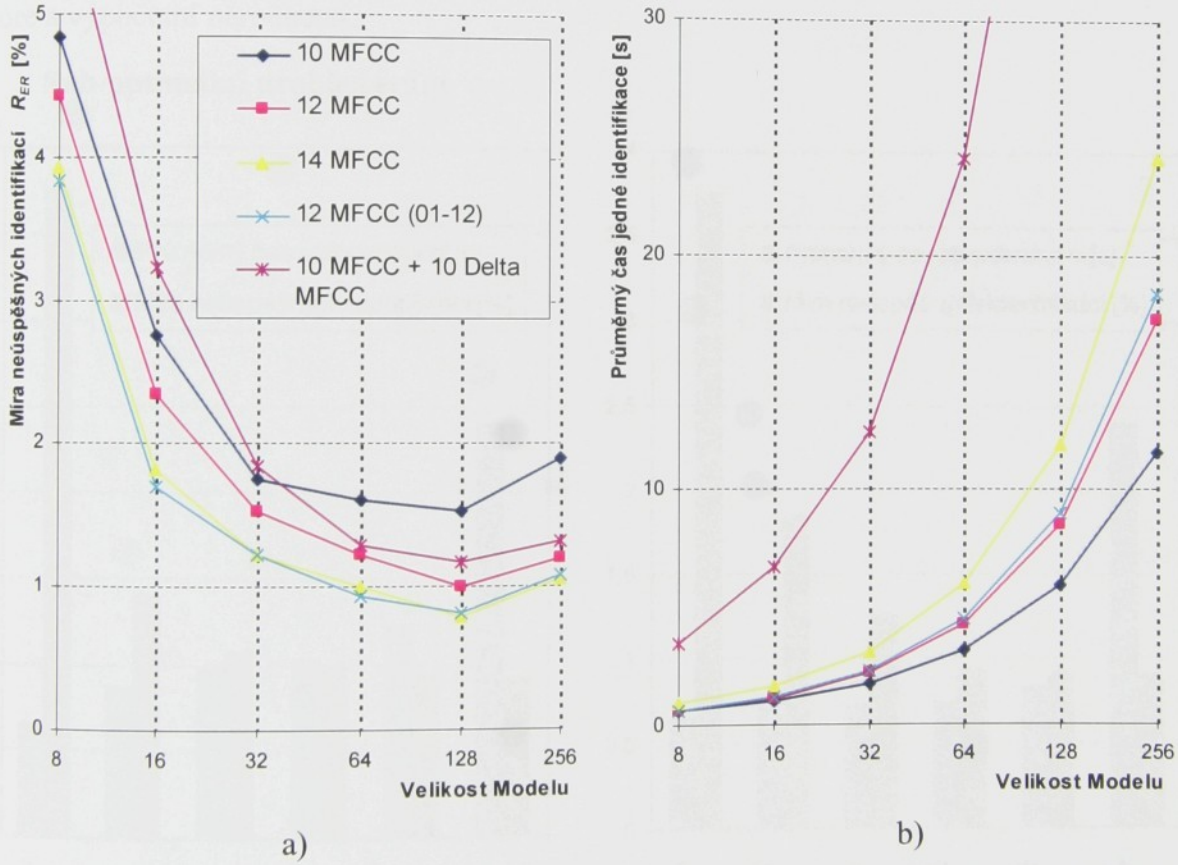
a) b)

Obr. 4.15: *Grafické znázornění průběhu míry úspěšných identifikací a průměrné kritické pravděpodobnosti GMM systému natrénovaného různými délkami promluv*

Na obr. 4.15b je vidět pokles kritických pravděpodobností[1] u totožného systému jako v případě obr. 4.15a. Můžeme vidět, že kritické pravděpodobnosti přímo souvisí s rozpoznávacím skóre. Tam kde stoupají kritické pravděpodobnosti, tam roste i míra R_{SUC}.

Tyto experimenty jsou pro systém přepisující zpravodajské pořady důležité z jednoho důvodu. Při trénování modelů mluvčích (IM, VM a adaptace řečových modelů na klíčové mluvčí) máme pouze omezený objem trénovacích dat. Bylo tedy nutné stanovit práh (délka dostupných řečových dat pro konkrétního řečníka), kdy mluvčí bude zařazen mezi klíčové mluvčí a kdy jeho data nebudou akceptována z důvodu nedostatečné délky. Nakonec byl práh stanoven na 100 sekund (kompromis mezi délkou nahrávek a množstvím mluvčích, kteří tento práh překonali). Rozborem experimentálních výsledků bylo pro tuto minimální délku trénovacích dat systém IM zvoleno 64 GMM složek.

Výběr optimálních kepstrálních příznaků



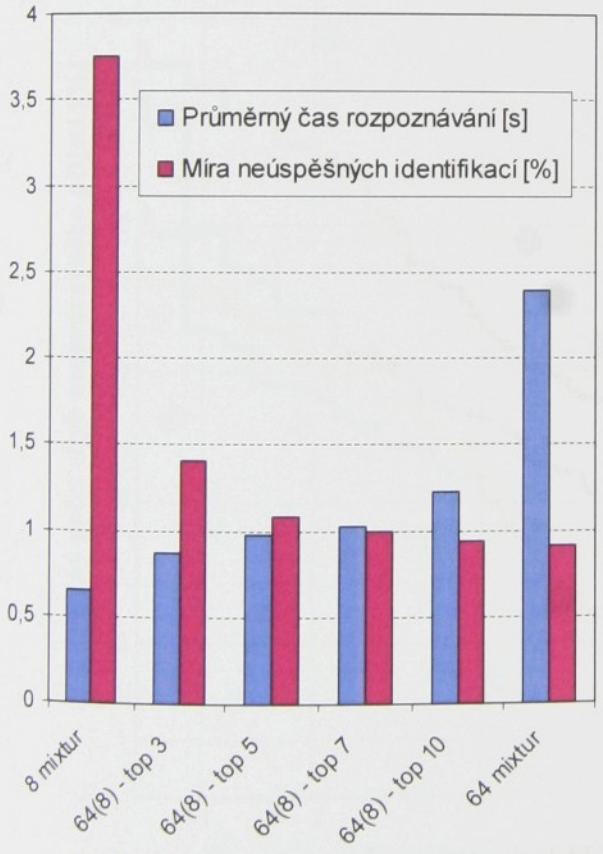
Obr. 4.16: *Grafické znázornění průběhu míry neúspěšných identifikací a průměrného času jedné identifikace GMM systému používajícího různé konfigurace vektorů příznaků*

[1] Princip kritických pravděpodobností je blíže vysvětlen na obr. 4.21.

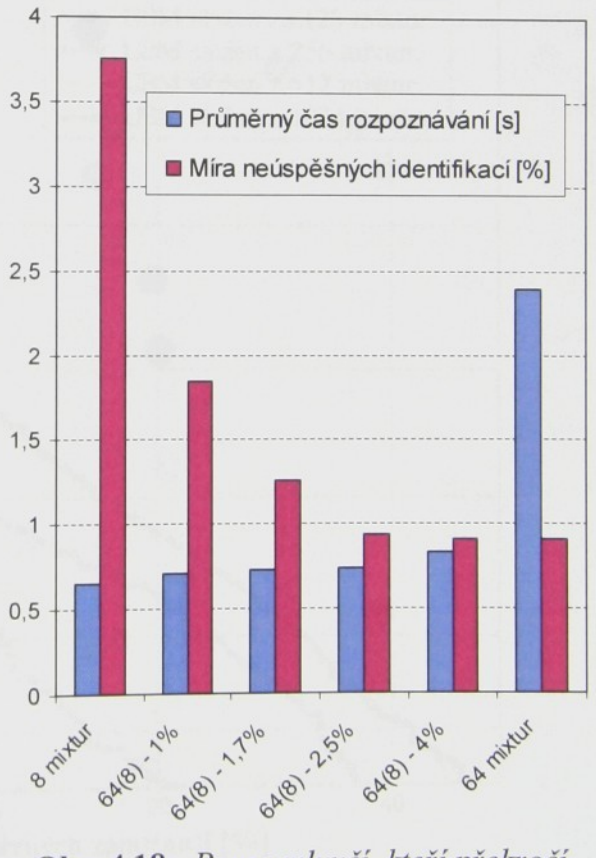
Pro tuto sadu experimentů bylo zvoleno 12 trénovacích nahrávek od každého mluvčího (přibližně 100 sekund spojité řeči), pro testy tedy zbylo 62 nahrávek (celkem 3410 testů, tj. správná identifikace nám zvýší míru úspěšných identifikací R_{SUC} o 0,029 %).

I v této sadě experimentů se projevila skutečnost, že počet gaussovských složek modelu mluvčího je třeba volit s ohledem na délku trénovacích dat (zvýšení R_{ER} u modelů s 256 složkami viz obrázek 4.16a). Nakonec jsme pro systém IM a VM zvolili 12 MFCC příznaků, které zaručují optimální výkonnost. Je třeba dodat, že tak jak je tomu u rozpoznávání řečníka zvykem, vynecháváme nultý kepstrální příznak, který nepřináší téměř žádnou aditivní informaci o identitě mluvčího. Dalším závěrem vyplývajícím z obrázků 4.16 je nepoužívat dynamické příznaky, které nepřinášejí takové zlepšení, které by ospravedlňovalo jejich nasazení vzhledem ke zvýšení výpočetní náročnosti (především v real-time aplikaci je nárůst výpočetní náročnosti kritickou otázkou). Takto zvolená konfigurace příznaků má pro náš systém IM a VM nejlepší poměr mezi rozpoznávacím skóre a výpočetní náročností.

Sub-optimální prohledávání

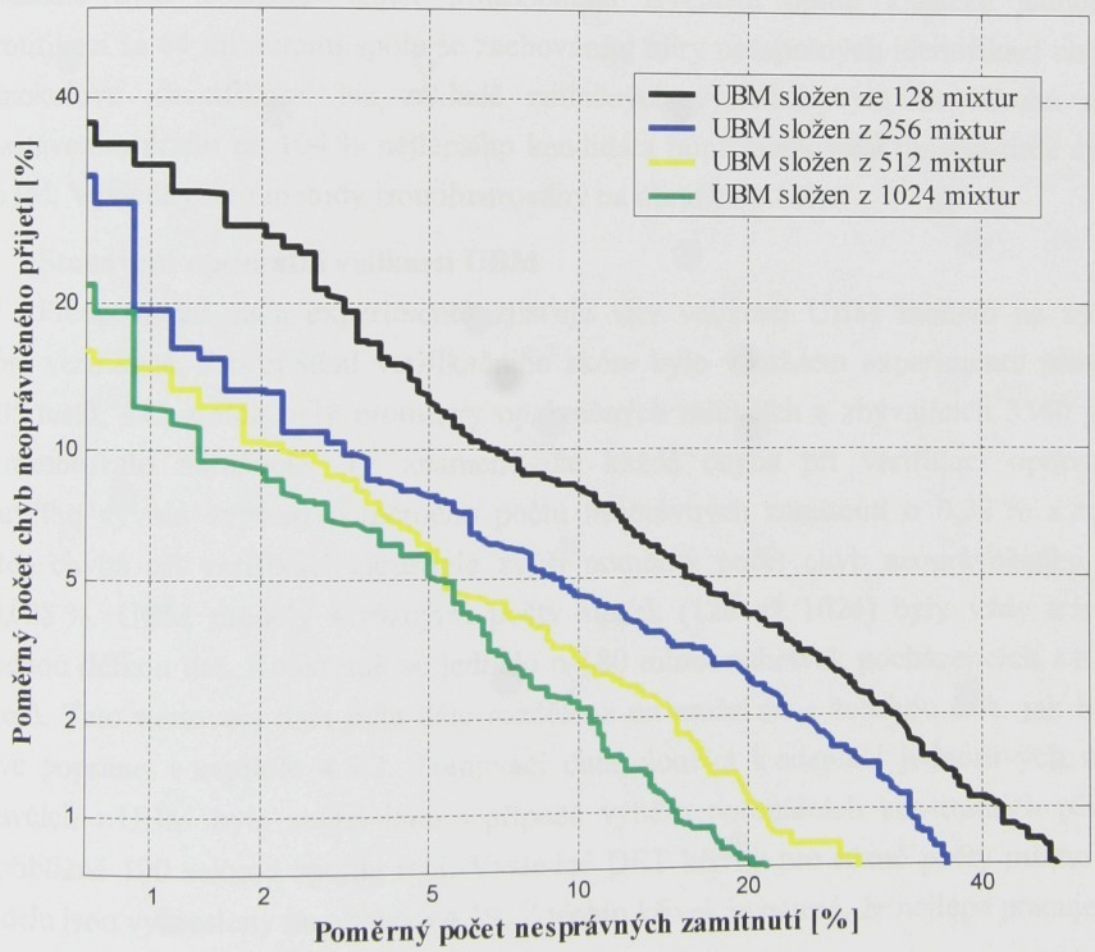


Obr. 4.17: *Pouze X nejlepších mluvčích postupuje do druhého kola*



Obr. 4.18: *Pouze mluvčí, kteří překročí X % postupují do druhého kola*

Pro tuto sadu experimentů byla zvolena stejná testovací a trénovací konfigurace dat jako v případě výběru optimálních kepstrálních příznaků. Jak je vidět z obr. 4.16b, roste čas potřebný na identifikaci jedné promluvy s počtem složek modelů téměř exponenciálně (samozřejmě, že použitím technik jako je například pruning a paměťová optimalizace lze tuto výpočetní náročnost do jisté míry redukovat). Vzhledem k tomuto faktu bychom měli používat jen tolik složek modelu mluvčího, kolik je jich nezbytně nutných. Naším řešením je rozdělení procesu identifikace do dvou průběhů. V prvním z nich je testovaná promluva porovnávána se všemi modely v databázi. Tyto modely jsou složeny pouze z omezeného počtu gaussovských složek (v našem případě z osmi), čímž je identifikace provedena se zmenšenou přesností, avšak velmi rychle. Druhý identifikační průběh je proveden s omezeným počtem kandidátů, jejichž zkrácené modely nejlépe reprezentují testovanou promluvu. Protože počet postoupivších kandidátů do druhého kola zřídka kdy přesáhne hodnotu deset, používáme již v tomto druhém kole optimální velikost GMM modelů (z předchozích experimentů vyplynulo 64 složek).



Obr. 4.19: *DET křivky pro verifikační systémy s různými velikostmi UBM modelů*

Postupů, jak vybrat X nejlepších mluvčích je hned několik. Na obr. 4.17 a 4.18 jsou vykresleny výsledky dvou přístupů (modré sloupce znázorňují průměrný čas potřebný pro provedení jedné identifikace v sekundách, fialové sloupce ukazují míru neúspěšných identifikací v procentech). První z těchto přístupů je velice jednoduchý. Napevno stanovíme počet kandidátů, kteří na základě skóre věrohodnosti vždy projdou do druhého kola. Výsledky této metody pro 3, 5, 7 a 10 vybraných kandidátů jsou na obrázku 4.17. Nalevo a napravo od těchto kandidátů jsou pro porovnání jednokolové výsledky identifikace s 8 mixturami (předvýběrová identifikace) a s 64 mixturami (finální identifikace). Výběr kandidátů u druhého přístupu spočívá ve vyhodnocení věrohodnosti všech mluvčích v prvním kole. Na základě mluvčího s nejvyšší věrohodností a stanoveného prahu jsou vybráni všichni kandidáti, kteří splňují podmínku dostatečné blízkosti k tomuto nejlepšímu mluvčímu. Tento přístup se velmi podobá výběru mluvčích do kohorty reprezentující alternativní hypotézu H_1 při VM (více v kapitole 4.5.1). Tento druhý přístup se ukázal jako velmi vhodný pro IM, neboť při stanoveném 104 % prahu akceptovaných kandidátů bylo dosaženo téměř třínásobného zrychlení oproti klasické jednokolové identifikaci se 64 mixturami spolu se zachováním míry neúspěšných identifikací na úrovni jednokolové identifikace. Na základě zmíněných výsledků byla tato druhá metoda s nastavením prahu na 104 % nejlepšího kandidáta implementována do real-time systému pro IM. Výsledky této metody jsou ilustrovány na obrázku 4.18.

Stanovení optimální velikosti UBM

Předposlední sada experimentů zjišťuje vliv velikosti UBM modelu na výsledné skóre verifikace. Pro zjištění verifikačního skóre bylo v každém experimentu provedeno 4005 testů, 445 z nich byly promluvy oprávněných mluvčích a zbývajících 3560 pokusů reprezentovalo narušitele. To znamená, že každá chyba při verifikaci oprávněného mluvčího vyvolá zvýšení poměrného počtu nesprávných zamítnutí o 0,22 % a naopak, každá chyba při verifikaci narušitele zvýší poměrný počet chyb neoprávněného přijetí o 0,028 %. UBM modely s různými počty složek (128 až 1024) byly vždy trénovány shodnou délkou dat. Konkrétně se jednalo o 180 minut nahrávek pocházejících z různých zdrojů. Tato trénovací data byla dále rozdělena na mužskou a ženskou část, jak bylo již dříve popsáno v kapitole 4.5.2. Trénovací data sloužící k adaptaci jednotlivých modelů mluvčích z UBM byla stejná jako v případě výběru optimálních kepstrálních příznaků, tj. přibližně 100 sekund spojité řeči. Výsledné DET křivky pro různé počty mixtur UBM modelu jsou vykresleny na obrázku 4.19. Z těchto křivek je patrné, že nejlépe pracuje UBM model s 1024 složkami.

4.6.4. Reálný testovací systém

Obr. 4.20: *Postup porovnávání promluvy neznámého mluvčího s jedním GM modelem*

Cílem rozpoznávání je určení identity mluvčího či její ověření. V případě identifikace mluvčího pomocí GMM je dosaženo určení identity výpočtem posteriorní pravděpodobnosti, s jakou konkrétní model z databáze mluvčích generuje testovaný vektor příznaků. Ve fázi rozpoznávání je tedy vypočítávána celková pravděpodobnost vstupní promluvy vůči všem složkám daného GMM (jednotlivé posteriorní pravděpodobnosti se pro všechny vstupní vektory vynásobí a vznikne tak celková posteriorní pravděpodobnost). Mluvčí, který dosáhne se svým modelem nejvyšší pravděpodobnosti vůči zparametrizované testované promluvě, je prohlášen za testovaného mluvčího (v tomto konkrétním příkladě se jedná o identifikaci nad uzavřenou množinou). Na obr. 4.20 je tento postup naznačen i graficky.

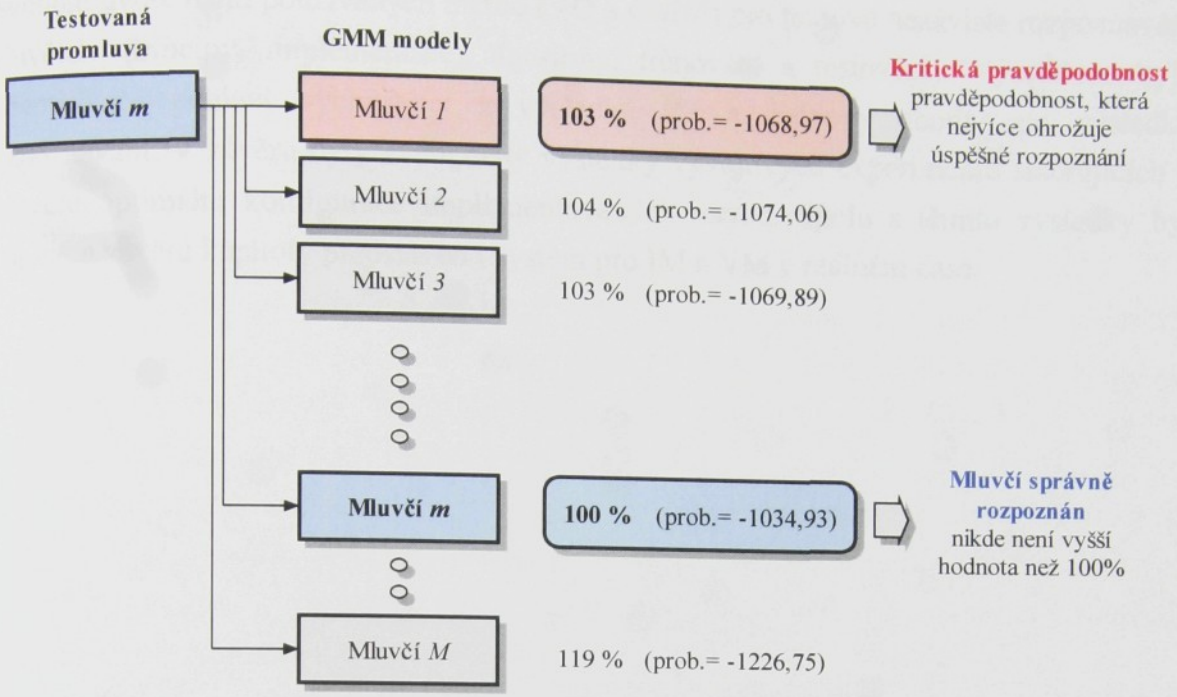
Jak tedy probíhá vlastní rozpoznávání?

V reálném světě bude náš systém pro určování totožnosti mluvčího pracovat následujícím způsobem (pro jednoduchost uvádíme jednokolovou identifikaci):

0. Na začátku je třeba vytvořit databázi mluvčích, kteří budou identifikační systém (dále již jen IS) používat. Tito mluvčí budou reprezentováni GMM modely, které vytvoříme postupem uvedeným v kapitolách 4.4.2 až 4.4.4. Takto vytvořené modely budou uloženy v IS a používány při identifikaci.

---

1. Osoba, která chce být identifikována, přistoupí k našemu systému pro identifikaci mluvčího a vysloví jakoukoliv promluvu o přibližné délce 2 až 5 sekund.
2. IS v části určené k extrakci příznaků tuto promluvu převede na posloupnost vektorů příznaků.
3. IS tyto příznakové vektory porovná s prvním GMM, který má uložen ve své databázi. Výsledkem porovnání bude pravděpodobnost, která charakterizuje podobnost testované promluvy s GMM (čím vyšší je pravděpodobnost, tím si je promluva a model podobnější).



Obr. 4.21: *Naznačení postupu IM proti celé databázi mluvčích*

4. Krok 3 opakujeme tak dlouho, dokud nejsou vypočítány pravděpodobnosti testované promluvy vůči všem GMM mluvčích v databázi (viz obr. 4.21). Jediné, co je třeba si zapamatovat, je věrohodnost a index modelu, ke kterému tato vypočtená věrohodnost náleží.
5. IS nalezne nejvyšší pravděpodobnost. Index spojený s touto pravděpodobností ukazuje na konkrétního mluvčího, jehož hlasem byl GMM natrénován. Tento mluvčí je následně určen jako identifikovaná osoba.

Další informace o tomto programu spolu s grafickým uživatelským rozhraním lze nalézt v příloze 1.

4.7. Shrnutí kapitoly

Na tomto místě by bylo vhodné krátce shrnout nejpodstatnější informace uvedené v této obsáhlé kapitole.

V úvodu kapitoly bylo provedeno vymezení problematiky rozpoznávání mluvčích vzhledem k ostatním biometrickým metodám identifikace osob (kam zapadá rozpoznávání mluvčího vzhledem k dosahovaným úspěšnostem rozpoznávání, pořizovacím nákladům a možnostem praktického použití). Úvodní část kapitoly byla zakončena rozdělením rozpoznávání mluvčího do dvou základních kategorií – IM a VM s popisem jejich základních částí. V další části kapitoly byly představeny některé přístupy pro textově závislou a nezávislou identifikaci, resp. verifikaci mluvčích. Poté následovalo podrobné

osvětlení dvou, námi používaných metod (VQ a GMM) pro textově nezávislé rozpoznávání mluvčích (principy, implementace, algoritmy, trénování a testování). Vysvětlena byla i identifikace pohlaví vycházející z výsledků IM a způsoby ohodnocení výsledků rozpoznávání. V závěru byly zveřejněny výsledky vývojových experimentů směřujících k nalezení optimální konfigurace implementovaných metod. Spolu s těmito výsledky byl v úplném závěru kapitoly představen i systém pro IM a VM v reálném čase.

5. Metody identifikace řečových segmentů

5.1. Úvod

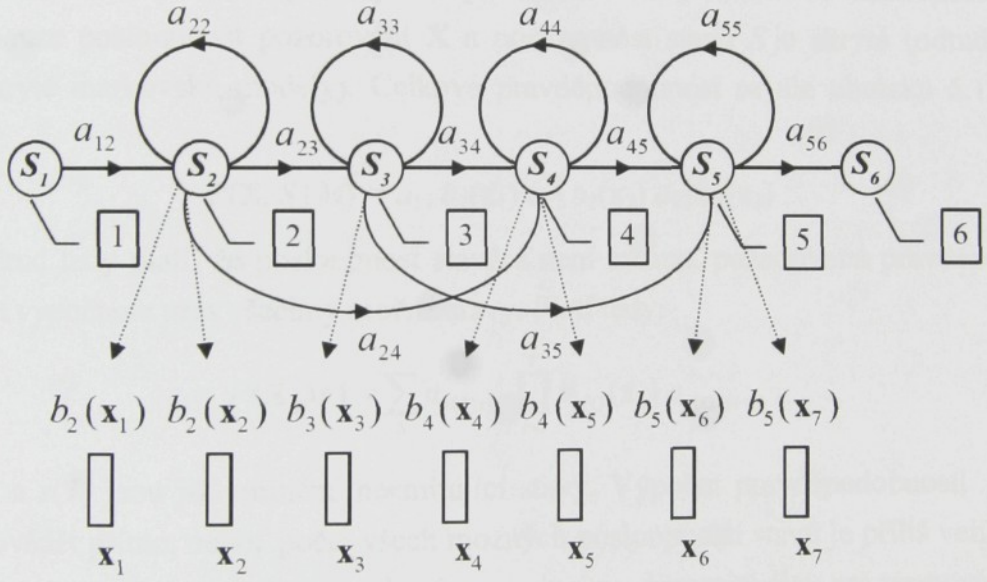
Snahou mnoha vývojových týmů pracujících v oblasti automatického přepisu zpravodajství je oddělení audiosegmentů obsahujících řeč od neřečových segmentů ještě před vlastním zpracováním v LVCSR. Důvody jsou přinejmenším dva. Tím prvním je zvýšení rozpoznávacího skóre, neboť i když LVCSR může obsahovat speciální modely pro různé ruchy, vždy je určitá část neřečových segmentů rozpoznána jako text. Navíc, tento nesmyslný text může dále negativně ovlivnit rozpoznávání následujících řečových segmentů. Druhým důvodem je čas rozpoznávání, který se zpracováním nesmyslných neřečových segmentů neúměrně prodlužuje.

