VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DETEKCE ZMĚNY JAZYKA PŘI HOVORU

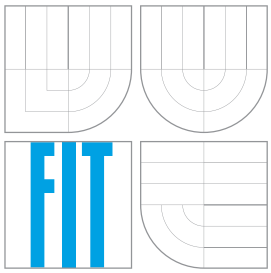
DIPLOMOVÁ PRÁCA
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

FILIP POVOLNÝ

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

DETEKCE ZMĚNY JAZYKA PŘI HOVORU

CODE SWITCHING DETECTION IN SPEECH

DIPLOMOVÁ PRÁCA
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

FILIP POVOLNÝ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. PAVEL MATĚJKA, Ph.D.

BRNO 2015

Abstrakt

Tato práce se zabývá problematikou detekce změny jazyka při hovoru. V první části jsou popsány v současnosti používané metody diarizace jazyků. K implementaci byla vybrána metoda založená na akustickém přístupu identifikace jazyka s využitím směsi Gaussovských rozložení, i-vektoru a lineární diskriminační analýzy. Pro experimenty byla vytvořena mandarínsko-anglická databáze se střídáním jazyků. Na této databázi zvolený systém dosahuje úspěšnosti 89,3 % správně klasifikovaných segmentů.

Abstract

This master's thesis deals with code-switching detection in speech. The state-of-the-art methods of language diarization are described in the first part of the thesis. The proposed method for implementation is based on acoustic approach to language identification using combination of GMM, i-vector and LDA. A new Mandarin-English code-switching database was created for these experiments. Using this system, accuracy of 89,3 % is achieved on this database.

Klíčová slova

střídání kódů, diarizace jazyků, identifikace jazyka, rozpoznávání jazyků

Keywords

code switching, language diarization, language identification, language recognition

Citace

Filip Povolný: Detekce změny jazyka při hovoru, diplomová práca, Brno, FIT VUT v Brně, 2015

Detekce změny jazyka při hovoru

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením pana Ing. Pavla Matějky, Ph.D. a uvedl všechny použité zdroje a publikace, z kterých jsem čerpal.

.....
Filip Povolný
27. mája 2015

Poděkování

Chtěl bych se poděkovat vedoucímu práce Ing. Pavlovi Matějkovi, Ph.D. za cenné rady na konzultacích v průběhu psaní této zprávy.

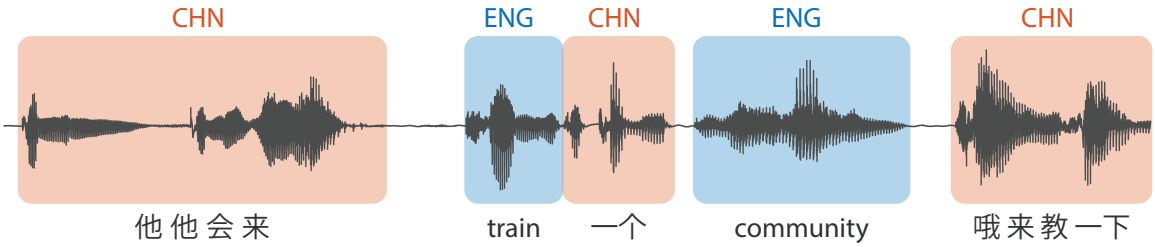
Obsah

Kapitola 1

Úvod

Zmena jazyka pri hovore sa v lingvistike nazýva *striedanie kódov* (angl. *code switching*[1]). Ide o jav, kedy rečník môže počas konverzácie striedať dva a viac jazykov alebo jazykových útvarov. Toto je bežné v multilingválnej spoločnosti, napr. v Indii, kde sa používa zmes hindčiny a angličtiny [12], na Taiwane zas kombinácia mandarínčiny, taiwančiny a angličtiny [3], alebo v Hong Kongu striedanie kantončiny a angličtiny [4]. Ďalším z príkladov je situácia, kedy sekretárka príjme hovor napr. v angličtine, no po prepojení pokračuje rozhovor v inom jazyku.

Existuje niekoľko typov striedania jazykov, no v tejto práci je riešené striedanie jazykov v rámci jednej vety (intra-sentential switching). Tento typ je najrozšírenejší a tiež sa ním zaoberá najviac štúdií. Obrázok 1.1 ukazuje úryvok reči, kde sa v rámci jednej vety striedajú jednojazyčné segmenty: čínština, angličtina, čínština, angličtina a čínština s dĺžkou 1,2; 0,35; 0,4; 0,75 a 0,7 sekundy v tomto poradí.