První pokusy s klasifikací audiosegmentů na řečové a neřečové byly většinou založeny na jednoduchých příznacích (např. počet průchodů nulou nebo krátkodobá energie). Jak je známo, tento přístup není příliš robustní s ohledem na snižující se odstup signálu od šumu (Signal to Noise Ratio – *SNR*) a selhává také při větší variabilitě audiosignálu. V [SCH97] byly následně představeny další, lépe diskriminující příznaky spolu s testy několika dalších metod zahrnující GMM, ANN se zpětným učením a *k*NN. V práci [ZHA92] byly představeny metody využívající k diskriminaci do skupin jako je hudba nebo řeč plus hudba hlasivkové periody spolu s heuristickým modelem. Pozměněný algoritmus založený na metodě nejbližšího souseda a LSP-VQ (Linear Spectral Pairs-Vector Quantization) je použit v práci [LIE02a]. Spojitý audiosignál je tam tříděn do čtyř základních kategorií – řeč, hudba, okolní hluk a ticho za pomoci jednosekundového okna, které je menší než okna do té doby standardně používaná (viz například [SCH97]). Dále jsou používány metody jako 64 složková GMM diskriminující audiosignál na řeč, čistou řeč a ostatní [GAU02], vektorová kvantizace plus nové akustické příznaky [LIN05], SVM a pět audiokategorií (čistá řeč, řeč plus okolí, zvuk na pozadí, ticho a hudba) doplněné novými spektrální příznaky [LIE03] nebo hybridní struktura HMM a neuronových sítí plus nové příznaky (entropie, dynamika a další) [WIL99].

Naše práce nejvíce navazuje na systém vytvořený týmem Vandecatseye & Martens [VAN03], kde jsou pro klasifikaci do pěti audiokategorií (řeč, řeč+hudba, řeč+ostatní, hudba, ostatní) použity HMM a Viterbiho dekodér. Po rozčlenění spojitého audiosignálu tímto postupem ještě následuje logický postprocessing, který má za úkol odstranit

nesprávně vložené krátké pauzy (< 2 s) a řečové segmenty (řečový úsek < 0,2 s mezi dvěma neřečovými segmenty).

Volba rozpoznávací metody vycházela z myšlenky použít shodných rozpoznávacích příznaků v celém systému pro přepis zpravodajských pořadů. Jelikož metoda HMM ve spojení s MFCC kepstrálními příznaky byla úspěšně použita jak při rozpoznávání spojité řeči, tak i v jednostavové alternativě při IM, VM a IP, rozhodli jsme se využít HMM i v úloze identifikace řečových segmentů. Pouze standardně používané levo-pravé uspořádání HMM, typické především pro úlohu rozpoznávání řeči, bylo v případě identifikace řečových segmentů nahrazeno strukturou ergodickou (podrobněji v kapitole 5.3). Tato struktura totiž dokáže lépe popsat chování a především variabilitu řečových a neřečových segmentů, kdy jsou v některých extrémních případech použity totožné *n*-stavové HMM pro délku segmentu 2 s a jindy pro délku 20 s. Dále je tak možné lépe modelovat opakování některých pasáží v rámci jednoho segmentu, k čemuž dochází například u znělek nebo upoutávek na další pořady.



Obr. 5.1: *Ukázka principu generování vektorů pozorování markovským modelem*

Pro vytvoření zde popsaného systému identifikujícího řečové segmenty v nahrávkách televizního zpravodajství byl použit systém HTK (Hidden Markov Model Toolkit). HTK je soubor nástrojů vytvořených pro experimentování s HMM a spolu s propracovanou implementací umožňuje sestavení kompletního rozpoznávacího systému od parametrizace až po vyhodnocení obdržených výsledků. Protože je systém HTK navržen podle obecně platných pravidel, je s ním možné zpracovávat jakékoliv v čase se vyvíjející procesy, ovšem primárně je určen na vytváření systémů pro rozpoznávání řeči. V následujících odstavcích této kapitoly zmíníme pouze obecné principy použitých metod a algoritmů,

neboť jejich přímá implementace je provedena v HTK, podrobnější popis je možno nalézt v [YOU02].

5.2. HMM klasifikátor

Navažme na informace o HMM uvedené v kapitole 4.2.1 o skutečnosti související přímo s rozpoznáváním řeči, neboť do této kategorie patří i zde diskutovaný rozpoznávač. Na HMM můžeme nahlížet jako na pravděpodobnostní konečný automat modelující stochastické procesy tak, že v pevně stanovených diskrétních časových okamžicích generuje náhodnou posloupnost vektorů pozorování. Na základě souboru předem stanovených pravděpodobností přechodů, změní v každém kroku model svůj stav (přechodem do jiného stavu nebo setrváním ve stávajícím). Tento stav, do kterého model přejde, dle své výstupní pravděpodobnosti vygeneruje jeden vektor pozorování[1].

Pravděpodobnost řady $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_T\}$, která je generována daným modelem M pohybem přes posloupnost stavů S, se jednoduše spočítá jako součin přechodových pravděpodobností a_{ij} a výstupních pravděpodobností stavů $b_i(\mathbf{x}_t)$. Ve skutečnosti je však známá pouze posloupnost pozorování $\mathbf{X}$ a posloupnost stavů S je skrytá (odtud pochází název skryté markovské modely). Celková pravděpodobnost se dle obrázku 5.1 vypočte jako

$$P(\mathbf{X}, S \mid M) = a_{11}\, b_2(\mathbf{x}_1)\, a_{22}\, b_2(\mathbf{x}_2)\, a_{23}\, b_3(\mathbf{x}_3) \ldots \tag{5.1}$$

Pokud tedy platí, že posloupnost stavů S není známa, požadovaná pravděpodobnost může být vypočtena přes všechny možné stavy. Platí tedy:

$$P(\mathbf{X}|M) = \sum_{S} a_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\mathbf{x}_t)\, a_{s(t)s(t+1)}\,, \tag{5.2}$$

kde $s(0)$ a $s(T)$ jsou již zmíněné neemitující stavy. Výpočet pravděpodobnosti $P(\mathbf{X}|M)$ nelze provádět přímo, neboť počet všech možných posloupností stavů je příliš veliký. Proto se používají iterační techniky založené na principu dynamického programování. Silná stránka použití skrytých markovských modelů spočívá právě v existenci efektivních algoritmů pro estimaci jejich parametrů a algoritmů používaných při rozpoznávání.

Pravděpodobnost i-tého slova w_i ze slovníku (pro jednoduchost uvažujme úlohu rozpoznávání izolovaných slov) tedy určíme opět za pomoci Bayesova vzorce jako

$$P(w_i|\mathbf{X}) = \frac{P(\mathbf{X}|w_i)P(w_i)}{P(\mathbf{X})}, \tag{5.3}$$

[1] Výjimkou může být první a poslední stav, které v některých případech slouží pouze pro zahájení a ukončení generační posloupnosti (jedná se tedy o takzvané neemitující stavy).

kde apriorní pravděpodobnosti $P(\mathbf{X})$ a $P(w_i)$ můžeme v našem případě zanedbat, neboť v úloze identifikace řečových segmentů uvažujeme shodnou pravděpodobnost výskytu jednotlivých slov (audiotříd) a $P(\mathbf{X})$ je při hledání jednoho slova vždy stejná. Vítězné slovo, které nejlépe odpovídá analytickému vzoru $\mathbf{X}$, je nalezeno podle vztahu

$$w^* = \underset{i}{ArgMax}\, P(\mathbf{X}|w_i). \tag{5.4}$$

Výstupní pravděpodobnosti stavů $b_s(\mathbf{x}_t)$ jsou ve většině HMM případů popsány směsí gaussovských rozložení hustot pravděpodobnosti. Vztah pro výpočet pravděpodobnosti stavu s pro příznakový vektor $\mathbf{x}_t$ vzhledem k části modelu charakterizovaného střední hodnotou $\bar{\mathbf{x}}_{sm}$ a kovarianční maticí Σ_{sm} je dán sumou přes všechna gaussovská rozložení směsi. Vliv každé mixtury je vážen koeficientem c_{sm}. Zmíněný postup výpočtu výstupní funkce jednoho stavu HMM je v podstatě identický s postupem výpočtu u GMM z kapitoly 4.4.1 vztah (4.13), tj.

$$b_s(\mathbf{x}_t) = \sum_{m=1}^{M} c_{sm} N(\bar{\mathbf{x}}_{sm}, \Sigma_{sm}) = \sum_{m=1}^{M} c_{sm} \frac{1}{\sqrt{(2\pi)^P \det \Sigma_{sm}}} . \exp[-\tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_{sm})^T \Sigma_{sm}^{-1} (\mathbf{x}-\bar{\mathbf{x}}_{sm})]. \tag{5.5}$$

Viterbiho dekodér

Pravděpodobnost $P(\mathbf{X}|M)$, že promluva $\mathbf{X}$ byla vygenerována modelem M reprezentujícím slovo w je přitom vybrána jako maximální z pravděpodobností vypočítaných přes všechna přípustná přiřazení sekvence vektorů příznaků slova $\mathbf{X}$ ke stavům modelu daného slova a dá se získat Viterbiho algoritmem (Viterbiho dekodér je v HTK implementován do nástroje *HVite*). Při jeho implementaci je možné zavedením kumulovaného součinu $V(t,s)$ využít metody dynamického programování. Kumulovaný součin $V(t,s)$ je definován jako:

$$V(t,s) = b_s(\mathbf{x}_t)\, Max[a_{ss} V(t-1,s), a_{s-1s} V(t-1,s-1)]. \tag{5.6}$$

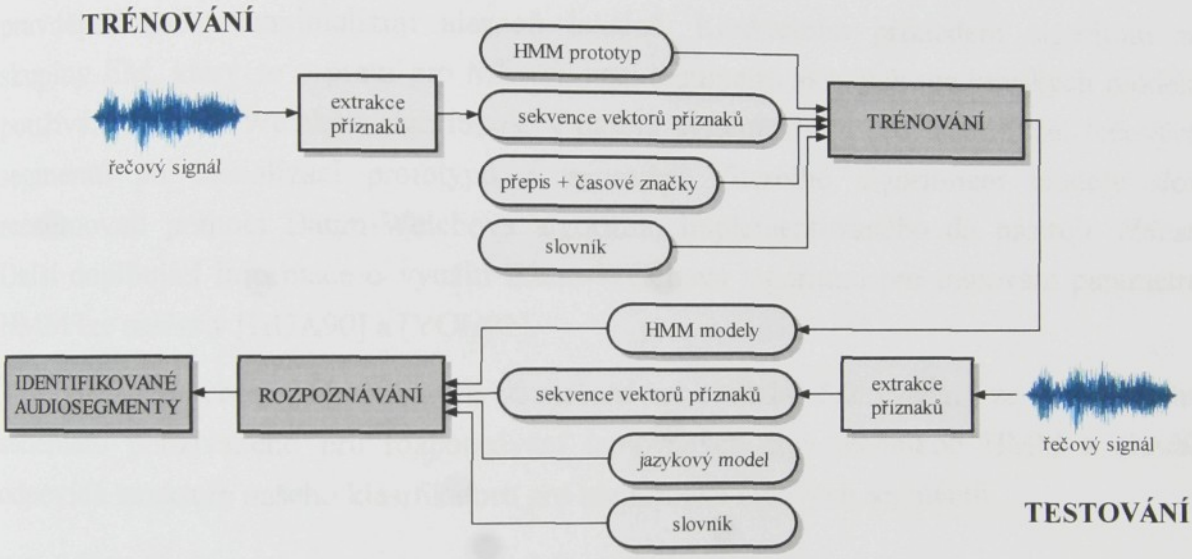
Tento koncept nalezení nejlepší cesty je velice důležitý a lze jej generalizovat i pro případ, kdy jsou jednotlivá slova ze slovníku rozpoznávače izolovaných slov reprezentována hláskovými modely. Model každého slova ze slovníku je v tomto případě sestaven z posloupností stavů modelů jednotlivých hlásek, které odpovídají fonetickému přepisu slova.

Rozpoznávání spojité řeči je mnohem složitější. Oproti rozpoznávání izolovaných slov se ve zpracovávaném signálu vyskytuje neznámý počet slov, která mohou začínat a končit v libovolném časovém okamžiku a následovat libovolně po sobě. V našem systému pro identifikaci řečových segmentů používáme celoslovní modely, které jsou za sebou zřetězeny pomocí gramatiky, viz kapitola 5.3. Tento přístup je ovšem použitelný pouze na malých slovnících (náš případ). S rostoucím slovníkem (nebo u aplikací s proměnnými

soubory slov) se tento postup stává neefektivním, ba přímo nepoužitelným. Proto se v těchto případech daleko častěji používají rozpoznávače spojité řeči založené na reprezentaci slov hláskovými modely. Podrobnosti lze nalézt v [NOU01].

Trénování

Stejně jako u GMM je třeba i při trénování HMM nalézt optimální parametry modelu. Navíc, při procesu generování řeči skrytým markovským modelem „prochází" model postupně posloupností svých stavů. Tato posloupnost je však pozorovateli skryta, neboť lze pozorovat jen posloupnost vektorů příznaků. Jak metoda užívaná při rozpoznávání, tak metoda užívaná při odhadu parametrů skrytého markovského modelu potřebuje umět nalézt nejlepší, resp. nejpravděpodobnější posloupnost stavů.



Obr. 5.2: *Blokové schéma postupu při trénování a identifikaci řečových segmentů pomocí HTK toolkitu*

Nejjednodušší metodou nalezení parametrů HMM je postup založený na možnosti použít Viterbiho algoritmus přiřazení framů trénovaného slova ke stavům jeho modelu. Tento algoritmus přeskupí pozorované vektory mezi jednotlivými stavy takovým způsobem, aby maximalizoval výslednou pravděpodobnost, jakou může konkrétní model vygenerovat s danými vektory. Poté jsou znovu přepočteny jednotlivé parametry. Tento proces se opakuje do té doby, dokud se mění odhady výstupních parametrů a dokud významně roste výsledná pravděpodobnost. Postup se pak opakuje pro všechna slova ze slovníku. Při trénování M-mixturových markovských modelů jsou v každém kroku výše popsaného postupu všechny framy přiřazené k danému stavu ještě iteračním algoritmem k-means rozděleny do M shluků. Každý z těchto shluků pak reprezentuje jednu mixturu. Z framů přiřazených k jednotlivým shlukům jsou pak určeny střední hodnoty a rozptyly

odpovídajících mixtur. Váhové koeficienty všech mixtur daného stavu jsou určeny jako poměr počtu framů přiřazených k odpovídajícímu shluku a celkového počtu všech příznakových vektorů přiřazených k danému stavu. Tento iterační postup trénování spojený s určením počátečního rozdělení je v HTK implementován do nástroje *HInit*.

Viterbiho algoritmus přiřazení odhadu parametrů modelu pracuje na principu konkrétního přiřazení pozorovaného vektoru jednomu stavu modelu s ohledem na *maximalizaci výsledné pravděpodobnosti*. Nabízí se druhý postup nazývaný *metoda maximální věrohodnosti*, kdy je každý vektor přiřazen každému stavu modelu pomocí pravděpodobnosti. Pro skryté markovské modely zatím nebyla nalezena žádná metoda, která by u tohoto přístupu umožnila dosažení globálního maxima výsledné pravděpodobnosti. Existuje však několik postupů založených na EM algoritmu, které tuto pravděpodobnost maximalizují alespoň lokálně. Konkrétním příkladem algoritmu ze skupiny EM, který se v praxi pro ML estimaci parametrů skrytých markovských modelů používá, je Baum-Welchův algoritmus. V našem systému jsme pro identifikaci řečových segmentů po inicializaci prototypů a trénování Viterbiho algoritmem modely slov reestimovali pomocí Baum-Welchova algoritmu implementovaného do nástroje *HRest*. Další doplňující informace o využití Baum-Welchova algoritmu pro trénování parametrů HMM lze nalézt v [HUA90] a [YOU02].

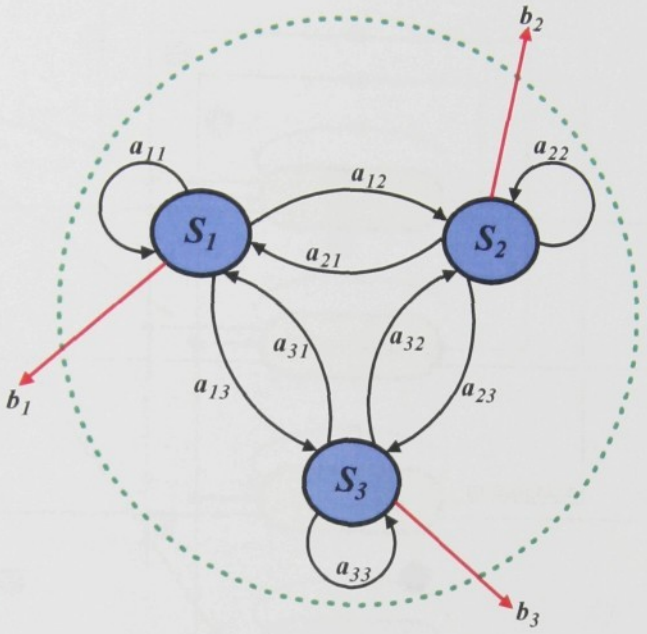
Blokové schéma klasifikátoru naznačené na obrázku 5.2 vychází ze standardního schématu používaného pro rozpoznávání izolovaných slov technikou HMM a zhruba odpovídá struktuře našeho klasifikátoru pro identifikaci řečových segmentů.

5.3. Topologie modelu

Pro rozpoznávací systém je nutné nejprve specifikovat strukturu skrytých markovských modelů. Pro všechny modely reprezentující námi stanovené audiokategorie byl zvolen jednotný typ markovského modelu. Experimentálně byla určena jeho ergodická pětistavová struktura sloužící k vlastnímu modelování příslušných akustických událostí. Na obrázku 5.3 je schematicky naznačen třístavový ergodický model, který byl také v našem systému testován.

Se strukturou modelů úzce souvisí i způsob parametrizace akustických dat. V našich experimentech jsme používali parametrizaci čítající až šestnáct melovských kepstrálních koeficientů, jejich první a druhé derivace, energii, CMS, liftraci a podobně. Testováno bylo ještě několik dalších konfigurací parametrů, jejichž některé výsledky je možné nalézt v kapitole 7.1. Nakonec byla zvolena konfigurace prvního až dvanáctého melovského koeficientu (stejně jako v úloze rozpoznávání řečníka nám nultý koeficient nedodává

žádnou aditivní informaci, která by nám pomohla při identifikaci) bez prvních a druhých derivací. Omezení pouze na kepstrální koeficienty jsme zvolili z důvodu použitelnosti vektorů příznaků ve všech modulech celého systému pro přepis zpravodajských pořadů, tj. parametrizace se provádí pouze jednou, na úplném začátku transkripčního procesu a jednotlivé moduly si již dále tyto vektory příznaků pouze načítají.



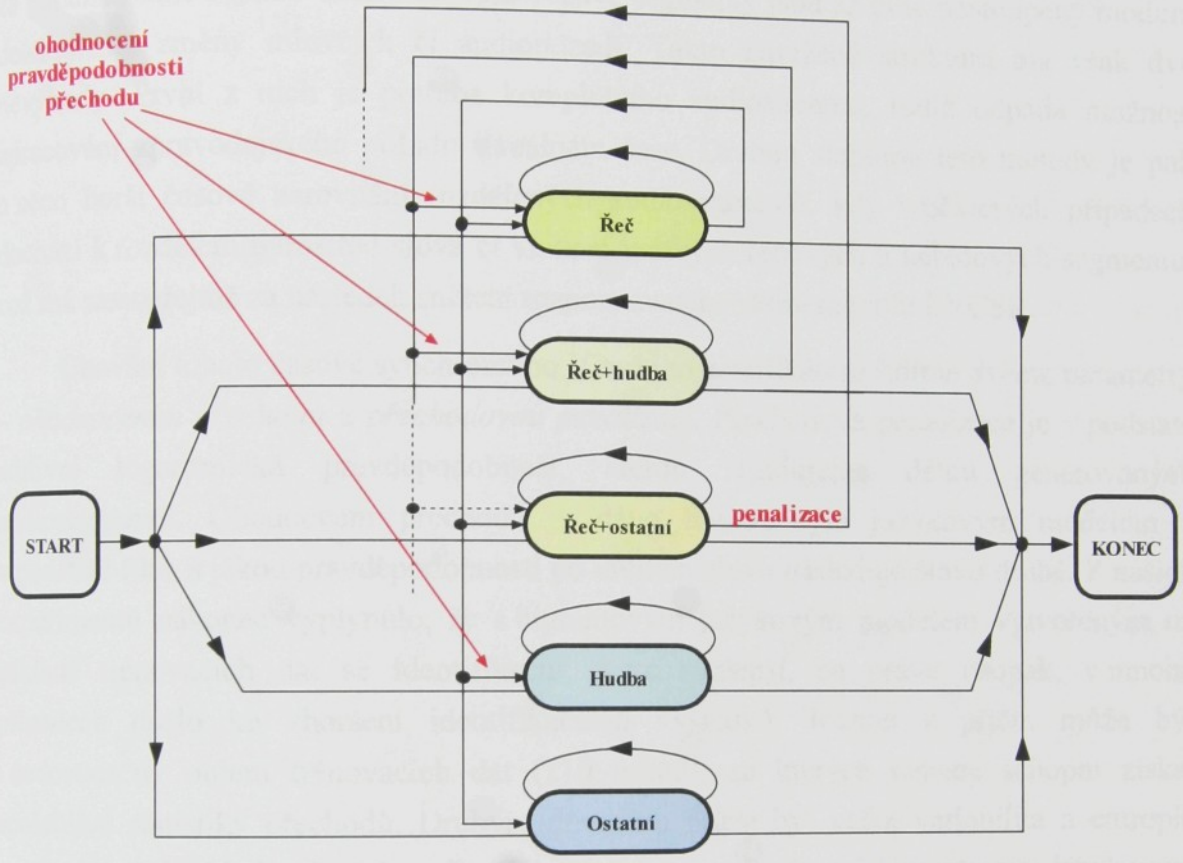
Obr. 5.3: *Schéma třístavového ergodického HMM*

Z literatury uvedené na začátku této kapitoly vyplývá, že počet audiokategorií bývá většinou volen od tří do šesti. V našem případě je řečový signál reprezentován třemi kategoriemi – řeč, řeč s hudbou na pozadí a řeč s ostatními ruchy na pozadí. Neřečový audiosignál spadá do jedné ze dvou kategorií – hudba a vše ostatní.

Rozpoznávání podle gramatik

Gramatika má obecně za úkol stanovit jazyková omezení, jež jsou kladena na různé posloupnosti slov, která budou rozpoznávána. Tato omezení mohou být buď deterministická (některé posloupnosti slov se vůbec nepřipouštějí) nebo pravděpodobnostní (některé posloupnosti slov jsou méně či více pravděpodobné). Pokud se chystáme řešit problém prostřednictvím statistického přístupu, budeme pracovat se stochastickou gramatikou, která vzniká přidáním pravděpodobností k jednotlivým odvozovacím pravidlům původní gramatiky. Příkladem je i struktura našeho systému pro identifikaci řečových segmentů (viz obrázek 5.4). Jedná se o schéma sestavené z HMM (tyto modely reprezentují jednotlivé audiokategorie), pomocí kterých chceme rozdělit spojitý audiozáznam. Jak z grafu vyplývá, rozpoznávač podle tohoto schématu předpokládá, že v příchozím signálu může být libovolná kombinace audiokategorií o libovolné délce. Pokud

si za audiokategorie (slova), která jsou v podstatě jen symboly označující příslušné markovské modely, tyto modely dosadíme, vznikne tak jedna velká síť a rozpoznávání vede na hledání nejlepší cesty touto sítí. Vedle dosažení nejlepšího skóre je zde ovšem nutné si ještě zapamatovat cestu, která tohoto skóre dosáhla, tedy výsledek rozpoznávání.



Obr. 5.4: *Rozpoznávací síť vytvořená z modelů audiokategorií, jež je použita pro identifikaci řečových segmentů ve spojitém audiozáznamu*

Takto vytvořenou strukturou lze modelovat různé úlohy rozpoznávání řeči, od těch nejjednodušších, až po velice komplikované. Příkladem jednodušší aplikace může být hlasová volba volaného, jak ji můžeme vidět u některých mobilních telefonů. Za složitějším příkladem aplikace si lze představit automatický hlasový informační systém, který s námi po telefonní lince vede dialog s pevně danou strukturou dotazů a odpovědí. Takovéto gramatiky jsou mocným nástrojem pro tvorbu pevně vázaných aplikací. Ovšem v případě plynulé a ničím nevázané promluvy nelze takovýchto gramatik použít. Přirozený jazyk obsahuje tak rozsáhlý slovník a tak širokou škálu všech možných slovních a větných kombinací, že není reálné vytvořit takovou gramatiku, která by dokázala celý tento rozsah pokrýt.

5.3.1. HMM klasifikátor na spojitém základu

Schéma naznačené na obrázku 5.4 přesně odpovídá struktuře našeho systému pro detekci řečových segmentů, který pracuje se spojitým audiosignálem. To znamená, že detekce řečových segmentů je z časového hlediska vykonávaných operací zařazena ihned za parametrizaci signálu. Identifikované řečové segmenty jsou až poté postoupeny modulu detekujícího změny mluvčích či audiopozadí. Takto navržená struktura má však dvě nevýhody. První z nich je potřeba kompletního audiostreamu, tudíž odpadá možnost zpracování zpravodajského pořadu v reálném čase. Druhou slabinou této metody je pak o něco horší časové zarovnání rozdělených audiosegmentů, kdy v některých případech dochází k rozdělení uprostřed slova či vložení krátkých řečových a neřečových segmentů, což má samozřejmě za následek snížení rozpoznávacího skóre modulu LVCSR.

Chování tohoto časově synchronního Viterbiho klasifikátoru řídíme dvěma parametry – *ohodnocením přechodu* a *přechodovou penalizací*. Přechodová penalizace je v podstatě aditivní logaritmická pravděpodobnost, kterou regulujeme délku generovaných audiosegmentů. Ohodnocení přechodů je dáno bigramovým jazykovým modelem a v podstatě říká, s jakou pravděpodobností po určitém slovu následuje slovo druhé. Z našich experimentů nakonec vyplynulo, že s bigramovým jazykovým modelem vytvořeným na základě trénovacích dat se identifikační skóre nezlepší, ba právě naopak, v mnoha případech došlo ke zhoršení identifikačních výsledků. Jednou z příčin může být i nedostatečný objem trénovacích dat (110 minut), ze kterých nejsme schopni získat spolehlivé statistiky přechodů. Druhým důvodem může být velká variabilita a entropie v případě střídání jednotlivých audioudálostí, jež nejsme schopni bigramovým jazykovým modelem uspokojivě popsat. V konečném důsledku to znamenalo vrátit se k původním hodnotám ohodnocujícím pravděpodobnosti přechodu, tj. použít stejné hodnoty pro všechny audiokategorie.