Obr. 1.1: Príklad reči so striedaním angličtiny a čínštiny spolu s označením a prepisom jednojazyčných segmentov.

Cieľom tejto práce je vytvoriť systém, ktorý detekuje miesto v nahrávke, kde rečník prepne z jedného jazyka do druhého, tzv. *changepoint*. Následne sú použité jazyky identifikované. Táto úloha, nazývaná aj *diarizácia jazyka*, je podobná úlohe identifikácie jazyka a preberá z nej mnoho postupov. Hlavným rozdielom medzi diarizáciou a identifikáciou jazyka je, že v rámci jednej nahrávky sa môže vyskytovať dva alebo viac jazykov. Jednojazyčné segmenty tvoriace túto nahrávku sú naviac niekoľkonásobne kratšie (1 –3 s) ako bežné nahrávky pre identifikáciu jazyka. Tabuľka 1.1 ukazuje prudký rast chyby identifikácie jazyka pri skracovaní segmentov.

V úlohe diarizácie jazyka sa bežne stretávame s jednojazyčnými úsekmi kratšími ako jedna sekunda, ktorých identifikácia je ešte náročnejšia kvôli nedostatku informácií v tomto

[1] V niektorých prípadoch je používaný výraz *code mixing* ako synonymum k výrazu code switching.

Dĺžka segmentu [s]	30 s	10 s	3 s
NIST LRE11 [%]	3,01	6,49	14,96

Tabuľka 1.1: Chyba identifikácie jazyka systémom z [1] na nahrávkach rôznej dĺžky v súťaži NIST LRE 2011.

segmente. K tomu prispieva aj silný prízvuk rečníka v úsekoch, kde používa sekundárny jazyk.

Údaje o hraniciach a identitách jazykov je možné využiť na zlepšenie úspešnosti automatického rozpoznávania súvislej reči na nahrávkach so striedaním kódov. Rozpoznávače reči štandardne predpokladajú jednojazyčný vstup. Napr. pri použití anglického rozpoznávača je nutné označiť a odstrániť nežiadúce segmenty s cudzím jazykom a spracovať len tie anglické. Rozpoznávač by inak spracoval aj cudzojazyčné segmenty a rozpoznal by ich chybne. V jazykovom modeli sa potom tieto chyby prenesú na okolité anglické segmenty, ktoré môžu byť tiež rozpoznané chybne. To vedie k značnému zníženiu celkovej úspešnosti rozpoznávača na viacjazyčnej reči.

V kapitole 2 sú porovnané metódy diarizácie jazyka používané v súčasnosti. Kapitola 3 popisuje základné princípy identifikácie jazyka. V kapitole 4 je popísaný implementovaný akustický systém diarizácie jazyka. Kapitola 5 bližšie popisuje použité dáta a experimenty s týmto systémom spolu s výsledkami. Kapitola 6 obsahuje zhrnutie dosiahnutých výsledkov a načrtáva ďalší postup práce.

[1] 2011 NIST Language Recognition Evaluation: http://www.nist.gov/itl/iad/mig/lre11results.cfm

Kapitola 2

Súčasné metódy diarizácie jazyka

V tejto kapitole je uvedený prehľad metód, ktoré sa v súčastnosti a nedávnej minulosti venujú diarizácií jazyka. Diarizácie jazyka je síce úloha, ktorá patrí k vedľajším a teda menej skúmaným úlohám spracovania reči, ale v poslednej dobe sa jej dostáva stále viac pozornosti. Dôvodom je narastajúce množstvo multilingválnej komunikácie a masívne využívanie systémov spracovania reči v každodennom živote. V súčasnosti je na diarizáciu jazyka zameraných mnoho výskumov, väčšina z Ázie, kde je striedanie jazykov používané denne. V závere kapitoly sú zhrnuté poznatky zo všetkých popísaných metód.