5.3.2. HMM klasifikátor pracující s předrozdělenými částmi

Použití HMM klasifikátoru pracujícího s předrozdělenými částmi audiosignálu umožnilo vyvinutí a nasazení do reálného provozu detektoru změn řečníka v audiosignálu [ŽĎÁ05]. Tento detektor, nejnověji založený na *modifikované metodě s adaptivním oknem*, v současné době tvoří nedílnou součást systému pro přepis televizního zpravodajství a umožňuje jeho plnou automatizaci. Jeho implementací a zařazením do procesu automatického přepisu odpadla jedna z nejnamáhavějších činností, jíž je ruční segmentace záznamů. Velkou výhodou této metody je jak její on-line pojetí, tj. minimální zpoždění detekce bodu změny vůči reálnému času, tak i její nízká výpočetní a paměťová náročnost.

Jak již bylo řečeno, HMM klasifikátor pracující s předrozdělenými částmi zpracovává již rozsegmentovaný audiosignál, tj. o každém segmentu můžeme tvrdit, že je z hlediska promlouvající osoby či audioobsahu homogenní. Proto není třeba, aby bylo prováděno časové zarovnání nebo dělení segmentů na menší části, neboť tato činnost je již zajištěna detektorem změn. Jednotlivé segmenty tedy mohou být zařazeny do jedné z kategorií řeč/neřeč jako celek. Odpadá nám tudíž potřeba vzájemného propojení modelů jednotlivých audiotříd a úloha se mění na klasické rozpoznávání izolovaných slov metodou HMM s ergodickými modely. Jednoznačnou výhodou tohoto přístupu je dodržení on-line koncepce v součinnosti s dalšími funkčními bloky kompletního systému pro přepis zpravodajských pořadů.

Konfigurace klasifikátoru je nastavena takovým způsobem, aby docházelo k správné detekci co možná největšího objemu řečových segmentů. Nezapomínejme, že identifikované řečové segmenty končí až v LVCSR bloku, kde každý výpadek řečového segmentu může výrazně ovlivnit úspěšnost rozpoznávacího skóre a srozumitelnost výsledného textového výstupu.

6. Databáze COST278-BN

6.1. Motivace vzniku databáze COST278-BN

Jedním z frekventovaných témat současného výzkumu v oblasti rozpoznávání řeči je i automatický přepis zpravodajských pořadů. Z toho vyplývá i potřeba vhodných dat pro trénování a testování těchto systémů.

Z těchto důvodů bylo vytvořeno mnoho národních databází televizního a rozhlasového zpravodajství. Jako příklad lze uvést korpus amerických zpravodajských pořadů Hub4 vytvořený konsorciem LDC[1] (Linguistic Data Consortium), jehož vznik měl podpořit výzkum a vývoj v oblasti transkripce zpravodajských pořadů. Bohužel, americké zpravodajství je v mnoha aspektech odlišné od evropského, a tudíž není vhodné tento zdroj dat použít pro vývoj „evropského" systému. Daleko závažnějším nedostatkem je však nemožnost vyhodnocování navržených systémů jako celku, tj. vytvoření kompletního textového výstupu zahrnujícího i rozpoznávání spojité řeči (důvodem je závislost rozpoznávače na jazyku dané země). Především tyto skutečnosti vedly několik institucí participujících na projektu COST278 k rozhodnutí vytvořit evropskou databázi televizního zpravodajství (viz společná publikace [VAN04]).

Stanoveným cílem nebylo vytvoření co nejobsáhlejší databáze vhodné k natrénování kompletního systému pro přepis zpravodajství. Naším úkolem bylo spíše vytvoření přiměřeně obsáhlého souboru nahrávek, který by byl vhodný k vzájemnému testování a porovnávání jednotlivých algoritmů mezi členy konsorcia, ať už se jedná o činnost modulů na jazyku nezávislých[2] (segmentace mluvčích, identifikace pohlaví, identifikace řečových segmentů), nebo na jazyku závislých (rozpoznávání řeči, identifikace tématu příspěvků a komplexní úloha automatické transkripce). Společně s databází je distribuován i software pro vyhodnocování dosažených výsledků v jednotlivých dílčích úlohách. Tak je zaručena vzájemná kompatibilita výsledků a lze i objektivně stanovit úspěšnost různých přístupů, které jsou používány partnery projektu v úlohách automatického zpracování řeči.

Podmínkou přistoupení každého nového člena ke konsorciu je dodání tří hodin ručně anotovaných nahrávek národního televizního zpravodajství libovolných komerčních či veřejných TV stanic. V současné době obsahuje celá databáze 31 hodin nahrávek

[1] „Linguistic Data Consortium" je organizace, která podporuje výzkum a vývoj v jazykové oblasti. Bližší info. na http://www.ldc.upenn.edu/Projects/Corpus_Cookbook/transcription/broadcast_speech/english/conventions.html.

[2] Tato nezávislost je pouze relativní. Ať je systém postaven na jakýchkoliv přístupech a algoritmech, různé jazyky při trénování a testování se do určité míry vždy promítnou do konečných výsledků.

pořízených z 13 TV stanic v 9 evropských jazycích. Jmenovitě to je vlámština, portugalština, galeo, čeština, slovenština, slovinština, řečtina, chorvatština a maďarština. Podrobnější statistika je znázorněna v tabulce 6.1 .

	Jazyk	TV stanice	Počet pořadů	Počet klíčových mluvčích	Délka pořadů (min.)	Orig. f_{vz} [kHz]
BE	vlámština	VRT	6	8	162	16
CZ	čeština	ČT1	7	11	187	44,1
GA	galeo	TVG	3	3	225	44,1
GR	řečtina	ERT	3	2	174	22,05
HR	chorvatština	HRT	6	11	201	44,1
HU	maďarština	MTV1, TV2, RTL KLUB	11	10	203	44,1
PT	portugalština	RTP1, RTP2	6	6	211	44,1
SI	slovinština	RTV-SLO1	3	6	182	16
SI2	slovinština	RTV-SLO1	3	4	151	16
SK	slovenština	TA3	9	7	190	44,1

Tab. 6.1: *Informace o jazykové skladbě databáze – tabulka ukazuje jména TV stanic a jejich komerční/veřejný status, počet pořadů, počet klíčových mluvčích v pořadech, délku nahraných dat v minutách a originální vzorkovací frekvenci nahrávek pořadů*

Soubory, které nebyly vzorkovány požadovanou frekvencí 16 kHz přímo při nahrávání, byly na tuto vzorkovací frekvenci převedeny později volně šiřitelným programem „sox" s parametry „-r 16000 polyphase". Současně s již zmíněnými audiozáznamy tvoří databázi i video soubory. Pomocí nich je možné zpětně ověřovat/opravovat identitu mluvčích, rozšířit vyvíjené algoritmy i na oblast multi-modálního zpracování řeči a používat tyto video soubory k prezentačním účelům. Z důvodu snížení paměťových nároků vizuální části databáze byl pro archivaci použit Real Media Video formát s rozlišením 352x288 obrazových bodů.

Každá národní databáze je rozdělena na trénovací část (přibližně 2 hodiny) a testovací část (přibližně 1 hodina). V evaluační kampani se později ukázalo, že ne vždy je vhodné dělit nahrávky na testovací a trénovací část a v experimentech jsou zvoleny i jiné varianty uspořádání dat. Všechna data jsou uložena na FTP serveru, který je přístupný všem členům konsorcia. FTP server má následující adresářovou strukturu: /<použití>/<jazyk>/<typ>/, kde <použití> reprezentuje adresář s trénovacími nebo testovacími daty, adresář <jazyk> je reprezentován mezinárodní značkou daného jazyka a <typ> rozděluje data na audio, video a transkripci. Na serveru je kromě řečových dat uložen i software s dokumentací pro tvorbu statistik a vyhodnocování výsledků rozpoznávání, který byl vytvořen na základě vzájemné spolupráce mezi jednotlivými členy konsorcia. Názvy souborů jsou v rámci databáze

jedinečné a identifikují TV stanici, ze které byly pořízeny, datum vysílání a název pořadu. Během evaluačního procesu byly v transkripčních souborech zjištěny drobné chyby přepisu (a je pravděpodobné, že se ještě další chyby objeví). Na základě těchto zkušeností jsou transkripce ukládány do podadresářů, jejichž jména obsahují datum revize. V těchto podadresářích jsou pak uloženy všechny transkripční soubory dostupné k danému datu revize. Tomu je samozřejmě přizpůsobeno i testování, kdy je pokaždé specifikována verze souborů používaných v testech. Podrobnější statistický pohled na databázi lze získat z přílohy 2, kde jsou shrnuty některé základní údaje o kompletní a české skupině dat. Distribuční politika databáze je nastavena takovým způsobem, aby umožňovala vědecko-výzkumné použití databáze všem stávajícím i novým členům, kteří splní podmínku dodání nových dat.

Hlavička

Seznam témat

Seznam mlučích

Transkripce

```
<?xml version="1.0" encoding="CP1250"?>
<!DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="(unknown)" audio_filename="03.05.29_Prima.wav" version="12" version_date="030714">
<Topics>
  <Topic id="to1" desc="Jaroslav Tvrdík nabídl svůj odchod z vedení ministerstva obrany"/>
  <Topic id="to2" desc="Vláda padne, pokud v referendu neschválíme vstup do EU"/>
</Topics>
<Speakers>
  <Speaker id="spk1" name="Martina Kociánová" check="no" type="female" dialect="native" accent="" scope="local"/>
  <Speaker id="spk2" name="03.05.29P - Muž 1" check="no" type="male" dialect="native" accent="" scope="local"/>
  <Speaker id="spk3" name="Pavel Šuba" check="no" type="male" dialect="native" accent="" scope="local"/>
</Speakers>
<Episode program="" air_date="">
(...)
  <Section type="nontrans" startTime="18.503" endTime="33.958">
    <Turn startTime="18.503" endTime="33.958" mode="planned" fidelity="high" channel="studio">
      <Sync time="18.503"/>
      <Event desc="jingle" type="noise" extent="begin"/>
      <Event desc="jingle" type="noise" extent="end"/>
    </Turn>
  </Section>
  <Section type="filler" startTime="33.958" endTime="55.683">
    <Turn speaker="spk1" startTime="33.958" endTime="49.102" mode="planned" fidelity="medium" channel="studio">
      <Sync time="33.958"/>
      ^Jaroslav ^Tvrdík nabídl brzy ráno svůj odchod z vedení ministerstva obrany.
      <Sync time="38.473"/>
      Pokud v referendu neschválíme vstup do ^Evropské ^Unie tak současná vláda nejspíš padne.
    </Turn>
  </Section>
(...)
 </Episode>
</Trans>
```

Obr. 6.1: *Ukázka formátu TRS*

6.2. Konverze do NIST STM formátu

Nativním formátem pro ukládání výsledků segmentace jsou pro program Transcriber (bližší popis je možno nalézt v příloze 3), ve kterém byly všechny ruční přepisy pořízeny, *.trs soubory. Ukázka přepisu uloženého ve zmíněném formátu je na obrázku 6.1. Transkripční formát (Transcriber Transcription format – TRS), který je používán programem Transcriber, je podobný univerzálnímu transkripčnímu formátu (Universal Test

Framework – UTF) používanému NISTem (National Institute of Standards and Technology) s nepatrnými rozdíly ve zpracování charakteristik a pojmenování entit mluvčího, které jsou pouze nepovinně obsaženy v UTF. Aby byla zachována kompatibilita s Hub-4 testovacími specifikacemi pro vyhodnocování rozpoznávacího skóre řeči, bylo třeba zajistit konverzi atributů mluvčího, módu řeči a přenosového kanálu z TRS/UTF do NIST podmínek pozadí „Focus conditions“[1] definovaných v (Pallett, 2002). Konverzní algoritmus převádí dialekt mluvčího (rodilý či nerodilý mluvčí), techniku řeči (plánovanou či spontánní), šířku přenosového pásma (telefonní či studiovou), kvalitu záznamu (vysoká, střední či nízká) a rozdílné audiopodmínky pozadí (hudba, řeč, hluk a jiné) stanovené TRS formátem do sedmi „Focus conditions“. Převod některých z výše zmíněných řečových atributů (technika řeči, šířka pásma a kvalita audiozáznamu) na odpovídající „Focus conditions“ je logický a pevně daný. Zvláštní pozornost je třeba věnovat úpravám dialektu mluvčího, odlišným akustickým podmínkám pozadí a při výskytu překrývající se řeči.

Neřečové segmenty

Podbarvení hudbou

Transkripce s vyznačením mluvčích, akustických podmínek a časů v sériovém časovém proudu

```
03.05.29_Prima.wav 1 "no_name" 0.000 18.503 <o,f3,unknown>
03.05.29_Prima.wav 1 "no_name" 18.503 33.958 <o,f0,unknown> [jingle-] [-jingle]
03.05.29_Prima.wav 1 "Martina Kociánová" 33.958 38.473 <o,f3,female> ^Jaroslav ^Tvrdík nabídl brzy ráno svůj
odchod z vedení ministerstva obrany.
03.05.29_Prima.wav 1 "Martina Kociánová" 38.473 43.653 <o,f3,female> Pokud v referendu neschválíme vstup do
^Evropské ^Unie tak současná vláda nejspíš padne.
03.05.29_Prima.wav 1 "Martina Kociánová" 43.653 49.102 <o,f3,female> [i] Obětí chladnokrevné vraždy se dnes ráno
stal třicetiletý řidič z ^Lužné u ^Rakovníka.
03.05.29_Prima.wav 1 "no_name" 49.102 51.854 <o,f3,unknown>
03.05.29_Prima.wav 1 "Martina Kociánová" 51.854 54.290 <o,fx,female> Vítejte u nás připravili jsme zpravodajský
deník.
03.05.29_Prima.wav 1 "03.05.29P - Muž 1" 54.290 55.683 <o,f1,male> [bb] Hezký večer.
03.05.29_Prima.wav 1 "Martina Kociánová" 55.683 59.971 <o,f0,female> [i] ^Jaroslav ^Tvrdík nabídl svůj odchod z
vedení ministerstva obrany.
03.05.29_Prima.wav 1 "Martina Kociánová" 59.971 64.485 <o,f0,female> [i] Kvůli reformě veřejných financí nedostane
jeho resort tolik peněz, kolik by potřeboval.
03.05.29_Prima.wav 1 "Martina Kociánová" 64.485 67.090 <o,f0,female> [i] Armádní reforma je od dnešního dne
zmrazena,
03.05.29_Prima.wav 1 "Martina Kociánová" 67.090 70.000 <o,f0,female> [i] Tvrdík ale možná nakonec ve funkci zůstane.
03.05.29_Prima.wav 1 "Pavel Šuba" 70.000 81.486 <o,fx,male> Do armády měla jít každý rok dvě celá dvě desetiny
procenta ~HDP, usnesla se na tom vláda a slíbili jsme to alianci. Protože chce ale stát ušetřit, má ročně
armáda dostat o čtyři miliardy méně. Bez nich ^Tvrdík ministrem zůstat nechce.
03.05.29_Prima.wav 1 "Jaroslav Tvrdík" 81.486 86.768 <o,f4,male> Předal jsem panu [e] premiérovi rezignaci podle
příslušného článku Ústavy ^České ^Republiky.
03.05.29_Prima.wav 1 "Vladimír Špidla" 86.768 93.022 <o,f4,male> Po krátkých konzultacích se rozhodnu do úterý, zda
podmíněnou rezignaci přijmu, nebo +nepřijdu.
03.05.29_Prima.wav 1 "Pavel Šuba" 93.022 98.780 <o,f4,male> Za ^Tvrdíka se postavila ~ODS, na reformě armády se
prý šetřit nedá. Otázkou však je, jestli ^Tvrdík skutečně skončí.
03.05.29_Prima.wav 1 "Petr Nečas" 98.780 106.524 <o,f4,male> Zda to není pouze hra, na základě které vláda a ministr
obrany nechce ztratit tvář před vnitřkem armády.
03.05.29_Prima.wav 1 "Bohuslav Sobotka" 106.524 110.117 <o,f4,male> Může to být určitá forma, forma nátlaku.
03.05.29_Prima.wav 1 "Cyril Svoboda" 110.117 113.388 <o,f4,male> Já myslím, že pan ministr ^Tvrdík setrvá ve funkci.
```

Obr. 6.2: *Ukázkový STM soubor s totožným textovým přepisem, který je i na obrázku 6.1 (zvýrazněné pasáže na obou ukázkách si navzájem odpovídají)*

Na základě konverzního algoritmu byl vyvinut softwarový nástroj pro parsing uložených XML (Extensible Markup Language) dat z prostředí Transcriberu a jejich

[1] Tento termín je natolik specifický a zavedený v literatuře (je možné setkat se i s termínem „F-conditions“), že jsme se z důvodu srozumitelnosti rozhodli nenahrazovat ho českým ekvivalentem.

převod na NIST STM (Sclite Segment Time Mark) formát. Vytvořený konverzní nástroj je mírně odlišný od konvertoru implementovaného přímo v programu Transcriber. Formát STM rozlišuje časové intervaly záznamu zároveň s informacemi o mluvčím, „Focus conditions" a textovými přepisy jednotlivých intervalů. STM formát byl stanoven jako implicitní formát v referenčních přepisech používaných při vývoji systému pro automatický přepis zpravodajství. Na tomto formátu je zároveň postavena i evropská databáze televizního zpravodajství COST278, na jejímž vývoji jsme se aktivně podíleli.

Veškeré textové přepisy vytvořené v prostředí programu Transcriber byly uloženy ve formátu XML používajícím odlišné národní znakové sady. Při pokusech o využití implicitní STM konverze, kterou nám umožňuje Transcriber, jsme velmi často naráželi na problémy s chybným exportem určitých znaků (důvodem bylo selhání podpory některých národních kódovaných stránek). Vytvořili jsme proto vlastní konverzní nástroj „TRS2STM" pro správnou konverzi Transcriber XML souborů (obr. 6.1) do NIST STM souborů (stejné jaké používá „American Hub4" viz obr. 6.2). Tento nástroj je přístupný spolu s databází COST278. Vytvoření vlastního konverzního nástroje navíc umožnilo provádět plně automatickou konverzi bez nutnosti spouštět program Transcriber a ručně otevírat jednotlivé soubory. V okamžiku, kdy jsme měli plně funkční a spolehlivý nástroj pro konverzi do STM, bylo potvrzeno, že referenčním vstupním formátem pro vyhodnocování všech dosažených výsledků v rámci celé BN skupiny bude formát STM. Abychom se vyhnuli případným problémům s různými znakovými stránkami, konverzní utilita ukládá STM soubory s podporou UNICODE standardu.

F-conditions – informace o audiokvalitě promluvy
F0 – normální řeč (většinou čtená, nebo jinak předem připravená)
F1 – spontánní řeč (odpovědi na otázky, ankety – předem nepřipravená)
F2 – telefonní řeč
F3 – hudba na pozadí
F4 – promluva obsahuje šum, nekvalitní řeč, překrývající se řeč
F5 – česky hovořící cizinci
Fx – jakékoliv další případy, případně směs předchozích

Tab. 6.2: *Výčet značek „F-conditions" popisujících audiokvalitu promluvy*

6.3. Společné značky, originální značky a pravidla databáze

6.3.1. Společná pravidla a značky

Jak již bylo předesláno v předchozích odstavcích, primárním zdrojem pravidel při přepisu zpravodajských pořadů se staly konvence stanovené LDC konsorciem pro databázi HUB4. Některé nejasnosti musely být předem vyřešeny, hlavně z důvodů separace autorů anotací, kteří museli pracovat nezávisle jeden na druhém na různých místech bez přímého kontaktu a konzultací. Tento důvod vedl k potřebě prodiskutovat některé sporné úseky, sjednotit pravidla a tím i zlepšit přesnost anotací. Tomuto úkolu byla věnována velká pozornost, která měla zaručit co možná nejmenší odchylky jednotlivých anotátorů od dohodnutého standardu přepisu. Jako příklad lze uvést zkušenosti portugalských kolegů. Jejich nahrávky byly pořízeny a anotovány v průběhu delšího časového úseku rozdílnými anotátory. Důsledkem byly nekonzistentní přepisy lišící se podstatným způsobem v umístění a četnosti dělících značek mezi pauzami promluv, nestejnorodé značky citoslovcí, ruchů a podobně. Tyto nahrávky musely být nakonec znovu překontrolovány a upraveny do jednotného formátu respektujícího stanovená pravidla. Úprava pravidel pro segmentaci a transkripci byla ukončena během workshopu organizovaného lisabonskou institucí INESC-ID, kde se sešlo šest ze sedmi zakládajících členů BN skupiny zabývající se transkripcí zpravodajských pořadů pro databázi COST278. Všichni anotátoři, kteří se v pozdější době podíleli na rozšíření databáze svými nahrávkami, byli se stanovenými pravidly seznámeni a řídili se jimi.

Tento modifikovaný soubor konvencí je určen k řešení některých otevřených otázek, které nepokrývají konvence LDC určené pro anglickou databázi HUB4. Během lisabonského workshopu byl sestaven soubor bodů, které bylo třeba upřesnit a sjednotit:

- atributy charakterizující kvalitu nahrávky promluvy mluvčího a přenosovou cestu,
- segmentace dlouhých vět na menší úseky,
- pauzy uvnitř promluv jednotlivých mluvčích,
- způsob, jakým budou pojmenovávány sekce (zpravodajské příspěvky),
- identifikace znělek („jingle" segmenty),
- způsob, jakým budou přepisovány promluvy v cizích jazycích,
- značky pro přepis různých citoslovcí, nestandardních slov a zkratek.

Hlavními atributy popisujícími audiokvalitu promluvy mluvčího je druh zařízení, kterým byla promluva zaznamenána (studiová či telefonní nahrávka) a zvuková věrnost,

s jakou byla nahrávka daným zařízením zaznamenána. Zvuková věrnost nahrávky vyjádřená pojmy nízká/střední/vysoká má odlišné významy pro nahrávku pořízenou ve studiové či telefonní kvalitě. Studiová kvalita s vysokou zvukovou věrností se využívá pouze pro nahrávky pořízené přímo ve studiu. K této situaci dochází obvykle tehdy, jestliže hovoří klíčoví mluvčí nebo je ze záznamu přehrávána ve studiu komentovaná reportáž. Střední zvukovou věrností jsou označovány ty nahrávky, kdy je komentář reportéra zachycen v terénu (příkladem může být žurnalista provádějící rozhovor přímo na ulici). Nízkou zvukovou věrností označujeme takové audiozáznamy, které zahrnují hluk přenosového kanálu. Téměř ve všech případech mohou být promluvy identifikovány jako studiové. Existuje pouze jedna výjimka – telefonní záznam. V případě hovoru přenášeného po telefonní lince jsou vysokou zvukovou věrností označeny ty nahrávky, které jsou velmi dobře srozumitelné a na jejichž pozadí se nevyskytuje téměř žádný hluk. Pokud jsou nahrávky stále ještě srozumitelné, ale nejsou již tak kvalitní kvůli rušení hlukem (šumem) v přenosovém kanálu, označujeme je střední zvukovou věrností. Nízká kvalita zvukové věrnosti je užívána pouze pro nejvíce znehodnocené případy, kdy je telefonní hovor téměř nesrozumitelný kvůli hluku (šumu) v přenosovém kanálu (viz tabulka 6.3).

		Přenosová cesta	
		Studio *[Šířka pásma > 4kHz]*	**Telefon** *[4kHz Šířka pásma]*
Kvalita	**Nízká**	rušený signál	nesrozumitelné
	Střední	venkovní nahrávky	rušený signál
	Vysoká	studiové nahrávky	kvalitní nahrávky

Tab. 6.3: *Atributy popisující audiokvalitu zaznamenané promluvy*

Promluva by neměla být příliš dlouhá, protože každý nádech mluvčího může být považován za potenciální konec řeči. Abychom zamezili nestandardnímu chování jednotlivých anotátorů (pro rozdělení promluvy mluvčího může někomu připadat dostatečně dlouhá pauza trvající 0,2 sekundy, někomu nedostačuje ani celá sekunda), bylo nutné stanovit jednoznačná pravidla pro dělení promluv. Tato pravidla jsou následující: v případě, že je ticho kratší než 0,5 sekundy, není dělící bod vůbec zaznamenán. Pokud se délka pauzy pohybuje mezi 0,5–1,5 sekundou, je pomlka zaznamenána uprostřed tohoto úseku. Pro odmlčení delší než 1.5 sekundy jsou zapsány dvě časové značky přesně vymezující začátek a konec úseku ticha.

Přehledněji si dělení mezi pauzami můžeme znázornit následujícím popisem. Pokud je délka ticha v promluvě pocházející od jednoho mluvčího:

- kratší než ½ sekundy:

  => PONECHÁNÍ PROMLUVY V CELKU,

- delší než ½ sekundy a kratší než 1+ ½ sekundy:

  => ROZDĚLENÍ PROMLUVY V PŮLI PAUZY,

- delší než 1+ ½ sekundy:

  => VLOŽENÍ DVOU ČASOVÝCH ZNAČEK, KTERÉ BUDOU VYMEZOVAT ZAČÁTEK A KONEC ÚSEKU TICHA.

Zpravodajský pořad je při anotaci rozdělen na menší bloky, které jsou zařazeny do jedné z následujících kategorií:

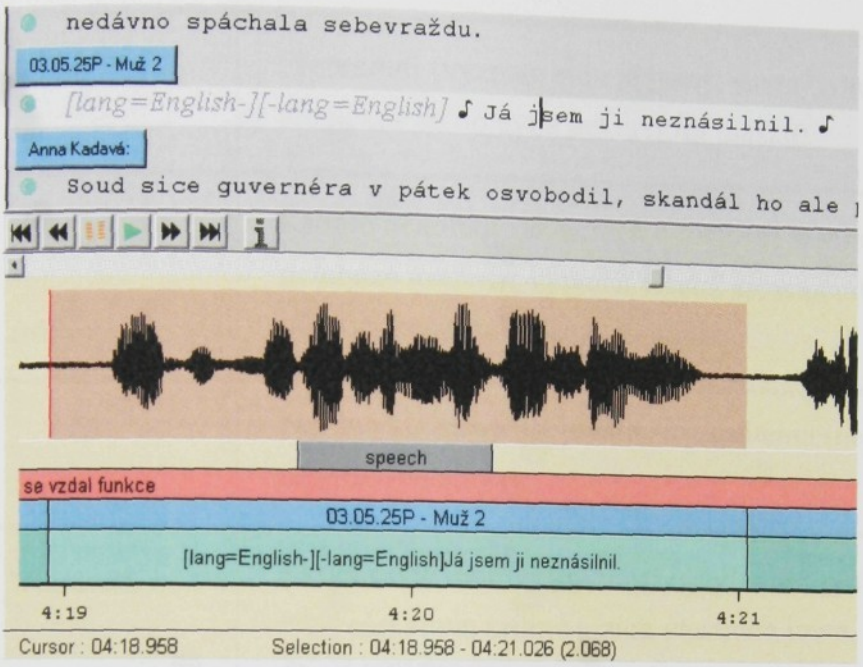
- reporty – „reports" (zpravodajské příspěvky, komentované reportáže, příběhy),
- upoutávky – „fillers" (titulky, krátké upoutávky na nadcházející vysílání),
- úseky kompletně vyňaté z přepisu – „nontrans" (komerční segmenty, zahajovací a ukončovací znělky).

Všechny znělky (jingly) jsou ponechány bez textového přepisu a je jim přidělen atribut „noise event tag", který danou oblast označí jako neřečový úsek a vyjme ji z evaluačního procesu. Jestliže televizní stanice používá rozdílné jingly na začátku a na konci pořadu, každému jinglu je přidělena dodatková přípona označující, zda jde o znělku počátku či konce segmentu.

Projevy v cizí řeči jsou označeny značkou „foreign language tag" a nejsou překládány. Stejně jako jingly jsou i tyto úseky vyřazeny z evaluačních experimentů. Příkladem přepisu cizí promluvy může být i obr. 6.3. Z tohoto obrázku je patrné, jak je někdy složité cizí promluvu přepsat a časově zarovnat, pokud se prolíná s nativním jazykem. Tento konkrétní příklad je složen z promluvy v českém a anglickém jazyce. Originální anglický jazyk je po chvíli potlačen do pozadí a je přes něj puštěn český simultánní překlad. Chvíli jsou oba jazyky přehrávány současně (v Transcriberu je tento případ vyznačen šedě jako řeč na pozadí) a celá promluva je dokončena čistým českým překladem. Právě v těchto sporných případech (tato promluva by se dala přepsat nejméně třemi dalšími způsoby) je velice důležité dodržovat dohodnutá pravidla přepisu.

Dalším jevem, i když ne tak častým, je výskyt podobně hlasité simultánní řeči dvou a více osob v jednom okamžiku. V takovém případě nelze označit jednu promluvu jako řeč na pozadí a přepsat pouze tu z nich, která je dominantní. V Transcriberu tuto situaci řešíme

přidáním nové změny mluvčího a za řečníka dosadíme obě promlouvající osoby. Při konverzi anotované nahrávky do STM formátu je simultánní řeč označena jako „inter-segment-gap", čímž je vyřazena z testování i trénování.



Obr. 6.3: *Přepis promluvy obsahující cizí jazyk v programu Transcriber*

Jak již bylo předesláno dříve, pro všechny jazyky COST278-BN databáze existuje pevně stanovená množina značek a anotačních postupů popisující citoslovce, zkratky, ruchy a neobvyklé fráze. V tabulkách 6.4 a 6.5 se pokusíme tyto značky a pravidla přehledně shrnout a vysvětlit jejich význam v kontextu k anotovanému zpravodajství.

Speciální značky a jejich význam

Příklad použití značek	Popis významu
@NATO *@DARPA* *@AIDS*	Zkratky názvů institucí, slovních spojení, medicínských a jiných termínů čtené dohromady.
~FBI *~CEO* *~YMCA*	Zkratky názvů, jejichž písmena čteme odděleně.
příležitost- *-stoupení*	Nedokončená slova a slova s nesrozumitelným nebo chybějícím začátkem.

^Homer ~L ^Simpson *^Praha* *^Petr*	Vlastní jména.
+pravděpodobně *+včerejší*	Přeřeknutí (význam slov je jasný, slova byla dokončena, ale někde se objevila nekorektní hláska).
%ehm %hmm %pp (zvuk rtů)	Kromě několika základních citoslovcí si v rámci národní databáze můžeme vytvořit vlastní slovník citoslovcí, pomocí kterého budeme anotovat.
*všici *začla	Neobvyklá slova ale přesto srozumitelná (nejčastěji nespisovná slova).
^^Rafjanii ^Agrawal **summit **schmates	Slova s odlišnou výslovností (především přejatá slova), ** – cizí slova kromě cizích vlastních jmen, ^^ – cizí vlastní jméno (jinak se čte a jinak se píše).

Tab. 6.4: *Přehled speciálních značek, příklady jejich použití*

Interpunkční znaménka

Při anotaci záznamu je dovoleno používat pouze čtyři interpunkční, nebo-li členící znaménka v jejich pravém významu (tečka, čárka, uvozovky a pomlčka). Ostatním znaménkům (středník, otazník, vykřičník, dvojtečka, tři tečky, závorky) je přidělena speciální řídící funkce, nebo nejsou povoleny v STM formátu, do kterého jsou všechny přepsané pořady zkonvertovány.

Upřesnění anotačních pravidel v některých sporných bodech

Otázka	Popis postupu
Jak naložit s identitou mluvčích, pokud nejsme schopni ji sami určit?	
↳	Zavedeme mluvčího jako nového. Každému novému mluvčímu je přiděleno jednoznačné jméno skládající se ze jména pořadu, TV kanálu, data vysílání a pohlaví s pořadovým číslem výskytu v daném pořadu.

Jak přepsat přeřeknutí mluvčího následované opravou (zopakováním téhož slova)?
↳ Chybné slovo označit odpovídající značkou (viz tabulka 6.4) a pokračovat dále v přepisu i s případnou korekcí mluvčího.
Co provést v případě, když skupina osob skanduje určité heslo?
↳ Vybrat ze skupiny skandujících osob konkrétního mluvčího (pokud takový existuje) a promluvě nastavit atribut „řeč na pozadí“.
Jak postupovat v případě, kdy je autentická promluva cizího řečníka simultánně překládána do češtiny (český překlad začíná téměř vždy s určitým zpožděním)?
↳ V době, kdy hovoří pouze cizinec označit řeč jako cizí. Jakmile začne český simultánní překlad, je třeba ukončit cizí řeč a nastavit atribut „řeč na pozadí“. Dále již standardně přepisujeme českou promluvu.

Tab. 6.5: *Ukázky řešení přepisu v několika sporných bodech*

6.3.2. Originální značky a anotační pravidla české podskupiny COST278-BN databáze

Originální značky a anotační pravidla doplňují pevně stanovené značky z minulé kapitoly, aby bylo možné podchytit specifika jednotlivých jazyků. Tyto specifické značky jsou omezeny pouze jednou podmínkou – je jich třeba užívat systematicky a konzistentně ve všech přepisech konkrétního jazyka.

V české podskupině COST278-BN databáze používáme pouze omezený počet značek. Konkrétně to jsou značky *[i]*, *[e]*, *[n]*, *[bb]*, *[rire]*, *[b]*, *[mic]*. Ostatní zvukové události odpovídající značkám z této skupiny se v českých nahrávkách neobjevují, a proto ani nebyly při anotaci použity.

Speciální značky z tabulky 6.4 rozšiřuje znak „ | “, který je určen pro uvozování neexistujících slov, která byla vyslovena mimo kontext a mají ve větě vysloveně rušivý vliv (ať již se jedná o hledisko slovosledu, nebo srozumitelnosti). Jako příklad může sloužit následující věta:

%ehm já nevím asi |des|, pět klubů z %ehm ~ODS.

Použití těchto značek se ukázalo potřebným ve fázi rozpoznávání spojité řeči na rozsegmentovaných částech, kde takto označená slova (části slov) jsou brána jako

automaticky chybná. Proto není chyba při rozpoznávání tohoto slova počítána do konečného rozpoznávacího skóre.

6.4. Vyhodnocovací software

Již dříve byl popsán STM formát, který je v rámci COST278-BN databáze používán k uložení referenčních transkripcí. Formát je podrobně popsán na webových stránkách organizací NIST a LDC. Vyhodnocovací software, který je produktem společného úsilí členů konsorcia, je používán v případě společných experimentů s databází COST278-BN. Je to soubor programů pracujících s referenčními STM soubory a s generovanými výstupy identifikace. Tyto výstupy mají strukturu zjednodušeného STM formátu.

Vygenerované soubory (výstupy rozpoznávání) by měly v jednom řádku obsahovat následující části v daném pořadí oddělené tabulátory (případně mezerami):

- počáteční čas segmentu (v milisekundách),
- konečný čas segmentu (v milisekundách),
- identifikace segmentu na řeč/neřeč (speech, nonspeech nebo unknown),
- číslo identifikovaného clusteru (číslo typu integer, -1 znamená neznámý),
- pohlaví mluvčího (male, female nebo unknown),
- identifikovaný druh pozadí (clean, noise, music nebo unknown),
- šířku přenosového pásma (broad, telephone nebo unknown),
- identifikovanou techniku řeči (read, spontaneous nebo unknown).

Neřečové segmenty jsou v kategoriích pohlaví mluvčího, šířka přenosového pásma a technika řeči vždy označovány jako „unknown". Není ovšem nutné vytvořit pokaždé kompletní záznam obsahující všech osm položek. Například:

- chceme provádět pouze segmentaci audiozáznamu dle změn mluvčích a audiopozadí, stačí vygenerovat pouze první dvě časové značky reprezentující začátek a konec segmentu,
- pokud nebudeme provádět identifikaci clusterů ani identifikaci pohlaví, ale zároveň budeme identifikovat druh pozadí segmentů, je nutné vyplnit prvních šest položek s tím, že jako číslo clusteru budeme zadávat -1 a pohlaví mluvčího bude „unknown".

Vyhodnocovací software samozřejmě obsahuje jednoduchou kontrolu syntaxe, která zaručuje shodný počet položek v každém sloupci. Pokud tomu tak není, vyhodnocení není provedeno a program je ukončen.

6.4.1. Metody vyhodnocování výsledků

V úlohách identifikace změn mluvčích a audiopozadí nebo při clusterování mluvčích je potřebné provádět vyhodnocování výsledků formou obousměrného hledání nejbližšího souseda mezi referenčními a vypočtenými body změn. Při identifikaci pohlaví (řeči/neřeči) máme situaci o něco jednodušší. Rozdělíme-li nahrávku opět (stejně jako v kapitole 4.6.1 a 4.6.2) na mikrosegmenty o určité velikosti (100 ms), stačí odečíst rozdíly ve vypočtených a referenčních mikrosegmentech. Uveďme si jako příklad identifikaci pohlaví, konkrétně úspěšnost rozpoznání mužského pohlaví. Označíme-li počet správně nalezených mužských mikrosegmentů n_{mk}, celkový počet referenčních mužských mikrosegmentů n_{mr}, celkový počet identifikovaných mužských mikrosegmentů n_{mc}, pak

$$RCL = \frac{n_{mk}}{n_{mr}} \times 100\%, \tag{6.1}$$

$$PRC = \frac{n_{mk}}{n_{mc}} \times 100\%, \tag{6.2}$$

$$F = \frac{2 \cdot PRC \cdot RCL}{PRC + RCL}. \tag{6.3}$$

Míra *RCL* se nazývá *recall* a značí procento správně identifikovaných mužských mikrosegmentů ze všech hledaných (referenčních) mužských mikrosegmentů. V mezním případě, pokud bychom naprosto všechny mikrosegmenty identifikovali jako mužské, dosáhli bychom ideálních 100 %. To ovšem znamená, že potřebujeme další míru, která nám poskytne informace o efektivnosti identifikace. Z tohoto důvodu se navíc používá míra zvaná *precision* označovaná jako *PRC*, která vyjadřuje procento správně nalezených mužských mikrosegmentů ze všech nalezených mužských mikrosegmentů. Tyto dvě míry jsou protichůdné, tj. roste-li jedna, klesá druhá a naopak. Avšak ani jedna z těchto měr nemá lokální maximum, a proto nejsou vhodné jako kritéria pro trénování. Toto je důvodem zavedení míry zvané *F-measure* (*F*), která tuto podmínku splňuje. Posledním hlediskem vyhodnocení kvality identifikátoru je míra *accuracy* (*ACC*). Ta jedním údajem popisuje úspěšnost identifikace celého systému pro identifikaci pohlaví (řeči/neřeči), a je proto použita pro porovnání jednotlivých systémů v evaluační kampani. Zavedeme-li shodným způsobem jako v předchozím textu počet korektně identifikovaných ženských mikrosegmentů $n_{žk}$, celkový počet referenčních ženských mikrosegmentů $n_{žr}$, celkový počet

mikrosegmentů prohlášených za ženské $n_{žc}$, pak dostaneme vztah pro výpočet míry accuracy:

$$ACC = \frac{n_{mk} + n_{žk}}{n_{mc} + n_{žc}} \times 100\%. \tag{6.4}$$

6.5. Experimenty s COST278-BN databází

Jak již bylo zmíněno dříve, česká národní podskupina COST278-BN databáze je složena ze tří hodin záznamu televizního zpravodajství veřejnoprávního kanálu ČT1. To ovšem nejsou všechna data, která budou v této práci využita pro vývoj a testování navržených algoritmů a metod. Současně se třemi hodinami oficiálně zařazenými do databáze bylo nahráno a ručně přepsáno dalších 181 minut (rozdělených z hlediska mluvčích do 991 homogenních promluv) televizního zpravodajství (záznamy z TV kanálů Prima, Nova a ČT1). Tyto nahrávky jsou používány jako vývojová testovací data našeho systému pro automatický přepis audiozáznamů. Protože výsledky dosažené při přepisu audiozáznamů (použití kompletního řetězce modulů), budou jednotlivě i globálně diskutovány v kapitole 7, uvedeme v této části pouze výsledky získané v rámci vyhodnocovací kampaně COST278 BN pořádané začátkem roku 2005. Podrobné výsledky lze nalézt v [ŽIB05] a v příloze 4 (tabulky P.3 až P.6).

Tato vyhodnocovací kampaň, které se zúčastnilo 8 z 10 stávajících členů konsorcia, si kladla za cíl otestovat algoritmy a přístupy jednotlivých členů konsorcia za co možná nejpodobnějších podmínek (stejná testovací data, vyhodnocovací nástroje a protokoly) a umožnit tak jejich objektivní porovnání. Zúčastněnými stranami byly instituce ELIS (Gent), INESC (Lisbon), TUB (Budapest), TUK (Kosice), TUL (Liberec), LJU (Ljubljana), UMB (Maribor) a UVIGO (Vigo).

Pro testování byly stanoveny čtyři úlohy, z nichž dvě se týkají této práce a budou dále diskutovány. Těmito úlohami jsou klasifikace audiosegmentů na řečové/neřečové a identifikace pohlaví mluvčích u řečových segmentů. Při použití dvou různých přístupů k trénování evaluovaných systémů dostaneme následující čtyři skupiny testů:

- **C1** : trénování na externích datech (individuální sady jednotlivých účastníků, tj. trénovací sady nejsou součástí COST278-BN databáze) a testování na kompletní COST278-BN databázi.
  - C1 + T1 = řeč/neřeč
  - C1 + T3 = identifikace pohlaví

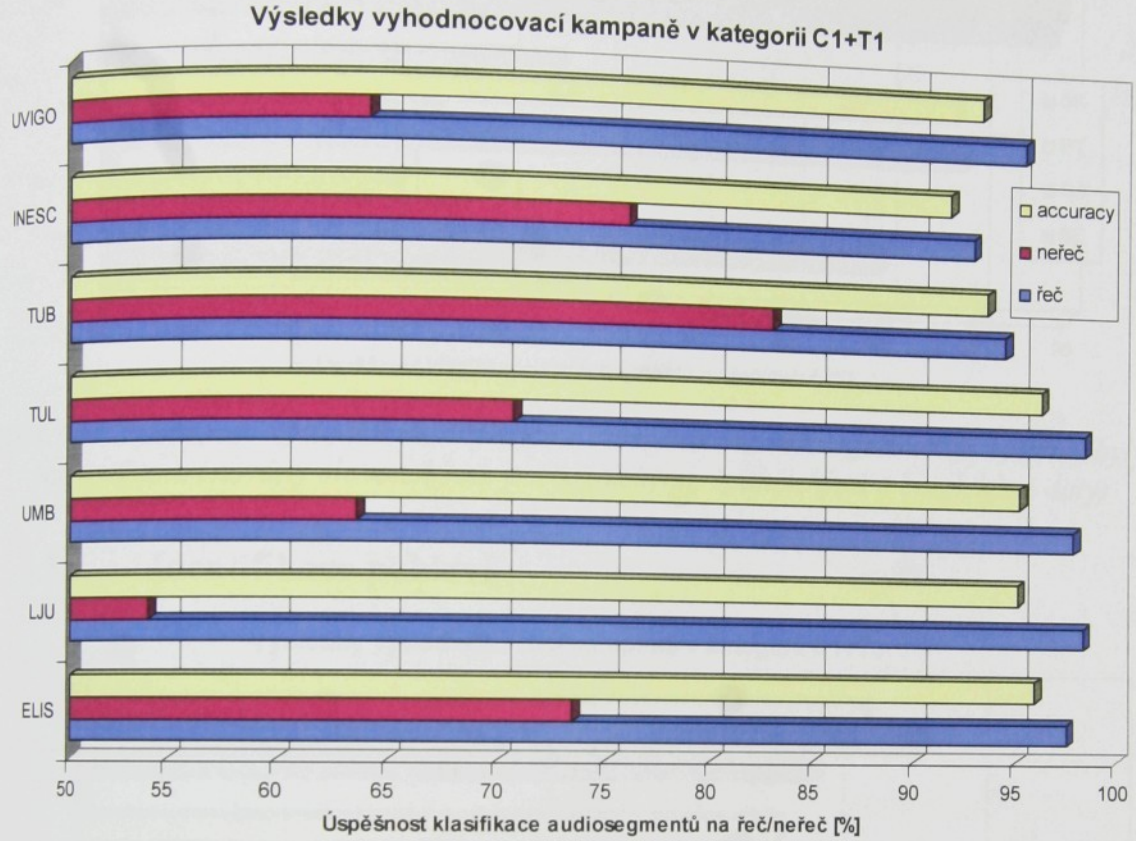
- **C2** : trénování na jedné vybrané národní sadě, testování na všech ostatních národních sadách COST278-BN databáze (daný postup opakovat pro všechny, nebo alespoň pro několik národních sad).
  - C2 + T1 = řeč/neřeč
  - C2 + T3 = identifikace pohlaví

Výhodou C2 jsou naprosto stejné podmínky pro všechny účastníky, a tudíž i zcela srovnatelné výsledky. Na druhou stranu výhodou C1 proti C2 je ten fakt, že délka trénovacích dat není limitována třemi hodinami tak, jak tomu je u C2, čehož většina zúčastněných institucí také využila. Především INESC, který jako jediný používal MLP (vícevrstvé perceptronové sítě), potřeboval pro natrénování svého systému daleko větší porci trénovacích dat než ostatní (46 hodin u detekce řeč/neřeč). Dalším důležitým pravidlem při testování v kategorii C2 je dodržení podmínek trénování pouze jednou určenou národní sadou pro všechny operace předcházející identifikaci řeč/neřeč nebo identifikaci pohlaví. To znamená, že v případě, kdy je před identifikací řeč/neřeč umístěn modul pro segmentaci na základě změn mluvčích a audiopozadí, musí být i tento modul natrénován stejnou národní podskupinou COST27-BN databáze. Pouze tímto způsobem zajistíme vzájemnou kompatibilitu mezi různými identifikačními přístupy (zaměněné pořadí zpracování audiozáznamu či případná integrace několika technik do jednoho bloku).

6.5.1. Identifikace řeč/neřeč

V principu můžeme metody pro identifikaci řečových/neřečových segmentů použité v této kampani rozdělit na dva koncepčně odlišné přístupy. První skupina identifikuje řečové/neřečové segmenty ze spojitého audiozáznamu (ELIS, UMB, TUB a TUK). Druhá skupina, do které patří i námi navržený algoritmus, používá již předsegmentovaný záznam (UVIGO, LJU, INESC a TUL). V českém přepisovacím řetězci toto předzpracování zajišťuje blok detekující změny mluvčích a akustického pozadí. Další rozdíly lze nalézt přímo v použitých metodách rozpoznávání. ELIS, LJU, UMB a UVIGO jako akustické modely používají GMM. TUL, TUK a TUB využívají HMM. Všechny tyto systémy pak pro reprezentaci řečových segmentů používaly modely: *řeč*, *řeč+hudba* a *řeč+ostatní*. Neřečové segmenty byly reprezentovány modely *hudba* a *ostatní*. Jak již bylo zmíněno, zcela odlišný způsob zvolil INESC s jejich MLP [WIL99] počítajícím posteriorní pravděpodobnostní vektor pro každý rámec. Na obrázku 6.4 jsou shrnuty výsledky (podrobněji viz příloha 4 tabulka P.3). O úspěšnosti identifikace nejvíce vypovídá míra *ACC*, případně *error rate* (*error rate* = 100 % – *ACC*), jež zohledňují jak procenta správně identifikovaných segmentů řeč/neřeč, tak i jejich délku vůči kompletnímu záznamu.

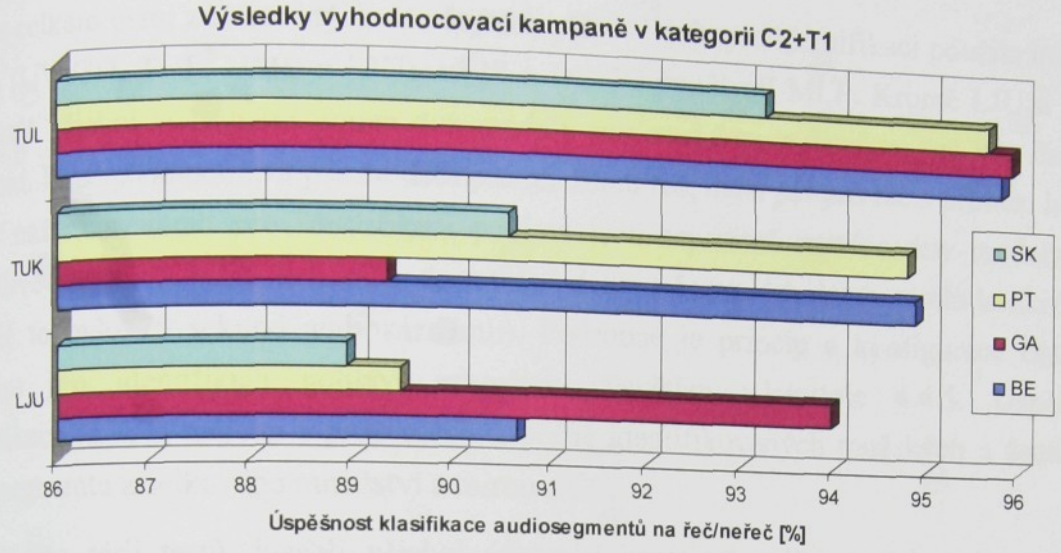
V tomto ohledu si náš systém (TUL) vedl nejlépe s 96,3 % *ACC*, za ním následoval ELIS s 95,9 % a UMB s 95,2 %.



Obr. 6.4: *Porovnání výsledků v kategorii řeč/neřeč (jednotlivé systémy byly natrénovány externími daty – každá zúčastněná instituce použila vlastní)*

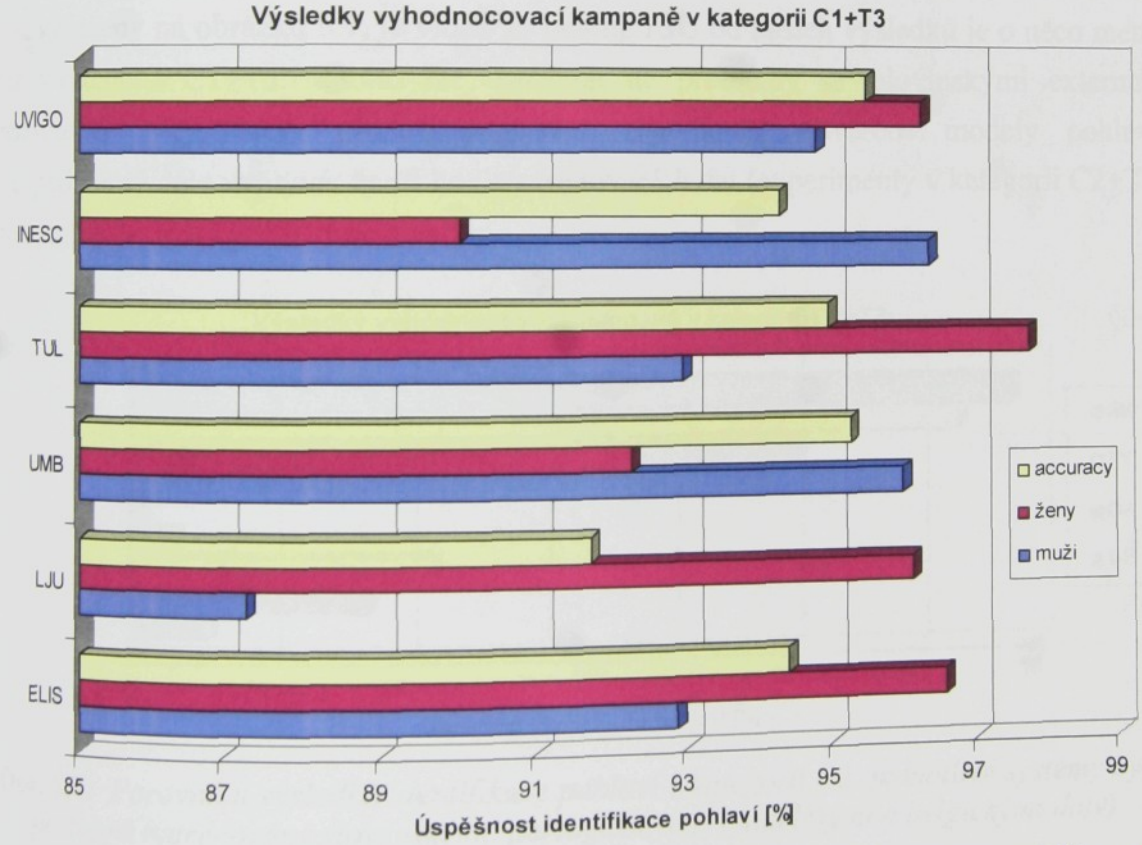
Kategorie experimentů C2+T1 se zúčastnily pouze tři instituce (TUL, TUK a LJU). Výsledky jsou znázorněny na obrázku 6.5. Pokud porovnáme naše a LJU výsledky v kategoriích C1 a C2 můžeme zkonstatovat, že i přes nižší identifikační skóre v kategorii C2 je vzájemný rozdíl přibližně shodný. Z toho můžeme usuzovat, že metody založené na GMM i HMM jsou na jazyku nezávislé. Vyšší *error rate* v kategorii C2 pravděpodobně způsobil nedostatek vhodných trénovacích dat (pro každý test tři hodiny národního BN záznamu) pro vytvoření robustního systému identifikujícího řečové/neřečové segmenty.

Dle dosažených výsledků navíc můžeme usuzovat, že obrácení pořadí při zpracování audiozáznamu (nejprve detekce změn mluvčích a následná identifikace řeč/neřeč a opačně) nemá ve svém důsledku podstatný vliv na výsledky identifikace. To samozřejmě platí pouze v případě, kdy modul zpracovávající audiozáznam jako první nemá výrazně horší výsledky z hlediska časového zarovnání vytvářených segmentů. Takové chyby by pak mohly vnášet nepřesnost do dalšího zpracování a zhoršit tak výsledky následujícího funkčního bloku.



Obr. 6.5: *Porovnání výsledků identifikace v kategorii řeč/neřeč (jednotlivé systémy byly postupně natrénovány slovenskými, portugalskými, španělskými a belgickými daty)*

6.5.2. Identifikace pohlaví

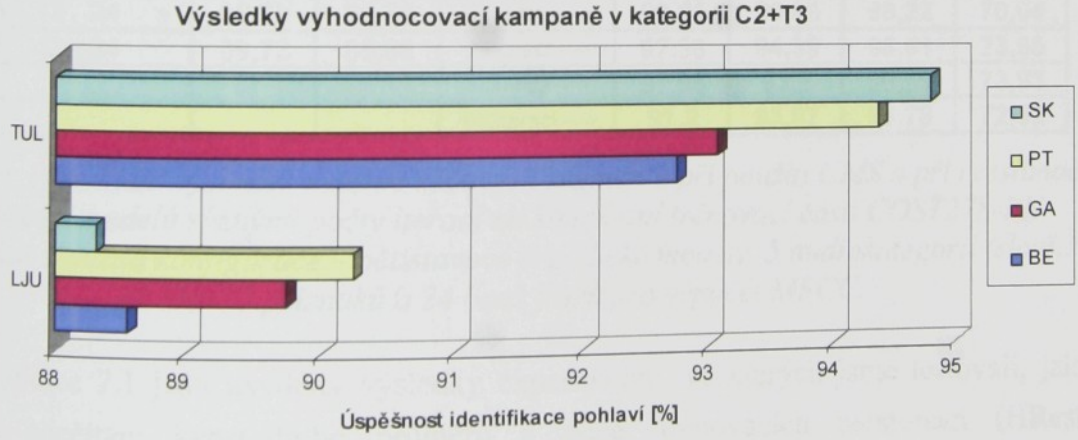


Obr. 6.6: *Srovnání výsledků identifikace pohlaví v kategorii C1+T3 (testy prováděny nad celou databází COST278-BN, systémy byly trénovány externími daty)*

Úlohy C1+T3 se v průběhu vyhodnocovací kampaně zúčastnilo šest institucí (UVIGO, INESC, TUL, UMB, LJU a ELIS) a v případě C2+T3 pouze dvě instituce (TUL a

LJU) z celkem osmi zúčastněných. Ve čtyřech případech je pro identifikaci použita metoda GMM (UVIGO, TUL, UMB a LJU), ELIS a INESC používají MLP. Kromě LJU a TUL používají všichni zúčastnění pouze dva modely pro mužskou a ženskou řeč bez dalšího rozlišení. LJU navíc používá pár modelů pro telefonní řeč, další pár pro řeč s příměsí hudby atd. V našem systému pro identifikaci pohlaví jsou separátně natrénovány modely pro jednotlivé mluvčí splňující kritérium dostatečné délky trénovacích dat (v tomto konkrétním případě to bylo 75 sekund audiozáznamu). Podrobně je princip a konfigurace českého systému pro identifikaci pohlaví mluvčích vysvětlen v kapitole 4.4.5. Úspěšnost identifikace je opět měřena v procentech úspěšně identifikovaných mužských a ženských mikrosegmentů z celkového množství a mírou *ACC*.

V této sérii testů dosáhli všichni účastníci kampaně velice podobných výsledků (výsledky se pohybují v rozmezí 94 % až 95 % *ACC*). Na obrázku 6.6 je vidět, že pouze výsledky LJU jsou o něco horší. Z toho ovšem nelze usuzovat, že by systém založený na MLP pracoval hůře než obvyklejší GMM. Výsledky kolegů z INESC, kteří používají stejný přístup jako LJU, byly srovnatelné s výsledky ostatních. Z výsledků kategorie C2+T3, které jsou vyneseny na obrázku 6.7, je vidět, že odstup LJU od našich výsledků je o něco menší než v kategorii C1+T3. Z toho lze usuzovat na problémy se slovinskými externími trénovacími daty, která mohou negativně ovlivňovat vytvářené modely pohlaví. Z experimentů dále vyplývá, že tři hodiny trénovacích dat (experimenty v kategorii C2+T3) jsou pro identifikaci pohlaví postačující.



Obr. 6.7: *Porovnání výsledků identifikace pohlaví v kategorii C2 (jednotlivé systémy byly postupně natrénovány slovenskými, portugalskými, španělskými a belgickými daty)*

Pokud bychom chtěli krátce shrnout dosažené výsledky, můžeme zkonstatovat značnou jazykovou robustnost všech použitých metod v obou prověřovaných kategoriích.

7. Experimentální výsledky

7.1. Identifikace řečových úseků

Pro vyhodnocování výsledků získaných v rámci identifikace audiosegmentů na řečové a neřečové byly využity postupy a software uvedené v kapitole 6.4. K trénování a testování byla použita COST278-BN databáze v poměru 2:1, tj. dvě hodiny od každé národní sady pro trénovací účely a jedna hodina určená k testování. Pod pojmem *mezinárodní data* budeme v této kapitole uvádět BE, CZ, GA, PT, SI a SK nahrávky, *českými daty* budeme mínit pouze CZ podskupinu mezinárodních dat. Ve všech experimentech byl každý stav ergodického modelu reprezentován 64 mixturami. V průběhu experimentů byl tento počet mixtur určen jako optimální volba jak z hlediska výkonnosti, tak i z hlediska výpočetní náročnosti. Ve všech tabulkách a grafech této kapitoly odpovídají hodnoty uvedené v tabulkách míře recall (RCL) v jednotkách procent.

CMS a trénování HMM

					Česká testovací data					Komplet. test	
Test. konfigurace					Init-CZ			Reest-All		Reest-All	
# stavů	# slov	MFCC	# filtrů	CMS	řeč	neřeč		řeč	neřeč	řeč	neřeč
5	5	12	24	•	**99,77**	**85,76**	25iterací --->	**99,31**	**90,68**	**98,22**	**70,04**
5	5	12	24		**99,73**	**85,66**	25iterací --->	**97,33**	**94,39**	**96,61**	**73,95**
							12iterací --->	**97,95**	**93,2**	**96,79**	**73,83**
							100iterací --->	**97,2**	**93,97**	**96,78**	**72,62**

Tab. 7.1: *Porovnání výsledků detekce řečových segmentů při použití CMS a při reestimaci parametrů modelů různými počty iterací na kompletní trénovací části COST278-BN databáze. Použitá konfigurace – pětistavové ergodické modely, 5 audiokategorií (slov), 12 MFCC příznaků a 24 bank filtrů pro výpočet MFCC.*

V tabulce 7.1 jsou uvedeny výsledky experimentů, ve kterých jsme testovali, jaký vliv má odečítání kepstrálního průměru a počet trénovacích reestimací (HRest) Baum-Welchovým algoritmem na úspěšnost identifikace řečových/neřečových segmentů HMM klasifikátorem pracujícím se spojitým audiosignálem. V případě CMS bylo prokázáno, že použití této metody má jednoznačně pozitivní dopad na výsledné hodnoty, jak ostatně vyplývá i z pravé části tabulky 7.1. Překvapivým zjištěním ovšem je, že tento nárůst míry recall se projeví pouze v případě reestimovaných modelů na datech celé trénovací části databáze. Pokud jsou modely pouze inicializovány Viterbiho algoritmem

(Hlnit) na českých trénovacích datech, ke zlepšení rozpoznávacího skóre téměř nedojde. Důvod je pravděpodobně skryt v různé kvalitě a obsahu nahrávek národních databází. CMS dokáže tyto jevy potlačit a systém se tak stává robustnějším a v jistém směru i univerzálnějším. Tím se dostáváme k otázce optimálního počtu reestimačních iterací, na níž se snažíme nalézt odpověď experimenty vyhodnocenými v pravé části tabulky 7.1. Pokud jsou modely reprezentující jednotlivé audiokategorie inicializovány pouze na českých datech, dokážeme se dostat na velmi vysokou míru recall. Pro 12 MFCC příznaků s provedeným CMS je to až 99,77 % správně detekovaných řečových segmentů vůči 14,24 % nesprávně prohlášeným neřečovým segmentům za řečové. V takovém případě je ovšem možné, že vznikne určitá závislost modelů na daném jazyce a takto postavený klasifikátor nemusí pracovat spolehlivě s jinými národními sadami. Jedním z dalších možných způsobů je inicializace modelů českými trénovacími daty a následná reestimace těchto modelů mezinárodními daty, jak je tomu i v případě výsledků v tabulce 7.1. Výsledky jsou uvedeny pro 12, 25 a 100 reestimačních iterací s mezinárodními trénovacími daty, v prvním sloupci pouze pro česká testovací data, v druhém pro mezinárodní testovací data (6 hodin). Výsledky takto trénovaného systému (inicializace modelů českými daty, reestimace celou trénovací částí databáze) jsou na českých datech o něco horší, ovšem při detekci řečových segmentů v rámci celé COST278-BN databáze se situace mění a tímto přístupem je dosaženo v průměru lepších a také vyváženějších výsledků.

HMM klasifikátor pracující se spojitým nebo rozsegmentovaným signálem

HMM klasifikátor pracující se spojitým signálem						
	1 stavový HMM		3 stavový HMM		5 stavový HMM	
	řeč	neřeč	řeč	neřeč	řeč	neřeč
BE	96,12	83,82	96,69	64,21	98,26	53,77
CZ	96,95	95,85	96,66	94,07	99,31	90,68
GA	95,4	80,4	96,33	68,46	98,26	62,44
PT	96,18	90,38	97,5	86,63	98,19	85,08
SI	94,76	94,44	97,81	94,37	98,4	89,73
SK	96,74	92,18	93,49	89,54	98,42	83,05
All	96,07	89,39	96,36	83,33	98,45	77,89

Tab. 7.2: *Výsledky detekce řeč/neřeč získané HMM klasifikátorem pracujícím se spojitým audiosignálem*

Na tomto místě budou porovnány a zhodnoceny oba implementované přístupy pro identifikaci řečových segmentů z kapitoly 5.3. Jako první vznikl systém pracující se spojitým audiosignálem, neboť v době prvních experimentů s detekcí řeč/neřeč ještě nebyl k dispozici detektor změn mluvčích. Tento klasifikátor tudíž plnil i roli jednoduchého detektoru změn mluvčích a audiopozadí. Jak již bylo výše uvedeno, body změn nalezené

tímto postupem byly v mnoha případech nepřesně umístěny, tudíž docházelo například k rozdělení promluvy uprostřed slova. To mělo za následek snížení rozpoznávacího skóre v modulu LVCSR. I přes tento handicap však klasifikátor dosahoval dobrých výsledků při odhalování řečových segmentů. Tyto výsledky jsou pro jednostavové, třístavové a pětistavové ergodické HMM uvedeny v tabulce 7.2. V posledním řádku jsou pak uvedeny výsledky vyjadřující míry recall řečových a neřečových segmentů všech šesti národních testovacích sad dohromady.

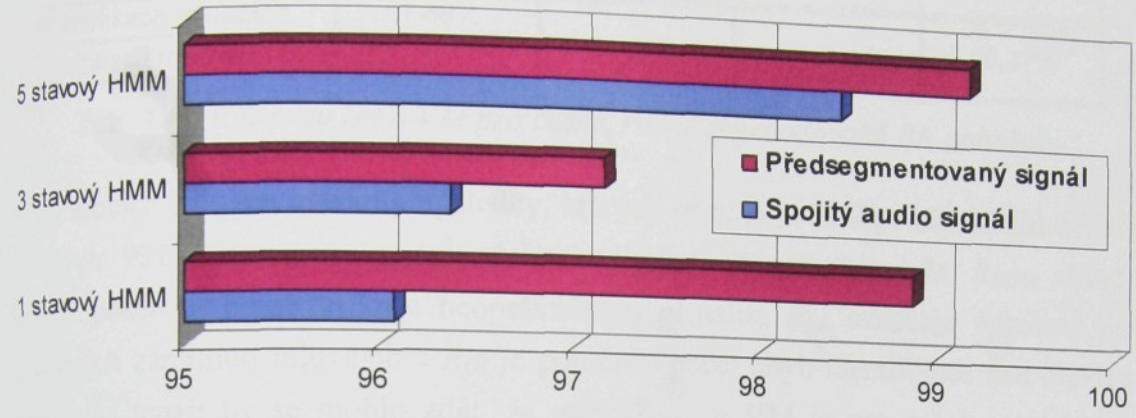
Tabulka 7.3 shrnuje výsledky získané druhou implementovanou metodou detekce řečových segmentů založenou na HMM, jež pracují s již s předpřipravenými segmenty. Tyto segmenty byly na rozdíl od tabulky 7.1 (ruční segmentace) získány automatickou segmentací (modul detekce změn v audiosignálu), založenou na metodě s adaptivním oknem, viz kapitola 2.1.2. Struktura tabulky 7.3 a testovací/trénovací data použitá v experimentech si navzájem odpovídají s tabulkou 7.2. V tabulce 7.3 jsou tedy opět vypsány výsledky pro jednostavové, třístavové a pětistavové ergodické HMM. Pro jednodušší srovnání je pod tabulkou umístěn poslední řádek tabulky 7.2 s výsledky klasifikátoru pracujícího se spojitým signálem.

HMM klasifikátor pracující s předrozdělenými částmi						
	1 stavový HMM		3 stavový HMM		5 stavový HMM	
	řeč	neřeč	řeč	neřeč	řeč	neřeč
BE	99,6	32,64	97,66	25,86	99,36	26,8
CZ	99,05	83,34	95,79	88,12	99,42	78,83
GA	99,31	45,41	97,95	49,81	99,53	49,07
PT	99,97	84,71	98,46	83,81	99,33	83,81
SI	98,73	78,19	99,68	86,58	99,61	89,31
SK	96,62	45,98	93,6	56,93	98,09	56,68
All	98,86	58,79	97,16	63,42	99,17	63,43
HMM klasifikátor na spojitém základu - srovnání						
All	96,07	89,39	96,36	83,33	98,45	77,89

Tab. 7.3: *Výsledky detekce řeč/neřeč HMM klasifikátorem pracujícím se segmenty předrozdělenými detektorem změn mluvčích a audiopozadí*

Z testů vyšel nejlépe pětistavový HMM pracující s předrozdělenými segmenty, jež se také stal součástí systému pro přepis zpravodajství. Pokud budeme bez bližší specifikace hovořit o detektoru řeč/neřeč, bude se jednat právě o tuto konfiguraci ve spojení s 12 MFCC příznaky a CMS. Na obrázku 7.1 je provedeno vzájemné srovnání míry recall identifikovaných řečových segmentů pro jednostavové, třístavové a pětistavové ergodické

HMM. Na tomto obrázku je zřetelně vidět, že nejvyššího skóre bylo dosaženo právě u pětistavových HMM.



Obr. 7.1: *Porovnání výsledků detekce řeči u metody pracující se spojitým audiosignálem a předrozdělenými segmenty pro všechny národní testovací sady*

U identifikace řeč/neřeč dáváme přednost vysoké úspěšnosti při detekci řečových segmentů, i když to má za následek určité snížení úspěšnosti při identifikaci neřečových segmentů. Navíc, se špatně identifikovanými neřečovými segmenty (označeny jako řečové a připuštěny k rozpoznávání spojité řeči) si dále může poradit LVCSR, který má 7 modelů reprezentujících různé kategorie ruchů a velkou část těchto neřečových událostí dokáže odfiltrovat sám. V neposlední řadě je neřečového signálu v televizním zpravodajství výrazně méně (na celé BN databázi je poměr řeč/neřeč přibližně 7:1), tudíž chybně identifikovaný neřečový segment zatíží míru recall vyšší chybou, ale v konečném důsledku není tato chyba natolik dramatická.

7.2. Identifikace a verifikace mluvčích

Veškeré experimenty v kapitolách 7.2 a 7.3 byly realizovány na českých testovacích promluvách čítajících 181 minut (rozdělených z hlediska mluvčích do 991 homogenních promluv) televizního zpravodajství (záznamy z TV kanálů Prima, Nova a ČT1). Tyto nahrávky jsou v úlohách rozpoznávání mluvčího používány jako vývojová testovací data našeho systému pro automatický přepis audiozáznamů (viz kapitola 6.5). Pro natrénování 117 modelů mluvčích (shodná trénovací data pro identifikaci i verifikaci) bylo použito 800 minut BN záznamu. Ze záznamu byli vybráni pouze ti mluvčí, jejichž celková délka promluv překročila 100 sekund. Pro trénování UBM byla použita data zbývajících mluvčích (zbytek z 800 minut BN nahrávek a dalších 180 minut nahrávek pocházejících z jiných zdrojů než BN).

Identifikace a verifikace mluvčích				
	R_{ER}	R_{OIFA}	$R_{OIFA} + R_{FR}$	R_{OI}
Identifikace mluvčích	1,86%	-	-	10,37%
Verifikace mluvčích	-	7,44%	9,03%	

Tab. 7.4: *Výsledky IM a VM pro české, ručně segmentované BN nahrávky*

V tabulce 7.4 jsou uvedeny výsledky, kterých se nám podařilo v oblasti IM a VM na zmíněných 991 testovacích promluvách dosáhnout. Připomeňme, že R_{OIFA} označuje poměrný počet chyb identifikace neoprávněným přijetím, R_{FR} označuje poměrný počet nesprávných zamítnutí mluvčích a R_{OI} je poměrný počet chyb identifikace nad otevřenou množinou. Čtenáři by se mohlo zdát, že míra R_{OIFA} u VM je pro ručně segmentované nahrávky příliš vysoká. Zvýšenou hodnotu míry R_{OIFA} si vysvětlujeme povahou verifikovaných nahrávek, kdy pouze v 45 % případů se jedná o čistý a kvalitní audiozáznam. V ostatních případech jsou promluvy podbarveny hudbou či jinými ruchy, případně se jedná o nahrávky s nižší kvalitou.

7.3. Identifikace pohlaví

V následujících několika odstavcích budou diskutovány výsledky vývojových experimentů s identifikací pohlaví řečníka, jejichž primárním cílem bylo ověření praktické použitelnosti navržených přístupů a nalezení nejlepšího postupu, který se stane součástí systému pro přepis zpravodajských pořadů. Hlavní část experimentů byla provedena v rámci mezinárodní evaluační kampaně COST278 BN a tyto výsledky jsou uvedeny v kapitole 6.5.2.

Pro vyhodnocování výsledků získaných při identifikaci pohlaví byly využity postupy a software uvedené v kapitole 6.4. Řádek v tabulce 7.5 s označením *počet chybně identifikovaných promluv* popisuje chování metod pro IP s ohledem na počet provedených identifikací. V tomto případě je procento chybně identifikovaných promluv vypočteno podle vztahu

$$\%\ chybně\ identifikovaných\ promluv = \frac{n_{err}}{n_{celk.}} \times 100\ \%, \tag{7.1}$$

kde n_{err} je počet chybně identifikovaných promluv a $n_{celk.}$ je počet všech identifikačních pokusů. Při 991 testovaných větách je jednou chybně identifikovanou promluvou zvětšeno procento chybně identifikovaných vět přibližně o 0,1 %. Hodnoty uvedené v tabulkách 7.5 a 7.6 pod označením *délka chybných segmentů* odpovídají míře $100\ \% - ACC$, poslední řádek pak ukazuje relativní zlepšení této míry vzhledem

k referenční hodnotě (záporná hodnota značí zhoršení identifikačního skóre). Větší procento chyb je u vyhodnocování po větách způsobeno chybami při identifikaci krátkých promluv, u kterých je vlivem nedostatečné délky sekvence příznakových vektorů větší pravděpodobnost nesprávné identifikace pohlaví, ale jejich délka menší měrou ovlivňuje míru *ACC*.

Identifikace pohlaví				
	nejlepší model	**všichni mluvčí**	***n* nejlepších**	***n* % nejlepších**
počet chybně identifikovaných promluv	2,93%	2,83%	6,16%	2,52%
délka chybných segmentů [*100 % - ACC*]	2,62%	1,46%	5,82%	2,12%
relativní snížení chybovosti	-	44,27%	-122,14%	19,08%

Tab. 7.5: *Výsledky experimentů hledajících nejlepší metodu použitelnou pro identifikaci pohlaví*

Tabulka 7.5 shrnuje výsledky IP dosažené metodami vyhodnocujícími výstupní pravděpodobnosti získané při IM. Podstatou zde ověřovaných metod IP je tedy přístup založený na GMM. V případě prvního sloupce s názvem *nejlepší model* je pohlaví autora promluvy určeno z pohlaví vítězného kandidáta. Výsledek tohoto postupu je stanoven jako referenční míra, ke které jsou vztaženy výsledky ukazující relativní snížení chybovosti ostatními metodami. Druhý přístup, označený v tabulce jmenovkou *všichni mluvčí*, určuje pohlaví promlouvající osoby na základě průměrné věrohodnosti testované promluvy vůči kohortě všech mužských kandidátů v databázi a poté vůči kohortě všech ženských mluvčích uložených v databázi. Pohlaví skupiny s vyšší průměrnou věrohodností je pak přiřazeno testované promluvě. Podobný přístup je zvolen i u posledních dvou metod, pouze pro vytvoření kohort mluvčích reprezentujících dané pohlaví používáme omezující pravidla a ne tvrdé přiřazení všech mluvčích shodného pohlaví z databáze jako v předchozím případě. Ve sloupci označeném jako *n nejlepších* jsou kohorty utvořeny z pevně daného počtu nejlepších kandidátů (v tomto případě z pěti nejlepších) každého pohlaví. Posledním způsobem, uvedeným v této tabulce, je vytvoření kohorty reprezentující pohlaví na základě *n % nejlepších* kandidátů. Podle mluvčího s nevyšší věrohodností (vítěz identifikace) a stanoveného prahu (experimentálně stanoven na 105 %) jsou vybráni všichni kandidáti, kteří splňují podmínku dostatečné blízkosti k tomuto nejpravděpodobnějšímu mluvčímu. Nejlepších výsledků bylo dosahováno s metodou všichni mluvčí, kde došlo k relativnímu snížení chybovosti vůči referenční identifikaci nejlepším modelem o 44 %.

Hlavním rozdílem proti evaluační kampani COST278 BN, kde byly segmenty vytvářeny automaticky modulem detekce změn v audiosignálu, je použití ručně

rozdělených segmentů. V případě manuálního rozdělení audiosignálu nám totiž nikde neproniká mužská promluva do ženské a naopak, vlivem nesprávně stanoveného bodu změny. To je i hlavní důvod zvyšující rozpoznávací skóre v tabulce 7.5 v porovnání s výsledky získanými v evaluační kampani COST278 BN.

Identifikace pohlaví založená na rozpoznávači spojité řeči s GD modely

IP mluvčích metodou GMM je s oblibou používána jak pro svou rychlost, tak i z důvodu textové a jazykové nezávislosti (viz evaluační kampaň COST278 BN, kapitola 6.5.2). I přes relativně dobré výsledky GMM IP jsme se pokusili o zlepšení tohoto klasického postupu. Použití LVCSR pro IP nás sice omezilo pouze na český jazyk, avšak z hlediska výkonnosti může být tento na jazyku závislý přístup výhodou. Zkombinování obou do značné míry odlišných přístupů pak mohlo potlačit slabiny jedné či druhé metody a v konečném důsledku přinést požadované zlepšení rozpoznávacího skóre. Tento předpoklad nakonec potvrdily i naše experimenty.

Identifikace pohlaví - fůze GMM a LVCSR				
	GMM	LVCSR	ideální fůze	dosažená fůze
počet chybně identifikovaných promluv	2,83%	2,12%	0,71%	1,51%
délka chybných segmentů [*100 % - ACC*]	1,46%	1,07%	0,71%	0,82%
relativní snížení chybovosti	-	26,71%	51,37%	43,84%

Tab. 7.6: *Vyhodnocení výsledků experimentů s LVCSR v úloze IP s českými BN daty. V tabulce jsou dále uvedeny výsledky fůze s klasickou GMM IP.*

V tabulce 7.6 jsou uvedeny výsledky GMM IP a LVCSR IP. Předposlední sloupec udává nejlepší možný výsledek dosažitelný pomocí fůze obou metod a v posledním sloupci je uvedeno reálné rozpoznávací skóre, které bylo dosaženo na naší testovací databázi čítající 991 promluv z BN oblasti. Nejlepší dosažitelný výsledek nám říká, kolik chyb IP nedokážeme fůzí obou metod ovlivnit, tj. kolikrát selžou obě metody IP při samostatné identifikaci. Pomocí koeficientů β_1 a β_2 ze vztahu (4.34) byla oběma metodám nastavena shodná váha, tj. $\beta_1 = \beta_2 = 0{,}5$. Série experimentů spočívajících ve změně poměru β_1 / β_2 neprokázala téměř žádný pozitivní vliv na výsledné rozpoznávací skóre.

Doplněním GMM IP o metodu založenou na LVCSR bylo dosaženo zvýšení míry *ACC* systému na hodnotu 99,18 %, což je relativní snížení chybovosti o 43,84 % v porovnání se standardní GMM IP. Nejvyšší teoreticky dosažitelnou hodnotou *ACC*, jež mohla být fůzí obou metod získána, bylo 99,29 %. Ve zbývajících 0,71 % případů totiž selhala jak identifikace pohlaví pomocí GMM, tak i pomocí LVCSR. Chybovost IP

použitím samotné GMM metody byla 1,46 %, chybovost identifikace pomocí LVCSR byla 1,07 % nesprávně rozpoznaných řečových segmentů.

I přes nesporný přínos LVCSR IP na zvýšení rozpoznávacího skóre při IP nebyla tato metoda doposud zařazena do řetězce operací sloužících k přepisu zpravodajských pořadů. Hlavním důvodem je vysoká výpočetní náročnost takovéto IP. Před reálným nasazením je třeba provést sérii dalších experimentů se zmenšeným slovníkem. Tímto způsobem by bylo možné snížit výpočetní náročnost na přijatelnou úroveň, čímž by se zvýšila atraktivnost tohoto přístupu z hlediska praktického nasazení.

7.4. Optimalizace nastavení parametrů vzhledem k celému rozpoznávacímu řetězci

V této kapitole bude objasněn a experimentálně ověřen přímý vliv metod používaných pro identifikaci audiosegmentů v kontextu s rozpoznávačem spojité řeči, respektive s přepisem zpravodajských pořadů do textové podoby. Potřeba identifikace audiosegmentů byla nejprve vyvolána automatizací segmentačního procesu, kdy bylo třeba před vlastním rozpoznáváním spojité řeči vyřadit nepotřebné, mnohdy i škodlivé neřečové audiosegmenty. V době, kdy byly implementovány metody pro adaptaci řečových modelů pro LVCSR, vyvstala potřeba identifikace audiosegmentů do dalších kategorií, jimiž byly pohlaví a skutečná totožnost promlouvajících osob. Kvalitním rozčleněním audiosegmentů do uvedených kategorií tak metody pro identifikaci audiosegmentů umožní nasazení pokročilých metod používaných při rozpoznávání spojité řeči a nepřímo tak zlepší kvalitu výstupního textového přepisu.

Vliv detekce řeč/neřeč na kompletní systém pro přepis zpravodajských pořadů

Výsledky uvedené v této a následující sekci (vliv procentuální úspěšnosti identifikace řečových segmentů na LVCSR) byly získány porovnáním automaticky přepsaného záznamu s ručně rozsegmentovanými a přepsanými nahrávkami metodou DTW (HTK nástroj *HResults*). Tímto způsobem bylo nalezeno optimální přiřazení automatického přepisu zpravodajství k ručnímu (referenčnímu) textovému výstupu. Ve všech následujících experimentech zabývajících se zcela automatickým přepisem zpravodajství, tj. od rozsegmentování až po rozpoznávání řeči v modulu LVCSR, byla vždy vyhodnocována jedna hodina televizního zpravodajství složená ze tří hlavních zpravodajských relací na kanálech ČT1, Nova a Prima.

V kapitole 7.1 byly uvedeny výsledky experimentů s metodami identifikace audiosegmentů na kategorie řeč/neřeč. Na základě těchto experimentů pak byla zvolena optimální konfigurace, pětistavový ergodický HMM pracující s předrozdělenými segmenty,

12 MFCC příznaky a CMS. Nyní je třeba zjistit, jaký vliv má identifikace řeč/neřeč na textový výstup LVCSR systému.

Tabulka 7.7 ukazuje přínos zařazení detektoru řeč/neřeč do rozpoznávacího řetězce končícího rozpoznáváním spojité řeči v modulu LVCSR. Hodnota *WER* (Word Error Rate), vyjadřující míru chybně rozpoznaných slov (vynesena v posledním sloupci tabulky), je s využitím nástroje HResults z HTK toolkitu vypočtena podle vztahu

$$WER = \frac{H - I}{N} \times 100\ \%, \tag{7.2}$$

kde H (hity) jsou slova odpovídající slovům v referenčním přepisu, I (inserce) jsou slova vložená do výstupního textového přepisu a N (počet slov) je celkový počet slovních položek v referenčním přepisu. Kromě těchto hodnot jsou v tabulce 7.7 uvedeny hodnoty D (delece) – označuje vynechaná slova a S (substituce) – značí slova nahrazená jinými.

Vliv identifikace řeč/neřeč na celý rozpoznávací řetězec						
	H	D	S	I	N	*WER*
kompletní signál	7005	333	1114	249	8452	20,07%
signál bez neřečových segmentů	7001	347	1104	224	8452	19,82%

Tab. 7.7: *Porovnání textových výstupů LVCSR získaných bez použití identifikace řeč/neřeč a s použitím detektoru založeného na HMM pracujícího s předrozdělenými segmenty*

První řádek s výsledky je získán rozpoznáváním celého signálu, který je pouze rozsegmentován detektorem změn mluvčích a audiopozadí. Výsledky v druhém řádku již byly získány se zařazeným detektorem řečových signálů, který odhalené neřečové segmenty nepřipustil do LVCSR. Jak vyplývá z výsledků tabulky 7.7, největší redukce chyb bylo dosaženo u insercí, jejichž počet byl odstraněním neřečových segmentů nejvíce snížen a to o 10 %. Tak byl potvrzen i náš předpoklad, že některé ruchy jsou i přes přítomnost ruchových modelů v LVCSR chybně rozpoznány jako slova. Na druhou stranu přibyly 4 % delecí, což jsou chyby jdoucí na vrub nesprávně odstraněným řečovým segmentům. K uvedeným výsledkům je třeba dodat, že nereprezentují nejlepší dosažené rozpoznávací skóre, pouze ukazují zlepšení, ke kterému dojde po zařazení identifikace řeč/neřeč do rozpoznávacího řetězce. Stejných experimentů bylo provedeno v průběhu vývoje systému přepisujícího zpravodajské pořady více a ve všech případech se absolutní snížení *WER* pohybovalo kolem hodnoty 0,25 % jako v tomto případě.

Vliv procentuální úspěšnosti identifikace řečových segmentů na LVCSR

Výsledky uvedené v předchozím odstavci byly dosaženy s konfigurací dosahující úspěšnosti 99,42 % odhalených řečových a 78,83 % odhalených neřečových segmentů. Otázkou zůstává, zda je tento poměr identifikovaných řečových/neřečových segmentů výhodný i pro LVCSR. Podívejme se na tabulku 7.8, kde jsou ve třech hlavních sloupcích vyneseny výsledky získané HMM systémem pracujícím s předrozdělenými segmenty, jehož výsledky byly uvedeny v kapitole 7.1. První řádek s výsledky pak odpovídá druhému řádku tabulky 7.3.

Vliv detekce řečových segmentů na rozpoznávací skóre LVCSR						
	1 stavový HMM		3 stavový HMM		5 stavový HMM	
	řeč	neřeč	řeč	neřeč	řeč	neřeč
počet identif. řečových segmentů	99,05%	83,34%	95,79%	88,12%	99,42%	78,83%
Accuracy	98,29%		95,41%		98,41%	
WER	19,96%		20,41%		19,82%	

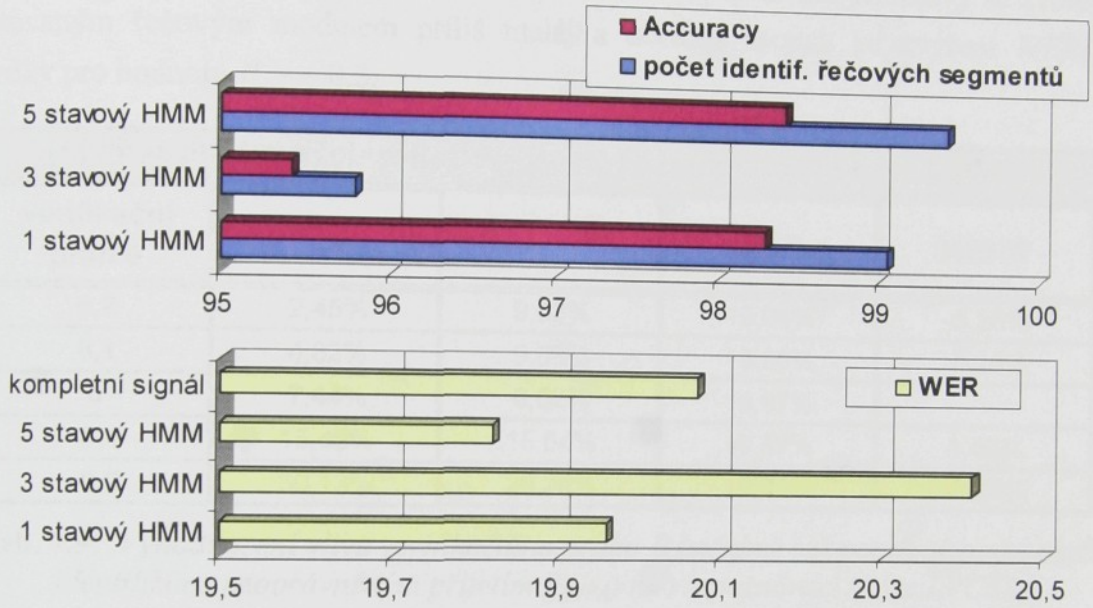
Tab. 7.8: *Výsledky uvedené v této tabulce znázorňují, jakým způsobem je ovlivněna úspěšnost přepisu zpravodajských pořadů správně identifikovanými řečovými segmenty*

Podívejme se na výsledky detekce řeč/neřeč z hlediska míry *ACC*. Mezi jednostavovým a pětistavovým HMM je v míře *ACC* minimální rozdíl 0,12 %. Oba výsledky se však liší poměrem identifikovaných řečových ku neřečovým segmentům. Rozdíl v úspěšnosti odhalování řečových segmentů, který lze vyčíst z prvního řádku tabulky 7.8 činí 0,37 % ve prospěch pětistavového HMM. Tento rozdíl má pak podstatný vliv na *WER* LVCSR, kde při odstranění neřečových segmentů pětistavovým HMM je snížena míra *WER* oproti jednostavovému HMM o 0,14 %. Tento výsledek pak potvrzuje naši hypotézu, že se vzrůstajícím procentem odhalených řečových segmentů klesá i *WER* LVCSR.

Na druhou stranu, aby nasazení detektoru řeč/neřeč mělo své opodstatnění, je třeba volit kompromis mezi maximalizací procenta identifikovaných řečových segmentů a přijatelnou úspěšností při odhalování neřečových segmentů. V případě třístavového HMM (prostřední sloupec tabulky 7.8) došlo k opačnému efektu, jímž je zvýšení *WER* LVCSR. Příčinou tohoto jevu je příliš nízká úspěšnost při odhalování řečových segmentů. Pokles je natolik výrazný, že výsledné rozpoznávací skóre je dokonce pod úrovní LVCSR bez identifikace řeč/neřeč (79,59 %). Diskutované varianty jsou graficky zobrazeny na obrázku 7.2.

Nastavování verifikačního prahu vzhledem k LVCSR modulu

V této a v následující sekci (Kombinace akustických modelů pro LVCSR) byly experimenty realizovány na českých testovacích promluvách čítajících 181 minut televizního zpravodajství (záznamy z TV kanálů Prima, Nova a ČT1). Tyto promluvy byly manuálně rozsegmentovány tak, že vzniklo 991 homogenních promluv (z hlediska mluvčích). Možnosti a důvody vedoucí k experimentování s nastavením verifikačního prahu již byly zmíněny v kapitole 4.6.1.



Obr. 7.2: *V horním grafu je spolu s hodnotou accuracy pro identifikaci řeč/neřeč vykreslen i počet úspěšně identifikovaných řečových segmentů. Na spodním grafu jsou pak vyneseny míry WER pro odpovídající počty stavů HMM. Pro porovnání byl přidán i výsledek LVCSR bez použití identifikace řeč/neřeč (nejvýše umístěný horizontální sloupec).*

V tabulce 7.9 jsou uvedeny výsledky experimentů, jež ověřují naši hypotézu o pozitivním vlivu určitého procenta neoprávněně akceptovaných mluvčích (modul pro VM) na *WER* LVCSR. Nově zavedenou veličinou je relativní zlepšení míry chybně rozpoznaných slov *WERR* (Word Error Rate Reduction). Podstata experimentu spočívala v posunutí verifikačního prahu směrem k vyšší, respektive nižší bezpečnosti[1] VM tak, abychom dosáhli snížení resp. zvýšení míry R_{OIFA}. Sekundárním jevem těchto posunů bylo i zvýšení celkové chyby verifikace ($R_{OIFA} + R_{FR}$), neboť posunem verifikačního prahu došlo k odchýlení od *EER* bodu, čímž zákonitě došlo i k nárůstu celkové chyby verifikace.

[1] Bezpečností je v tomto kontextu míněna velikost míry R_{OIFA}, tj. množství neoprávněně akceptovaných narušitelů.

Výsledky uvedené v tabulce 7.9 ukazují, že v případě mírného snížení verifikačního prahu θ může docházet k redukci *WER* LVCSR. Pravděpodobným důvodem tohoto jevu je použití adaptovaných řečových modelů, jež mají v případě vhodného nasazení lepší úspěšnost než GD řečové modely[1]. Vždyť GMM používané pro IM a VM jsou ve své podstatě zjednodušenou obdobou HMM používaných pro rozpoznávání spojité řeči. Spolu se shodnými MFCC příznaky pak volba nesprávného, přesto však podobného adaptovaného řečového modelu metodou GMM může znamenat vhodnou volbu i pro HMM. Pokud je však hodnota θ stanovena příliš nízko, je podobnost rozpoznávané promluvy se zvoleným adaptovaným řečovým modelem příliš malá a dochází naopak ke zvýšení *WER*, viz výsledky pro hodnotu $\theta = -0{,}2$.

Vliv neoprávněných přijetí R_{OIFA} na rozpoznávací skóre LVCSR				
verifikační práh θ	R_{OIFA}	$R_{OIFA} + R_{FR}$	***WER***	***WERR***
0,2	2,45%	9,63%	19,94%	-0,36%
0,1	4,82%	9,08%	19,90%	-0,18%
0	7,44%	9,02%	19,87%	-
-0,1	15,46%	15,64%	19,57%	1,49%
-0,2	30,13%	30,26%	19,91%	-0,24%

Tab. 7.9: *Vyhodnocení vlivu verifikačního prahu θ (měníme tak poměrný počet chyb identifikace neoprávněným přijetím R_{OIFA}) na rozpoznávací skóre LVCSR*

Na obrázku 7.3 jsou pak míry *WERR* a R_{OIFA} znázorněny i graficky. Dodejme, že v případě nahrazení manuální segmentace automatickou se rozpoznávací skóre LVCSR systému zhoršuje v průměru o 1–1,5 %.

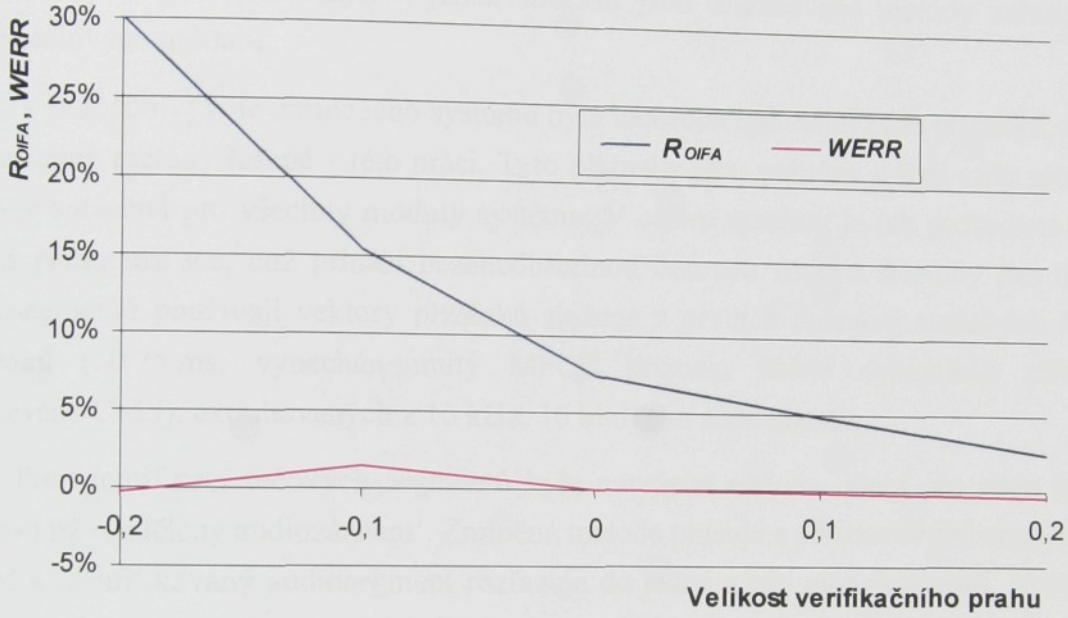
Kombinace akustických modelů pro LVCSR

Jak se později ukázalo, lepším řešením pro rozpoznávání řeči než použití nejpodobnějšího adaptovaného řečového modelu je sestavení modelu nového, získaného kombinací modelů kohorty mluvčích [ČER05]. V této sekci se zaměříme pouze na 393 řečových segmentů (71 minut), u nichž byli navržení mluvčí při standardně nastaveném verifikačním prahu θ zamítnuti. Kompletní postup přepisu audiozáznamů spolu se zde uváděným schématem adaptace řečových modelů je blokově rozkreslen na obrázku 2.2.

Výsledky získané testováním níže popsaného přístupu k adaptaci řečových modelů na konkrétního mluvčího jsou shrnuty v tabulce 7.10. Tato tabulka ukazuje míry *WER* a

[1] GD modely jsou pro rozpoznávání používány v tom případě, pokud je navržená identita mluvčího (blok IM) v průběhu verifikačního procesu zamítnuta.

WEER v závislosti na velikosti kohorty mluvčích (N), z níž jsou nové adaptované řečové modely vytvářeny. Protože způsoby adaptace řečových modelů nejsou tématem této práce, je postup adaptace popsán především z hlediska IM a IP. Pro rozpoznávání segmentů, jejichž navrhovaní kandidáti byli v průběhu VM akceptováni jsou standardně použity adaptované řečové modely vytvořené z jejich trénovacích dat. V okamžiku, kdy je navržený kandidát zamítnut, využíváme informací z identifikačního procesu pro vytvoření setříděného seznamu nejúspěšnějších N kandidátů z IM, kteří budou tvořit adaptační kohortu.



Obr. 7.3: *Graf závislosti poměrného počtu chyb identifikace neoprávněným přijetím R_{OIFA} u VM a relativního zlepšení rozpoznávacího skóre WERR u LVCSR*

Při vytváření kohorty je dobré, pokud jsou všichni mluvčí stejného pohlaví (jak z hlediska rozpoznávacího skóre, tak i z hlediska praktické realizovatelnosti, neboť mixtury obou pohlaví mohou mít rozdílné velikosti). Váha, s jakou se modely mluvčích na výsledném adaptovaném modelu podílejí, je určena pořadím z identifikace. Výsledný adaptovaný model je pak lineární kombinací vektorů středních hodnot HMM mluvčích z kohorty. Variance a váhy mixtur jsou jednoduše zkopírovány z odpovídajícího GD modelu.

Kombinace akustických modelů pro LVCSR								
N	***3***	***5***	***7***	***10***	***15***	***20***	***30***	***40***
WER [%]	21,7	21,2	21,2	21,2	20,9	20,9	21,3	21,4
WERR [%]	9,1	11,2	11,2	11,2	12,4	12,4	10,8	10,3

Tab. 7.10: *Výsledky testování adaptace řečových modelů na konkrétního mluvčího s využitím kohorty mluvčích. Za referenční výsledek byl zvolen výstup systému s SI modely, jež na totožných testovacích datech dosáhl WER = 23,9 %.*

8. Závěr

Primárním cílem této disertační práce bylo navržení, implementace a reálné ověření metod pro identifikaci audiosegmentů v úloze automatického přepisu zpravodajství. Prvotním impulsem ke stanovení tohoto cíle byla participace SpeechLabu pod vedením Prof. Ing. Jana Nouzy, CSc. na evropském projektu COST278 BN, který je zaměřen na textový přepis zpravodajských pořadů. V Liberci tak byl vytvořen funkční systém pro plně automatický přepis zpravodajství, jehož součástí jsou diskutované metody zařazené do jednoúčelových modulů.

V průběhu vývoje zmíněného systému byla nalezena optimální sada příznaků vhodná pro všechny metody řešené v této práci. Tyto příznaky jsou vybrány z širší sady příznaků, která je společná pro všechny moduly systému. V celém systému je tak provedena pouze jediná parametrizace, což přináší nezanedbatelnou časovou úsporu. Metody pro detekci audiosegmentů používají vektory příznaků složené z prvních dvanácti statických MFCC příznaků (10/25 ms, vynechán nultý MFCC příznak, žádné dynamické příznaky, aplikováno CMS), extrahovaných z 16 kHz, 16 bitového audiosignálu.

Pro identifikaci řečových segmentů byla navržena metoda, která pro svou činnost využívá již rozdělený audiozáznam[1]. Zmíněná metoda pracuje s pětistavovými ergodickými HMM a identifikovaný audiosegment rozřazuje do jedné z pěti audiokategorií. V případě, že systém je trénován i testován na manuálně segmentovaných českých datech, úspěšnost rozpoznávání stoupá až na 99,77 % správně detekovaných řečových segmentů při 14,24 % nesprávně prohlášených neřečových segmentech za řečové. Na celé databázi COST278-BN byla při použití automatické segmentace (v rámci mezinárodní evaluační kampaně) dosažena úspěšnost identifikace řečových částí 98,4 %, neřečových částí bylo nalezeno 70,7 %, což v souhrnu odpovídá 96,3 % *ACC*. Tohoto výsledku bylo dosaženo po natrénování systému externími (českými) daty. Pokud bychom brali v úvahu pouze výsledky získané na české podskupině databáze COST278-BN, procento nalezených řečových/neřečových segmentů by bylo podstatně vyšší (99,5 % / 76,0 %) a odpovídalo by míře 98,2 % *ACC*.

V době provádění experimentů pracovala identifikace mluvčích se 117 nejfrekventovanějšími mluvčími z české trénovací databáze (dnes to je přes 300 mluvčích). Identifikace je postavena na metodě GMM se zrychleným výpočtem. Zrychlení je provedeno

[1] Tyto části jsou nazývány segmenty, což jsou pokud možno homogenní části (z hlediska mluvčích či z hlediska audio pozadí) spojitého audiosignálu.

rozdělením identifikačního procesu do dvou průchodů. V prvním z nich je testovaná promluva porovnávána se všemi modely v databázi. Počet gaussovských složek modelů je však omezen na 8, což způsobí, že identifikace je provedena se zmenšenou přesností, avšak velmi rychle. Druhé kolo identifikace absolvuje pouze vybraný počet kandidátů (na základě výsledků v prvním kole) s finální velikostí GMM modelů (64 složek). Úspěšnost identifikace na ručně segmentovaných záznamech televizního zpravodajství (181 minut českých testovacích nahrávek rozdělených do 991 promluv) činí 92,45 % správně přiřazených segmentů k mluvčím v databázi. Daleko přesnější mírou je délka správně identifikovaných segmentů (*ACC*), jež činí 98,76 %. Tento rozdíl je způsoben různými délkami segmentů, přičemž největší procento chyb je způsobeno krátkými promluvami. Proto rychleji klesá počet správně identifikovaných segmentů než míra *ACC*. Modul pro identifikaci mluvčího také poskytuje informace o dalších kandidátech s nejvyššími věrohodnostmi. Tato informace slouží v pozdější fázi zpracování zpravodajství k adaptaci řečových modelů na neznámé mluvčí[1].

Identifikace pohlaví používá jako základ pro svou práci výsledky získané z modulu pro identifikaci mluvčích. Pro každého mluvčího z databáze je spočtena věrohodnost (tato pravděpodobnost je ve skutečnosti převzata z výsledků IM), s jakou by jeho model generoval daný segment. Konečné rozhodnutí o pohlaví mluvčího je provedeno na základě průměrné hodnoty věrohodnosti všech mluvčích obou pohlaví. Míra *ACC* u takto nakonfigurovaného systému je u ručně segmentovaných českých dat (již zmíněných 991 promluv) 98,54 %. Doplněním GMM IP o metodu založenou na LVCSR bylo dosaženo zvýšení míry *ACC* na hodnotu 99,18 %. Při experimentech s GMM IP na automaticky segmentované databázi COST278-BN byla v rámci mezinárodní evaluační kampaně dosažena míra *ACC* 94,9 % (totožná míra *ACC* jak pro česká tak i pro mezinárodní data).

Základem verifikace mluvčích jsou opět GMM. Pro reprezentaci hypotézy, že promluva nepochází od mluvčího určeného v průběhu IM byl zvolen UBM (1024 složek). Opačnou hypotézu, tzn. že promluva pochází od proklamovaného mluvčího, reprezentuje GMM adaptovaný z UBM (model byl adaptován daty proklamovaného mluvčího). Ve fázi rozpoznávání zjišťujeme pravděpodobnost obou hypotéz a na základě přednastaveného prahu sloužícího k nastavení míry verifikační jistoty je potvrzena nebo zamítnuta identita mluvčího. Na tradičních 991 testovacích promluvách byla při verifikačním prahu nastaveném na $\theta = 0$ dosažena míra celkové chyby verifikace ($R_{OIFA} + R_{FR}$) 9,03 %, přičemž míra R_{OIFA} činila 7,44 %. VM dále určuje, který řečový model bude použit při

[1] V tomto kontextu jsou neznámí mluvčí všichni ti, kteří v databázi nemají uložen model, byli špatně identifikováni a při verifikaci zamítnuti nebo je *SNR* příliš nízké a mluvčí byli při verifikaci zamítnuti.

rozpoznávání řeči. Pokud byl identifikovaný mluvčí také verifikován, jsou při rozpoznávání spojité řeči apriorně použity na něj adaptované řečové HMM. V opačném případě jsou řečové modely adaptovány na skupinu nejbližších osob. Výběr je prováděn pouze z osob, které mají stejné pohlaví jako verifikovaný kandidát.

Kvalitním rozčleněním audiosegmentů do výše diskutovaných kategorií tak metody pro identifikaci audiosegmentů umožňují nasazení pokročilých metod používaných při rozpoznávání spojité řeči a nepřímo tak zlepšují kvalitu výstupního textového přepisu. V závěru této práce byla provedena řada experimentů ověřujících skutečný vliv vyvinutých metod na konečný textový přepis zpravodajství. Použitím adaptovaných řečových modelů v modulu LVCSR (jejich nasazení umožnily metody pro IP, IM a VM) bylo při plně automatickém přepisu jedné hodiny televizního zpravodajství[1] dosaženo snížení *WER* až o 3,75 %. Zařazením detektoru řeč/neřeč do rozpoznávacího řetězce bylo za týchž podmínek dosaženo dalšího absolutního snížení *WER* o 0,25 %.

Shrnutí přínosů k rozvoji vědního oboru

- V práci je uceleným způsobem shrnuta problematika automatické transkripce zpravodajských pořadů. Zvláštní důraz je kladen na IM, IP, VM a detekci řečových segmentů, což jsou metody jejichž problematika je v této práci diskutována.
- Autor je spolutvůrcem panevropské databáze televizního zpravodajství COST278-BN, která slouží k vývoji a porovnání metod používaných při automatické transkripci zpravodajských pořadů. Autor se dále aktivně podílel na metodice použití této databáze a vyhodnocovacích nástrojích využívaných ve společných experimentech.
- V rámci mezinárodní evaluační kampaně bylo provedeno vyhodnocení úspěšnosti IP a detekce řečových segmentů na databázi COST278-BN. Ve třech ze čtyř vyhodnocovaných kategorií dosáhl autor nejlepších výsledků. Tyto výsledky pak byly společnou publikací prezentovány na prestižní světové konferenci.
- Vytvoření a praktické odzkoušení metody pro zrychlenou IM, jejíž principem je rozdělení identifikačního procesu do dvou rozpoznávacích průchodů.
- Nalezení velice rychlé a v praxi snadno použitelné metody pro IP založené na vyhodnocování výsledků získaných při IM.
- Ověření využitelnosti LVCSR s GD modely v úloze IP.

[1] Tato hodina je složena ze tří hlavních zpravodajských relací na kanálech ČT1, Nova a Prima.

- Všechny v práci diskutované metody byly postupně testovány v systému pro automatický přepis televizního zpravodajství[1] vyvíjeného na TUL. Tímto způsobem byla prokázána možnost jejich praktického nasazení.

Shrnutí přínosů pro praxi

- Vytvoření real-time systému pro IM a VM, který je za účelem reálného nasazení dlouhodobě testován softwarovou společností CIT.
- Fůzí výsledků dvou metod pro IP (GMM a LVCSR) bylo dosaženo relativního snížení chybovosti systému identifikujícího pohlaví o 43,84 % (proti do té doby používané GMM IP).
- Zařazením modulů pro IM, VM a IP do řetězce operací prováděných systémem pro přepis zpravodajských pořadů byl umožněn vývoj a použití metod pro adaptaci řečových modelů (modul LVCSR) na konkrétního mluvčího. Metody pro rozpoznávání mluvčího tak nepřímo přispěly ke zvýšení konečného rozpoznávacího skóre celého systému. Adaptované řečové modely spolu s detekcí řečových segmentů snižují míru *WER* až o 4 %.
- Identifikace řeč/neřeč přispěla ke zkvalitnění finálního textového výstupu nejen z hlediska výsledného rozpoznávacího skóre, ale i s ohledem na čitelnost a „prezentovatelnost" těchto výstupů.

Náměty pro další práci

Jedním z budoucích úkolů je ověření funkčnosti navržených metod a postupů (v součinnosti s ostatními moduly systému pro přepis zpravodajských pořadů) i pro jiné druhy audiosignálů, čímž bychom získali komplexní media-mining systém. Dosud byl náš systém, kromě přepisu zpravodajských pořadů, úspěšně nasazen i v případě některých rozhlasových pořadů a televizních či rozhlasových debat.

Další výzkum by se mohl věnovat implementaci a experimentálnímu ověření přínosů detekce znělých a neznělých fonémů v modulech pro rozpoznávání mluvčích. Postup by mohl být následující: na základě fonetického přepisu segmentu získaného z rozpoznávače spojité řeči bude fonémovým zarovnávačem provedeno časové přiřazení textu na úrovni jednotlivých fonémů. Ze zparametrizovaného signálu budou vybrány pouze ty úseky, které budou splňovat předem stanovená kriteria (například již zmíněnou znělost). GMM

[1] Výjimkou je pouze identifikace pohlaví pomocí rozpoznávače spojité řeči, jejímuž reálnému nasazení zatím brání přílišná výpočetní náročnost této metody.

identifikace a verifikace takto předzpracovaného segmentu by pak pokračovala standardním způsobem.

V neposlední řadě bychom se mohli pokusit o rozšíření využití LVCSR (doposud používané pouze pro IP s GD modely) i na ostatní úlohy rozpoznávání řečníka. Při rozpoznávání řeči za pomoci adaptovaných řečových modelů existuje pro každého mluvčího v databázi adaptovaný řečový model, který může reprezentovat jeho identitu. Na základě známých výsledků z experimentů s IP očekáváme zajímavé výsledky i u IM, případně VM. Fůze s již implementovanými postupy by pak, stejně jako u IP, mohla znatelně vylepšit dosahovaná rozpoznávací skóre.

Literatura

[BAC03] Backfried, G., Caldés, R.J.: *Spanish Broadcast News Transcription.* Proc. of Eurospeech2003, Geneva, Sept. 2003, pp.1561–1564.

[BAL99] Balchandran, R., Ramanujam, V., Mammone, R.: *Channel Estimation Normalization by Coherent Spectral Averaging For Robust Speaker Verification.* Proc. 6th European Conf. Speech Communication and Technology, Budapest 1999, pp.755–758.

[BIM04] Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Reynolds, D.A.: *A tutorial on text-independent speaker verification.* EURASIP Journal on Applied Signal Processing 2004, vol.4, pp.430–451.

[CAM97] Campbell, J.P. jr.: *Speaker Recognition: A Tutorial.* Proceedings of the IEEE, vol. 85, no. 9, September 1997, pp.1437–1462.

[DEM77] Dempster, A.P., Laird, N.M., Rubin, D.B.: *Maximum Likelihood from Incomplete Data via the EM Algorithm.* Journal of the Royal Statistical Society, 1977, pp.1–38.

[DUD01] Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification.* John Wiley & Sons, USA, 2001.

[DUN00] Dunn, R., Reynolds, D.A., Quatieri, T.F.: *Approaches to Speaker Detection and Tracking in Conversational Speech.* In Digital Signal Processing, vol.10, 2000, pp.93–112.

[FUR81] Furui, S.: *Cepstral Analysis Technique for Automatic Speaker Verification.* IEEE Trans. Acoustics, Speech and Signal Processing, vol.29, No.2, 1981, pp.254–272.

[GAU02] Gauvain, J.L., Lamel, L., Adda, G.: *The LIMSI Broadcast News Transcription System.* Speech Communication, vol.37(1–2), 2002, pp.89–108.

[GAU99] Gauvain, J.L., Lamel, L., Adda, G., Jardino, M.: *The LIMSI 1998 HUB-4E Transcription System.* Proc. of the DARPA Broadcast News Workshop, Herndon, 1999, pp.99–104.

[GIB97] Gibbon, D., Moore, R., Winski, R.: *Handbook of standards and resources for spoken language systems.* New York: Mouton de Gruyter, 1997.

[GON93] Gonzales, R.C., Woods, R.E.: *Digital Image Processing.* Addison-Wesley, Reading, Massachusetts, 1993.

[GRA82] Gray, R.M., Karnin, E.D.: *Multiple Local Optima in Vector Quantizers.* IEEE Trans. on Information Theory, 1982, IT-28, pp.256–261.

[HAR01] Harb, H., Chen, L., Auloge, J.Y.: *Speech/music/silence and gender detection algorithm.* In Proceedings of the 7th International conference on Distributed Multimedia Systems DMS01, September 2001, pp.257–262.

Literatura

[BAC03] Backfried, G., Caldés, R.J.: *Spanish Broadcast News Transcription*. Proc. of Eurospeech2003, Geneva, Sept. 2003, pp.1561–1564.

[BAL99] Balchandran, R., Ramanujam, V., Mammone, R.: *Channel Estimation Normalization by Coherent Spectral Averaging For Robust Speaker Verification*. Proc. 6th European Conf. Speech Communication and Technology, Budapest 1999, pp.755–758.

[BIM04] Bimbot, F., Bonastre, J.F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Reynolds, D.A.: *A tutorial on text-independent speaker verification*. EURASIP Journal on Applied Signal Processing 2004, vol.4, pp.430–451.

[CAM97] Campbell, J.P. jr.: *Speaker Recognition: A Tutorial.* Proceedings of the IEEE, vol. 85, no. 9, September 1997, pp.1437–1462.

[DEM77] Dempster, A.P., Laird, N.M., Rubin, D.B.: *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, 1977, pp.1–38.

[DUD01] Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. John Wiley & Sons, USA, 2001.

[DUN00] Dunn, R., Reynolds, D.A., Quatieri, T.F.: *Approaches to Speaker Detection and Tracking in Conversational Speech*. In Digital Signal Processing, vol.10, 2000, pp.93–112.

[FUR81] Furui, S.: *Cepstral Analysis Technique for Automatic Speaker Verification*. IEEE Trans. Acoustics, Speech and Signal Processing, vol.29, No.2, 1981, pp.254–272.

[GAU02] Gauvain, J.L., Lamel, L., Adda, G.: *The LIMSI Broadcast News Transcription System*. Speech Communication, vol.37(1–2), 2002, pp.89–108.

[GAU99] Gauvain, J.L., Lamel, L., Adda, G., Jardino, M.: *The LIMSI 1998 HUB-4E Transcription System.* Proc. of the DARPA Broadcast News Workshop, Herndon, 1999, pp.99–104.

[GIB97] Gibbon, D., Moore, R., Winski, R.: *Handbook of standards and resources for spoken language systems.* New York: Mouton de Gruyter, 1997.

[GON93] Gonzales, R.C., Woods, R.E.: *Digital Image Processing*. Addison-Wesley, Reading, Massachusetts, 1993.

[GRA82] Gray, R.M., Karnin, E.D.: *Multiple Local Optima in Vector Quantizers.* IEEE Trans. on Information Theory, 1982, IT-28, pp.256–261.

[HAR01] Harb, H., Chen, L., Auloge, J.Y.: *Speech/music/silence and gender detection algorithm*. In Proceedings of the 7th International conference on Distributed Multimedia Systems DMS01, September 2001, pp.257–262.

[HAR03] Harb, H., Chen, L.: *Gender Identification Using A General Audio Classifier*. Proc. of the IEEE International Conference on Multimedia & Expo ICME 2003, July 6–9, Baltimore, USA, vol.2, pp.733–736.

[HUA01] Huang, X., Acero, A., Hon, H.W.: *Spoken Language Processing*. In Prentice Hall PTR, New jersey, United States of America, 2001, ISBN 0-13-022616-5.

[HUA90] Huang, X., Ariki, Y., Jack, M.A.: *Hidden Markov models for speech recognition*. Edinburgh University Press 1990.

[ISO99] Isobe, T., Takahashi, J.: *Text-independent speaker verification using virtual speaker based cohort normalization*. Proc. of the European Conference on Speech Communication and Technology, 1990, pp.987–990.

[JAI04] Jain, A.K., et al.: *An Introduction to Biometric Recognition*. IEEE Trans. on Circuits and Systems for Video Technology, vol.14, No.1, 2004, pp.4–20.

[LEG95] Leggetter, C.J., Woodland, P.C.: *Flexible Speaker Adaptation Using Maximum Likelihood Linear Regression*. Proc. ARPA Spoken Language Technology Workshop, 1995, Morgan Kaufmann, pp.104–109.

[LIE03] Lie Lu, Hong-Jiang Zhang, Stan Li: *Content-based Audio Classification and Segmentation by Using Support Vector Machines*. ACM Multimedia Systems Journal 8 (6), 2003, pp.482–492.

[LIE02b] Lie Lu, Hong-Jiang Zhang: *Speaker Change Detection and Tracking in Real-Time News Broadcasting Analysis*. Proc. of the tenth ACM international conference on Multimedia, Juan-les-Pins, France 2002, pp.602–610.

[LIE02a] Lie Lu, Hong-Jiag Zhang, Hao Jiang: Content Analysis for Audio Classification and Segmentation", IEEE Trans. on Speech and Audio Processing, vol.10, No.7, Oct. 2002, pp.504–516.

[LIK88] Li, K.P., Porter, J.E.: *Normalizations and selection of speech segments for speaker recognition scoring*. Proc. of IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP'88), vol.1, New York, NY, USA, April 1988, pp.595–598.

[LIN05] Ling Guo, Ying-Chun Shi, Xian-Zhong Zhou, Feng Zhang: *Location and Extraction of Broadcast in News Video Based on QGMM and BIC*. The Fifth International Conference on Computer and Information Technology (CIT'05), 2005, pp.662–667.

[LIN80] Linde, Y., Buzo, A., Gray, R.M.: *An Algorithm for Vector Quantizer Design*. IEEE Trans. on Communication, 1980, COM-28(1), pp.84–95.

[MAT93] Matsui. T., Furui, S.: *Concatenated phoneme models for text variable speaker recognition*. Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Mineapolis, 1993, pp.II391–394.

[MCT03] McTait, K., Adda-Decker, M.: *The 300K LIMSI German Broadcast News Transcription System*. Proc. of Eurospeech 2003, Geneva, 2003, pp.213–216.

[NGU02] Nguyen, L., Gue, X., Schwartz, R., Makhoul, J.: *Japanese Broadcast News Transcription*. Proc. of ICSLP 2002, Denver, October 2002, pp.1749–1752.

[NOU95] Nouza, J.: *On the Speech Feature Selection Problem: Are Dynamic Features More Important Than the Static Ones?* Proc. of EUROSPEECH'95 Conference, Madrid, Spain, September 1995, pp.919–923.

[NOU01] Nouza, J.: *Počítačové zpracování řeči – cíle, problémy, metody a aplikace.* Liberec 2001.

[PAR96] Parris, E.S., Carey, M.J.: *Language independent gender identification.* In Acoustics, Speech and Signal Processing ICASSP-96, vol.2, 1996, pp.685–688.

[PON04] Pongtep, A., Sepideh, B., Hansen, J.H.L.: *Cluster-dependent modeling and confidence measure processing for in-set/out-of-set speaker identification.* In INTERSPEECH-2004, Jeju Island, Korea, pp.2385–2388.

[PRO96] Proakis, J.G., Manolakis, D.G.: *Digital Signal Processing: Principles, Algorithms and Applications.* Prentice Hall, USA, 1996.

[PSU95] Psutka, J.: *Komunikace s počítačem mluvenou řečí.* V nakladatelství Academia – Akademie věd České republiky, Česká republika, Praha, 1995.

[RAD04] Radová, V.: *Rozpoznávání řečníka.* Habilitační práce, Česká republika, Plzeň, 2005.

[RED84] Redner, R.A., Walker, H.F.: *Mixture Densities, Maximum Likelihood and the EM Algorithm.* SIAM review, 1984, vol. 26, No. 2, pp.195–239.

[REY00] Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: *Speaker Verification Using Adapted Gaussian Mixture Models.* Digital Signal Processing, 2000, vol. 10, pp.19–41.

[REY92] Reynolds, D.A.: *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification.* Ph.D. thesis, Georgia Institute of Technology, September 1992.

[REY94b] Reynolds, D.A.: Experimental Evaluation of Features for Robust Speaker Identification. IEEE Trans. Speech and Audio Processing, vol. 2, No. 4, 1994, pp.639–643.

[REY94a] Reynolds, D.A.: *Speaker Identification and Verification Using Gausian Speaker Mixture Models.* Proc. of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland, April 1994, pp.27–30.

[REY97] Reynolds, D.A.: *Comparison of background normalization methods for text-independent speaker verification.* Proc. of the European Conference on Speech Communication and Technology, September 1997, pp.936–966.

[ROS92] Rosenberg, A.E., DeLong, J., Lee, C.H.,Juang, B.H.: *The use of cohort normalized scores for speaker verification.* Proc. Intl. Conf. on Spoken Language Processing, 1992, pp.599–602.

[SAN03] Sanderson, C.: *Automatic Person Verification Using Speech and Face Information.* Ph.D. thesis, Griffith University, February 2003.

[SCH00] O'Schaughnessy, D.: *Speech Communications: human and machine.* IEEE Press, New York, 2000.

[SCH97] Scheirer, E., Slaney, M.: *Construction and evaluation of a robust multifeature music/speech discriminator*. In Proc. of ICASSP' 97, Apr. 1997, vol. II, pp.1331–1334.

[SLO97] Slomka, S., Sridharan, S.: *Automatic Gender Identification Optimised For Language Independence*. Proceeding of IEEE TENCON Speech and Image Technologies for Computing and Telecommunications, 1997, pp.145–148.

[SVO00] Svobodová, M.: *Identifikace mluvčího v češtině na základě akustického spektra hlásek*. konference: Przemiany fonetyki slowianskiej w latach 1944–2000 Toruń, 26–27.2.2000, Fonetický ústav FF UK Praha.

[TIN03] TingYao Wu, Lie Lu, Ke Che, Hong-Jiang Zhang: *UBM-based Incremental Speaker Adaptation*. Proc. of IEEE International Conference on Multimedia and Expo, vol. II, Baltimore, MD, July 6–9, 2003, pp.721–724.

[VAN03] Vandecatseye, A., Martens, J.P.: *A Fast, Accurate and Stream-Based Speaker Segmentation and Clustering Algorithm*. Eurospeech, vol.2-14, Geneva, 2003, pp.941–944.

[WIL99] Williams, G., Ellis, D.: *Speech/music discrimination based on posterior probability features*. Procs. of Eurospeech 1999, Budapest, 1999, pp.687–690.

[YOU02] Young, S. J. et al.: *The HTK Book*. Entropic Inc, 1995, revised 2002.

[ZHA92] Zhang, T., Kuo, C.C. J.: *Video content parsing based on combined audio and visual information*. Proc. SPIE, vol. IV, 1992, pp.78–89.

[ŽĎÁ05] Žďánský, J.: *Metody detekce změny mluvčího v akustickém signálu*. Disertační práce, Liberec 2005.

Seznam vlastních prací

[ČER05] Červa, P., David, P., Nouza, J.: *Acoustic Modeling Based on Speaker Recognition and Adaptation for Improved Transcription of Broadcast Programs*. In: Specom 2005, October, 2005, Patras, Greece, pp. 183–186, ISBN 5-7452-0110-X.

[NOU05b] Nouza, J., Červa, P., Žďánský, J., Kolorenč, J., David, P.: *Towards automatic transcription of parliament speech*. In: Electronic Speech Signal Processing 2005, Semtember, 2005, Prague, Czech Republic, pp. 237–244.

[NOU05a] Nouza, J., Žďánský, J., David, P., Červa, P., Kolorenč, J., Nejedlová, D.: *Fully Automated Process of Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon*. In: Interspeech 2005, September, 2005, Lisboa, Portugal, pp. 1681–1684, ISSN 1018-4074.

[ŽIB05] Žibert, J., Mihelič, F., Martens, J., P., Meinedo, H., Neto, J., Docio, L., Garcia-Mateo, C., David, P., Žďánský, J., Pleva, M., Cizmar, A., Zgank, A., Kačič, Z., Teleki, C., Vicsi, K.: *The COST278 Broadcast News Segmentation and Speaker Clustering Evaluation – Overview, Methodology, Systems, Results*. In: Interspeech 2005, September, 2005, Lisboa, Portugal, pp. 629–632.

[DAV05] David, P., Červa, P., Nouza, J.: *Optimized configuration of Speaker Recognition system for Broadcast News transcription.* In: 7th International Workshop on Electronics, Control, Modelling, Measurement and Signals, May 17–20, 2005, Tolouse, France.

[DAV04] David, P., Červa, P., Nouza, J.: *Speaker Recognition Applied for Enhanced Broadcast News Transcription.* In: Proc. of 14th Czech-German Workshop „Speech Processing", September 2004, Prague, Czech Republic, pp.72–76, ISBN 80-86269-11-6.

[ŽĎÁ04] Žďánský, J., David, P., Nouza, J.: *An Improved Preprocessor for the Automatic Transcription of Broadcast News Audio Stream.* In: Proc. of ICSLP 2004, October 2004, Jeju Island, Korea, pp.1065–1068, ISSN 1225-441x.

[NOU04] Nouza, J., Žďánský, J., David, P.: *Fully Automated Approach to Broadcast News Transcription in Czech Language.* In: Text, Speech and Dialogue. Lecture Notes in Artificial Intelligence. Springer-Verlag Berlin 2004, pp.401–408, ISBN 3-540-23049-1, ISSN 0302-9743.

[VAN04] Vandecatseye, A., Martens, J., Neto, J., Meinedo, H., Mateo, C., Dieguez, J., Mihelic, F., Zibert, J., Nouza, J., David, P., Pleva, M., Cizmar, A., Papageorgiou, H., Alexandris, C.: *The COST278 pan-European Broadcast News Database.* In: Proceedings of LREC 04, Lisboa, Portugal, May 2004, pp.873–876, ISBN 2-9517408-1-6.

[DAV03c] David, P.: *Using TRANSCRIBER tool for Broadcast News Transcription.* In: Proc. of 13th Czech-German Workshop „Speech Processing", Prague, September 2003, pp.116–120, ISBN 80-86269-10-8.

[DAV03b] 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, vol.IV, pp.372–376, ISBN 980-6560-01-9.

[DAV03a] David, P.: *Unsupervised Segmentation of Audio Recordings.* In: Proc. of 6th International Workshop on Elektronics, Control, Measurment and Signals-ECMS 2003. Liberec, June 2003, pp.17–20, ISBN 80-7083-708-X.

[DAV02c] David, P.: *Presentation of real-time system for automatic speaker identification.* In: Proc. of 12th Czech-German Workshop „Speech Processing", Prague, September 2002, pp.74–78, ISBN 80-86269-09-4.

[DAV02b] David, P.: *Diplomová práce – Rozpoznávání řečníka na základě hlasových charakteristik.* TU Liberec, Liberec 2002.

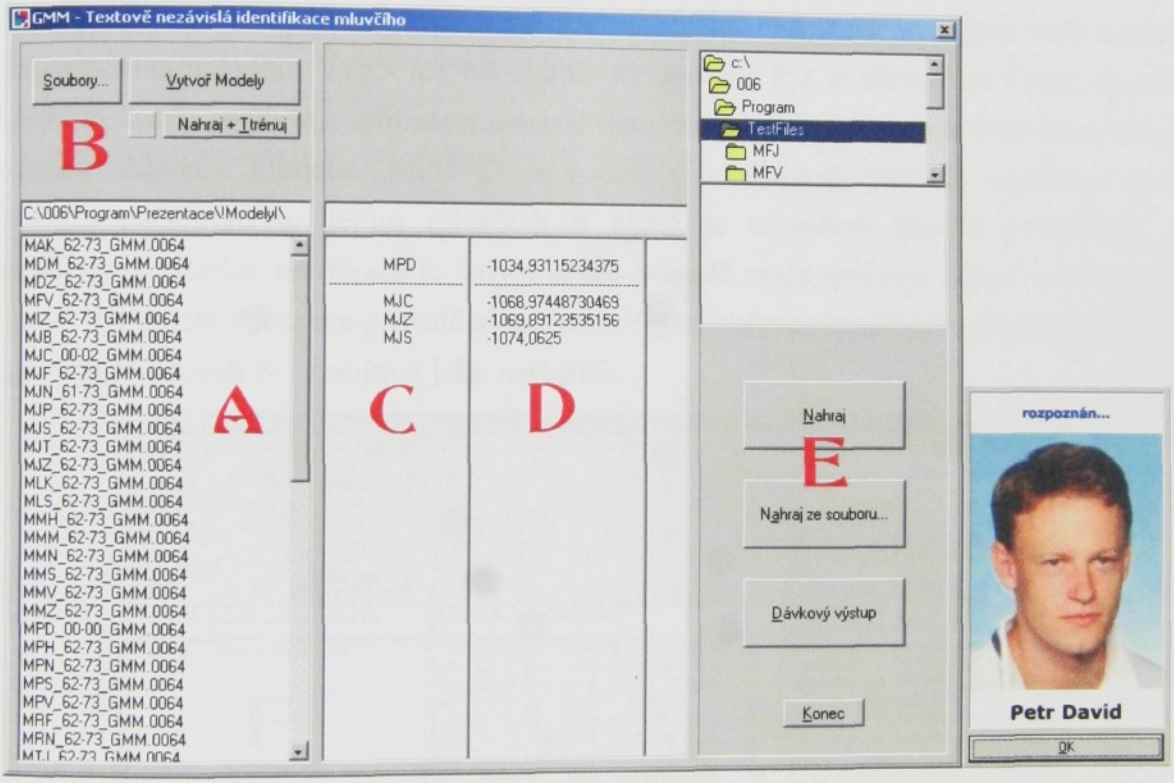
[DAV02a] David, P.: *Experiments with Speaker Recognition using GMM.* In: Proc. of Radioelektronika 2002,Bratislava,May 2002, pp.353–357,ISBN 80-227-1700-2.

[DAV01b] David, P., Nouza, J.: *Úloha rozpoznávání mluvčího*, In: Počítačové zpracování řeči – cíle, problémy, metody a aplikace, Liberec, December 2001, pp.95–105, ISBN 80-7083-551-6.

[DAV01a] David, P.: *Experiments with Speaker Recognition in Czech.* In: Proc. of 11th International Workshop "Speech Processing", Prague, September 2001, pp.59–62, ISBN 80-86269-07-8.

Příloha 1 – Demonstrační aplikace real-time IM a VM

Real-time systém pro identifikaci a verifikaci mluvčích vznikl v průběhu implementace metod pro IM a VM jako funkční aplikace ověřující chování rozpoznávacího softwaru v reálných podmínkách. V současné době je tento program zapůjčen softwarové firmě CIT k dlouhodobému testování, které má rozhodnout o spuštění pilotního projektu předcházejícího pozdějšímu reálnému nasazení aplikace. Grafické uživatelské rozhraní tohoto programu je k vidění na obrázcích P.1 (IM) a P.2 (VM).

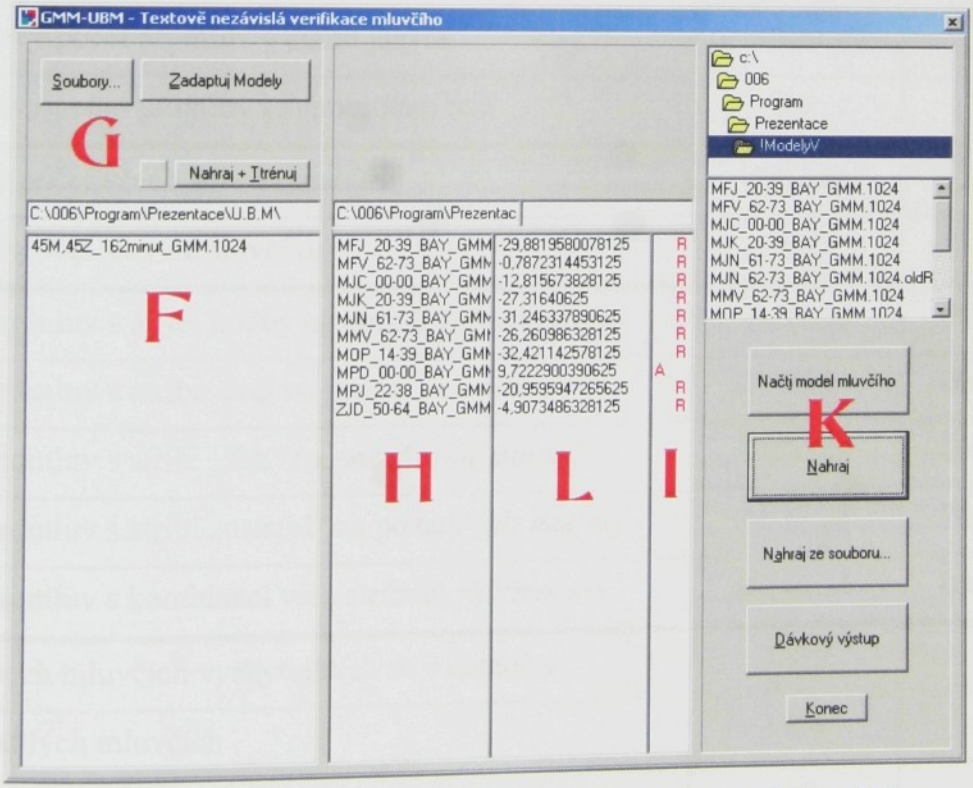


Obr. P.1: *Pohled na hlavní okno programu pro real-time IM*

Načtené GMM modely řečníků z databáze mluvčích jsou na obrázku označeny písmenem „A“, konkrétně se jedná o modely s 64 mixturami (určeny příponou souborů *GMM.0064). Tlačítka nad tímto oknem (sekce „B“), umožňují načtení dalších modelů, online záznam trénovací promluvy nového mluvčího nebo vytvoření modelu z této trénovací promluvy a okamžité zařazení do databáze, tzn. mluvčí může být ihned testován. Tlačítko „Nahraj“ v sekci „E“ zapíná VAD (Voice Activity Detector) detektor, který hledá začátek promluvy. V okamžiku, kdy tento začátek zachytí, spustí nahrávání. Po skončení promluvy sám vypne nahrávání a spustí vlastní identifikační proces. Ekvivalentně k předchozímu postupu, tlačítko „Nahraj ze souboru...“ vyvolá okno pro načtení offline

uložené nahrávky a její otestování. Písmena „C" a „D" označují místa, kam je vypsána identita a skóre identifikovaného mluvčího (pokud je k modelu přiřazena i fotografie mluvčího, je zobrazena zcela vpravo). Kromě vítězného řečníka jsou vypsáni ještě tři nejbližší alternativní mluvčí. Tak můžeme vizuálně zkontrolovat kritickou pravděpodobnost, v případě chyby pak možnou pozici mluvčího mezi alternativními kandidáty.

Obdobně jednoduše jako v případě IM vypadá i hlavní okno programu pro VM. Za zmínku stojí pouze několik odlišností. Do sekce „F" se načítá UBM model, zatímco adaptované modely mluvčích jsou načteny do sekce „H". V sekci „L" jsou vypsány rozdíly logaritmů věrohodností hypotéz H_0 a H_1. Nakonec je v „I" vypsán výsledek verifikace, kde písmeno „R" značí zamítnutí a písmeno „A" akceptaci testované promluvy vůči modelu v odpovídajícím řádku. Výpis jak ho vidíme na obrázku P.2 je netradiční v tom, že nám v jednom kroku poskytuje informaci nejen o testu proti proklamovanému mluvčímu (ten je označen zkratkou modelu „MPD"), ale současně ukazuje i výsledky verifikace proti ostatním načteným modelům mluvčích u nichž je testovaný mluvčí považován za narušitele. Výsledek verifikace je tedy v tomto případě zcela správný, neboť verifikovaný mluvčí byl pouze v jednom případě akceptován (proti svému vlastnímu modelu) a ve všech ostatních případech byl zamítnut jako narušitel.



Obr. P.2: *Pohled na hlavní okno programu pro real-time VM*

Příloha 2 – Struktura databáze COST278-BN

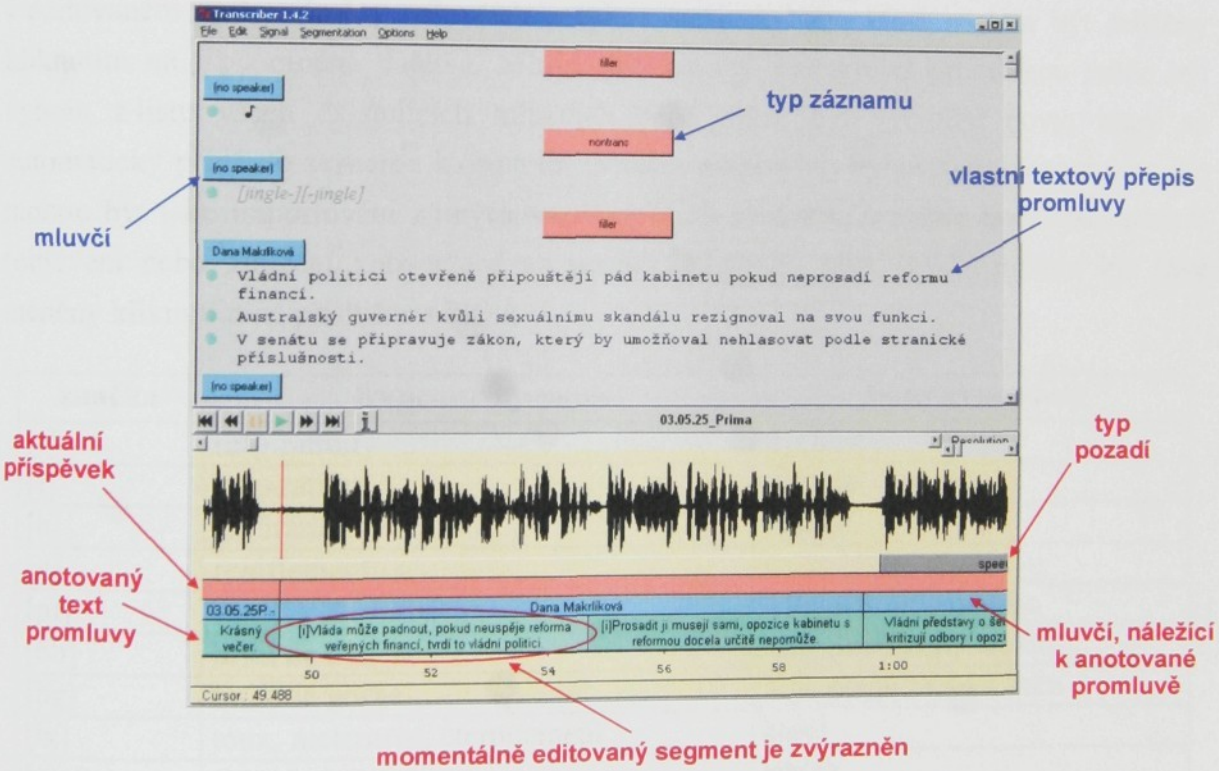
Tab. P.1: *Statistický pohled na COST278-BN databázi*

Textový popis významu statistických hodnot v následujících dvou sloupcích	**Globální hodnoty**	**Stat. české podskupiny**
Celková délka nahrávek v databázi *[hh:mm:ss]*	27:52:46	03:01:00
Délka upoutávkových segmentů *[hh:mm:ss]*	00:47:46	00:07:23
Délka report segmentů [hh:mm:ss]	24:30:59	02:49:35
Délka „nontrans" segmentů [hh:mm:ss]	02:34:00	00:04:02
Průměrná délka anotovaných promluv (vět) *[s]*	4,38	5,63
Celkový počet promluv (vět)	21356	1906
Celkový počet anotovaných promluv	20813	1886
Celkový počet promluv v cizím jazyce	129	15
Celkový počet promluv se simultánní řečí	86	8
Celkový počet všech slov v databázi	221643	26940
Celkový počet slov ve slovníku	60831	8983
Délka promluv s atrib. hudby na pozadí *[hh:mm:ss]*	02:19:22	00:19:56
Délka promluv s atribu. řeči na pozadí *[hh:mm:ss]*	01:39:57	00:12:19
Délka promluv s atrib. „shh" na pozadí *[hh:mm:ss]*	00:51:43	00:33:01
Délka promluv s atrib. „ostatní" na pozadí *[hh:mm:ss]*	04:37:38	00:31:06
Délka promluv s kombinací více atributů *[hh:mm:ss]*	02:45:54	00:00:30
Počet všech mluvčích vyskytujících se v databázi	1633	432
Počet rodilých mluvčích	1426	418
Počet nerodilých mluvčích	117	4
Počet mluvčích hovořících cizím jazykem	90	10

Počet mluvčích ženského pohlaví	442	128
Délka promluv mluvčích ženského pohlaví *[hh:mm:ss]*	09:21:26	01:05:51
Počet mluvčích mužského pohlaví	1092	285
Délka promluv mluvčích mužského pohlaví *[hh:mm:ss]*	14:33:55	01:45:06
Celková délka promluv s charakteristikou F0 *[hh:mm:ss]*	08:18:37	01:25:04
Celková délka promluv s charakteristikou F1 *[hh:mm:ss]*	03:20:58	00:23:54
Celková délka promluv s charakteristikou F2 *[hh:mm:ss]*	01:02:19	00:01:56
Celková délka promluv s charakteristikou F3 *[hh:mm:ss]*	01:46:17	00:16:25
Celková délka promluv s charakteristikou F4 *[hh:mm:ss]*	09:55:53	00:44:17
Celková délka promluv s charakteristikou F5 *[hh:mm:ss]*	00:15:02	00:01:18
Celková délka promluv s charakteristikou FX *[hh:mm:ss]*	00:39:37	00:04:01

Příloha 3 – Transcriber

Tento program určený pro anotaci rozsáhlých audionahrávek byl vyvíjen v prostředí interpretovaného jazyka TCL s podporou grafického rozhraní TK. Jazyk, který byl původně určen pro UNIX, je v současné době dostupný prakticky pro jakoukoliv platformu včetně MS WINDOWS, což umožňuje snadnou přenositelnost. Protože vývoj probíhal na základě licence GNU[1] (General Public License), jedná se o volně šiřitelný software. Pro účely přepisu audiozáznamu jsou do Transcriberu implementována transkripční pravidla, která jsou specifikována konsorciem LDC. Kromě přenositelnosti má Transcriber další významný klad, kterým je podpora Unicode kódování. Díky tomu je možné v souboru s přepisem zpravodajství kromě národního jazyka, ve kterém je zpravodajství pořízeno, použít téměř jakýkoliv jiný jazyk podporovaný standardem Unicode. Nativním formátem pro ukládání souborů s přepisy audiosouborů je XML.



Obr. P.3: *Ukázka hlavního okna programu Transcriber. Uživatelská část je vertikálně rozdělena na tři sekce. V horní části je zapisována ortografická transkripce, jména mluvčích, názvy reportáží a typy segmentů. Uprostřed je umístěno okno pro práci s audiosignálem (selekce, zvětšování/zmenšování, posun). Dole je pak synchronně se střední částí zobrazen výsledek segmentace do čtyř nezávislých proudů.*

[1] Transcriber je možné stáhnout z internetových stránek „http://www.etca.fr/CTA/gip/Projets/Transcriber/".

Na obrázku P.3 je znázorněno hlavní okno programu Transcriber. Uživatelské rozhraní je rozděleno na tři hlavní části: sekce pro editaci textu, sekce pro navigaci v audiosignálu a sekce zobrazující stav segmentace do čtyř nezávislých proudů. Všechna tři okna nám v podstatě poskytují náhled na přepisovaný audiosoubor z různých úhlů pohledu. Zobrazení jsou navzájem synchronizována tak, abychom byli v každém okamžiku schopni velmi rychlé navigace v přepisovaném záznamu. Pokud je například v průběhu přehrávání zvuku kurzor přesunut do jiného segmentu, je zvuk automaticky zastaven a spuštěn od nově vybraného segmentu. To je užitečné zejména při kontrole již hotových přepisů, kdy je velmi snadné přejít, případně znovu poslechnout určitý segment pouhým kliknutím myši. Kompletní přepis audiozáznamu neobsahuje pouze ortografickou transkripci, ale také všechny informace o změnách mluvčích, jejich jménech, sekcích, akustických podmínkách, názvech reportáží a podobně. Nová sekce je v textovém okně indikována tlačítkem uprostřed řádku spolu s názvem tématu. Změna mluvčích je znázorněna tlačítkem nalevo od textu a změněné akustické podmínky jsou reprezentovány ikonou noty přímo v editovaném textu. Změny mluvčích a sekce mají atributy, které mohou být měněny kliknutím na již zmíněná tlačítka. Mluvčí asociovaný s konkrétní promluvou může být vybrán z listu všech existujících mluvčích nebo může být vytvořen nový, který se automaticky přidá do seznamu k ostatním. Mluvčí mohou být vyhledáváni v seznamu, ale mohou být také importováni z jiných transkripčních souborů. Zvukové podmínky pozadí (objevení nebo zmizení konverzace na pozadí, jakýkoliv hluk apod.) mohou být také měněny kliknutím na příslušnou ikonu.

značka	francouzský popis	český ekvivalent
[r]	respiration	dýchání
[i]	inspiration	nádech
[e]	expiration	výdech
[n]	reniflement	nosní hluky
[pf]	soufflé	dýchnutí na mikrofon
[bb]	bruit de bouche	hluk pocházející ze rtů
[bg]	bruit de gorge	hluk pocházející z krku
[tx]	toux, raclement, eternuement	kašel
[rire]	rires du locuteur	smích
[sif]	sifflement du locuteur	šepot
[b]	bruit indetermine	neznámý hluk
[conv]	conversations de fond	promluva na pozadí
[paf]	froissement de papiers	hluk pocházející od papírů
[shh]	soufflé electrique	hluk od přenosové cesty
[mic]	bruits micro	mikrofonní hluk

Tab. P.2: *Značky pro ruchy standardně nabízené programem Transcriber*

Některé zvuky nemohou být rozumě přepsány do textové podoby[1]. Transcriber řeší tento problém speciálními kódy zobrazovanými v hranatých závorkách (seznam nabízených kódů je uveden v tabulce P.2). Jako příklad můžeme uvést [i] pro nádech v průběhu řeči. Další neřečové události, které trvají delší dobu a zasahují do řeči (stává se velmi často), popisujeme jiným způsobem. Jako příklad si můžeme vzít úvodní znělku zpravodajství nazývanou „jingle". Její začátek a konec bychom vyznačili „[jingle-] ... [-jingle]". V současné verzi Transcriberu nejsou tyto značky zatím časově synchronizovány se zvukem, ale autoři o tom do budoucna uvažují. Nová událost může být na současnou pozici kurzoru vložena pomocí položek z menu, nebo klávesovou zkratkou (přednastavena je klávesa „enter", která vytváří nový řádek v textovém editoru). Protože je tato funkce dostupná i při přehrávání audia, jde velice snadno vytvořit hrubá segmentace audiozáznamu pouhými stisky klávesy „enter" při spojitě přehrávaném zpravodajství.

[1] Do této kategorie spadá většina lidských „neřečových" zvuků (nádechy, výdechy, zakašlání), ruchy vzniklé při pořizování nahrávky apod.

Příloha 4 – Výsledky evaluační kampaně COST278-BN

Tab. P.3: *Výsledky testování v kategoriích C1 + T1*

	ELIS			LJU			UMB			TUL		
	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC
BE	97,8	77,9	97,0	98,9	49,7	96,8	98,7	61,8	97,2	99,1	61,5	97,6
CZ	98,4	93,5	98,1	98,3	82,0	97,4	99,5	77,6	98,2	99,5	76,0	98,2
GA	96,4	84,6	95,9	98,6	42,3	96,3	95,2	89,9	95,0	97,4	94,4	97,3
GR	96,7	67,3	93,3	98,3	30,8	90,5	98,4	32,9	90,8	98,2	63,6	94,2
HR	95,5	71,9	94,2	98,4	73,5	97,0	95,8	56,9	93,6	98,2	63,6	94,2
HU	97,3	72,5	95,1	97,2	58,2	93,6	98,4	53,5	94,3	97,3	64,5	94,3
PT	98,6	47,4	96,1	98,1	32,8	94,8	98,9	49,4	96,4	99,5	43,2	96,7
SI	97,4	87,1	96,8	97,9	73,5	96,4	97,3	86,7	96,6	98,1	91,5	97,7
SI2	98,5	66,6	98,2	98,4	72,5	98,2	97,1	73,6	96,9	97,8	83,2	97,7
SK	98,7	65,3	94,5	99,3	25,3	89,8	99,0	53,7	93,2	99,0	65,5	94,8
mean	97,5	73,4	**95,9**	98,3	54,1	**95,1**	97,8	63,6	**95,2**	98,4	70,7	**96,3**

Americké BN	8h. slovinských BN	8,7h. slovin. BN	1,8h. českých BN

	TUB			INESC			UVIGO		
	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC
BE	95,7	87,3	95,4	93,9	74,0	93,1	92,8	74,3	92,0
CZ	96,7	85,4	96,3	96,4	87,8	96,0	99,3	69,1	97,7
GA	93,3	91,3	93,2	92,0	78,4	91,4	93,6	95,5	93,7
GR	93,6	76,6	91,6	89,9	76,3	88,3	99,3	31,9	91,4
HR	91,9	96,8	92,2	89,6	99,0	90,2	96,7	80,0	95,8
HU	93,7	80,8	92,5	88,8	76,2	87,7	92,3	71,9	91,0
PT	94,7	62,5	93,1	95,3	46,3	92,9	90,8	54,7	89,1
SI	95,4	93,2	95,2	94,1	83,2	93,5	98,5	72,2	96,7
SI2	94,0	84,9	93,9	95,4	80,8	95,3	99,0	37,3	98,1
SK	95,3	71,1	92,2	93,3	58,3	88,9	92,1	55,0	87,1
mean	94,4	83,0	**93,6**	92,9	76,0	**91,7**	95,4	64,2	**93,3**

2h. řeč + 1h. neřeč	46h. portugal. BN

Tab. P.4: *Výsledky testování v kategoriích C2 + T1*

LJU

	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC
BE	96,6	69,7	92,4	95,3	56,6	93,7	97,3	52,7	95,5	90,5	49,8	88,8
CZ	98,6	80,1	97,6	98,3	72,9	96,8	96,6	62,6	94,6	96,0	87,2	95,5
GA	69,3	43,0	68,2	96,5	82,0	95,9	87,1	19,9	84,4	85,0	80,6	84,9
GR	94,8	61,0	90,8	96,7	51,2	91,4	94,7	27,4	86,8	89,0	65,9	86,3
HR	90,1	72,0	89,1	96,1	70,4	94,6	84,0	68,1	83,0	83,6	96,9	84,3
HU	94,7	72,2	92,6	95,9	67,0	93,2	93,3	45,4	88,9	87,3	80,5	86,7
PT	96,2	38,7	93,4	95,2	42,9	92,6	92,0	51,8	90,0	93,0	40,1	90,4
SI	94,7	86,4	94,2	93,0	91,2	92,9	96,5	68,5	94,8	90,3	91,2	90,4
SI2	96,0	63,5	95,7	96,5	69,4	96,3	97,1	58,3	96,7	92,9	64,8	92,6
SK	98,6	50,2	92,4	98,0	56,6	92,3	86,2	37,4	80,0	93,5	62,2	89,5
mean	93,0	63,7	**90,6**	96,2	66,0	**94,0**	92,5	49,2	**89,5**	90,1	71,9	**88,9**
train	BE			GA			PT			SK		

TUK

	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC
BE	99,4	31,3	96,6	90,4	27,5	87,8	99,5	17,3	96,1	96,1	52,4	94,3
CZ	99,6	40,8	96,2	94,1	78,0	93,2	99,6	40,2	96,2	93,8	74,1	92,7
GA	98,8	40,7	94,1	94,5	63,7	92,0	99,3	39,1	94,4	90,8	62,7	88,5
GR	98,3	48,3	92,3	90,2	61,4	86,8	99,2	36,3	91,7	90,0	65,0	87,1
HR	99,3	44,2	94,8	92,6	73,3	91,1	99,4	51,0	95,4	91,6	66,1	89,5
HU	99,2	39,7	93,1	91,0	60,8	87,9	99,3	34,5	92,7	94,3	68,3	91,7
PT	99,6	32,8	95,1	89,6	40,2	86,4	99,5	29,3	94,9	91,8	51,8	89,2
SI	99,1	54,6	96,4	92,9	62,1	91,0	99,1	53,9	96,2	93,4	73,2	92,1
SI2	99,6	33,9	99,0	91,5	29,7	91,0	99,7	34,2	99,2	93,5	70,3	93,3
SK	99,4	36,5	91,1	93,7	37,2	86,2	99,5	36,4	91,1	90,8	62,7	87,1
mean	99,2	40,3	**94,9**	92,1	53,4	**89,3**	99,4	37,2	**94,8**	92,6	64,7	**90,6**
train	BE			GA			PT			SK		

TUL

	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC	řeč	neřeč	ACC
BE	99,0	74,1	98,0	99,3	21,1	96,1	99,1	59,7	97,5	98,6	50,8	96,7
CZ	99,4	45,2	96,4	99,4	74,7	98,0	98,0	87,4	97,4	93,6	93,7	93,6
GA	97,6	76,0	96,7	99,7	72,2	98,6	97,5	92,9	97,3	97,3	53,7	95,5
GR	98,1	58,9	93,5	98,4	44,2	92,1	98,3	53,8	93,1	99,1	23,2	90,3
HR	96,2	85,1	95,6	99,0	53,3	96,4	98,0	41,6	94,7	96,9	22,6	92,6
HU	97,0	65,1	94,1	98,9	45,5	94,0	96,1	65,9	93,4	98,2	56,6	94,4
PT	99,3	44,5	96,6	99,8	16,8	95,8	99,6	44,6	96,9	87,0	61,9	85,7
SI	97,8	92,6	97,5	99,4	78,1	98,1	98,4	79,1	97,2	94,4	90,9	94,2
SI2	97,8	86,5	97,7	99,7	34,2	99,2	98,9	73,4	98,7	98,4	75,3	98,2
SK	99,2	43,4	92,0	96,8	50,6	90,9	93,6	68,7	90,4	98,9	39,3	91,3
mean	98,1	67,1	**95,8**	99,0	49,1	**95,9**	97,8	66,7	**95,7**	96,2	56,8	**93,3**
train	BE			GA			PT			SK		

Tab. P.5: *Výsledky testování v kategoriích C1 + T3*

	ELIS			LJU			UMB		
	muži	ženy	ACC	muži	ženy	ACC	muži	ženy	ACC
BE	91,9	96,3	93,0	93,0	95,2	96,5	99,9	70,2	92,5
CZ	88,6	98,2	92,2	92,2	93,0	87,3	98,3	95,4	97,2
GA	86,1	96,2	91,6	91,6	68,8	96,3	87,8	95,1	91,8
GR	91,2	98,7	93,8	93,8	90,4	96,8	98,4	89,7	95,4
HR	98,4	91,9	95,9	95,9	83,6	94,4	98,1	90,3	95,1
HU	95,6	97,5	96,3	96,3	93,2	92,9	96,0	96,7	96,3
PT	96,3	95,5	96,1	96,1	85,8	97,8	97,7	93,4	96,7
SI	87,4	96,7	91,4	91,4	85,3	97,9	86,8	97,9	91,6
SI2	95,4	97,4	96,1	96,1	90,9	95,8	99,1	94,5	97,6
SK	97,7	97,2	97,5	97,5	86,8	99,3	97,6	99,2	98,3
mean	92,9	96,6	**94,4**	87,3	96,1	**91,7**	96,0	92,2	**95,3**

	TUL			INESC			UVIGO		
	muži	ženy	ACC	muži	ženy	ACC	muži	ženy	ACC
BE	95,5	98,2	96,1	98,4	84,1	94,9	95,1	94,9	95,0
CZ	93,0	98,0	94,9	98,1	91,8	95,7	96,2	97,7	96,8
GA	86,1	98,1	92,6	91,3	95,4	93,5	84,1	95,0	89,9
GR	96,8	94,2	95,9	96,3	95,1	95,9	94,1	97,6	95,7
HR	97,3	95,5	96,6	98,7	76,2	90,1	98,7	91,3	96,0
HU	92,4	97,9	94,5	99,3	84,8	94,0	98,4	93,2	96,1
PT	93,5	98,6	94,7	97,7	94,8	97,1	98,2	99,5	98,4
SI	80,3	98,4	88,1	88,2	92,8	90,1	84,1	94,7	88,0
SI2	97,2	98,5	97,6	96,0	93,6	95,2	99,1	98,8	99,0
SK	97,3	99,9	98,4	99,4	91,4	96,1	99,8	99,7	99,8
mean	92,9	97,7	**94,9**	96,3	90,0	**94,3**	94,8	96,2	**95,5**

Tab. P.6: *Výsledky testování v kategoriích C2 + T3*

LJU

	muži	ženy	*ACC*	muži	ženy	*ACC*	muži	ženy	*ACC*	muži	ženy	*ACC*
BE	92,7	86,7	91,3	88,5	89,1	88,6	98,9	82,5	94,9	91,7	92,0	91,8
CZ	92,1	89,4	91,1	93,0	90,9	92,2	95,8	85,6	91,9	90,9	92,8	91,6
GA	78,8	81,1	79,7	82,4	92,8	88,0	97,3	86,8	91,6	86,3	79,4	82,7
GR	91,6	78,4	87,1	87,3	93,6	89,5	93,7	81,1	89,3	91,4	91,4	91,4
HR	89,8	81,3	86,9	89,5	88,3	89,0	91,1	71,3	82,7	93,1	77,8	87,5
HU	94,5	79,9	89,3	92,9	90,0	91,8	96,5	81,9	91,0	93,9	83,9	90,5
PT	91,0	91,5	91,1	82,6	97,2	86,0	90,5	86,4	89,5	79,4	87,3	81,3
SI	85,0	93,0	88,5	80,0	97,8	87,8	87,3	89,6	88,3	79,3	91,7	84,6
SI2	90,2	92,5	91,0	88,2	95,5	90,6	96,6	79,1	90,7	88,5	89,3	88,8
SK	86,7	93,7	89,7	92,6	96,0	94,0	95,9	88,3	92,5	89,2	98,4	93,0
mean	89,2	86,8	**88,6**	87,7	93,1	**89,8**	94,4	83,3	**90,2**	88,4	88,4	**88,3**
train	BE			GA			PT			SK		

TUL

	muži	ženy	*ACC*	muži	ženy	*ACC*	muži	ženy	*ACC*	muži	ženy	*ACC*
BE	94,3	97,1	95,0	93,9	97,2	94,7	97,2	96,3	97,0	98,2	85,1	94,9
CZ	91,7	97,3	93,8	87,9	97,6	91,6	94,1	97,8	95,5	94,6	97,2	95,6
GA	95,1	96,7	96,0	84,0	97,3	91,4	84,5	98,2	92,1	86,6	95,6	91,5
GR	88,0	97,9	91,4	94,7	91,7	93,6	93,6	93,5	93,6	97,1	91,2	95,1
HR	91,9	91,2	91,6	95,8	94,5	95,3	95,0	93,8	94,5	94,7	90,7	93,2
HU	91,6	98,4	94,2	90,6	96,8	92,9	92,2	97,7	94,3	92,0	97,7	94,2
PT	87,1	96,6	89,3	92,7	96,9	93,7	95,9	97,9	96,3	97,0	94,6	96,5
SI	79,0	95,8	86,2	74,7	98,2	84,7	81,7	96,7	88,1	86,6	97,3	91,2
SI2	91,5	92,7	91,9	94,8	98,2	95,9	95,4	97,0	95,9	98,1	96,1	97,4
SK	97,0	99,5	98,0	95,2	99,1	96,9	95,5	97,6	96,4	97,5	99,6	98,4
mean	90,7	96,3	**92,7**	91,1	96,8	**93,1**	92,5	96,7	**94,4**	94,2	94,5	**94,8**
train	BE			GA			PT			SK		