2.1 Detekcia hraníc jazyka s využitím bi-fonémových pravdepodobností

V roku 2004 Chan ai. prezentovali metódu detekcie hraníc jazyka založenú na bi-fonémovom rozpoznávaní jazyka [2]. Fonémový rozpoznávač je natrénovaný pomocou HTK na jednojazyčných korpusoch CUSENT pre kantónčinu a TIMIT pre angličtinu. Kantónčina je v tomto prípade dominantný jazyk, takže väčšina anglických foném zdieľa model s podobnými kantónskymi fonémami. Systém používa iba kantónsky pravdepodobnostný bi-fonémový model. Výpočítaná pravdepodobnosť potom udáva mieru, s akou daná fonéma patrí do kantónčiny. Anglické úseky sú identifikované, len ak je pravdepodobnosť nižšia ako zvolený prah.

Systém je vyhodnotený na databáze so striedaním jazykov nahranej jediným rečníkom, kde v každej nahrávke je jeden anglický segment. V 76 % prípadov systém správne detekuje hranice tohto segmentu s odchýlkou do 0.3 s. Kvôli použitiu jediného bi-fonémového modelu systém dosahuje najlepšie výsledky na kantónskej monolinguálnej reči, no zlyháva na čistej angličtine.

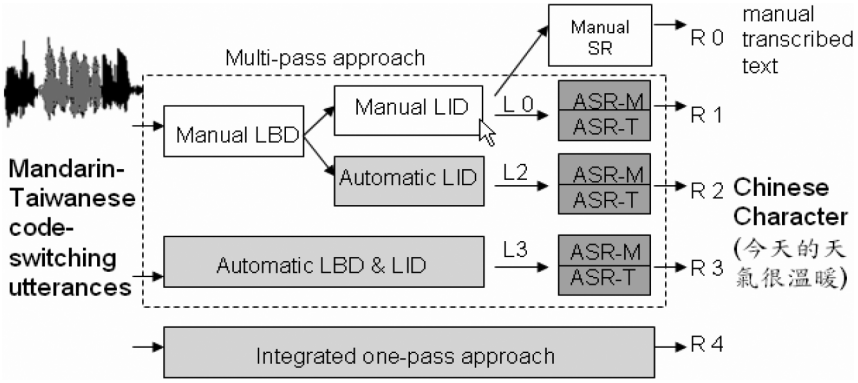
2.2 Rozpoznávanie reči so striedaním čínskych dialektov

Obvyklý postup pri rozpoznávaní reči so striedaním jazykov obsahuje detektor hraníc jednojazyčných segmentov a identifikátor jazyka v týchto seqmentoch. Následne je na jednotlivé segmenty použitý jednojazyčný rozpoznávač reči.

Roku 2006 predstavujú Lyu ai. metódu [7], ktorá využíva skutočnosť, že všetky čínske dialekty až na niekoľko výnimiek zdieľajú spoločnú formu zápisu. Reč so striedaním jazykov je v tomto prípade spracovaná dvojjazyčným rozpoznávačom reči. Ten obsahuje dvojjazyčný akustický, výslovnostný a jazykový model. Hranice ani identita jazykov nie sú známe. Zmes jazykov je vnímaná ako jeden jazyk, ktorý je možné v jednom priechode previesť na

postupnosť čínskych znakov – slabík. Výhodou je, že s jedným znakom môže byť spojených niekoľko výslovností v rôznych jazykoch vrátane japončiny, kórejčiny alebo vietnamčiny, ktoré tiež využívajú čínske znaky.

Systém je natrénovaný na jednojazyčných taiwanských a mandarínskych dátach. Na testovanie sú použité dáta, kde je hlavným jazykom mandarinčina, do ktorej sú vložené krátke taiwanské frázy. Výkonnosť rozpoznávača je daná mierou chybne rozpoznaných čínskych znakov (CER – Chinese character Error Rate). V práci je porovnaný navrhovaný jednopriechodový systém s niekoľkými variantami tradičného viacpriechodového systému. Všetky systémy sú zobrazené na obrázku 2.1 a v tabuľke 2.1 možno vidieť, že jednopriechodový systém (R4) dosahuje na testovacích dvojjazyčných dátach najmenšiu chybovosť.



Obr. 2.1: Porovnávané systémy z [7]: Výstup **R0** je referenčný ručný prepis reči. **R1**–**R3** sú výstupy tradičných viacpriechodových rozpoznávačov reči, ktoré majú k dispozícii dáta anotované na rôznej úrovni. **R4** je výstup plne automatického jednopriechodového systému.

CER [%]	Slovná zásoba	
Subsystém	10 K	20 K
R1	14,22	22,59
R2	20,7	28,61
R3	23,2	31,76
R4	**13,31**	**20,02**

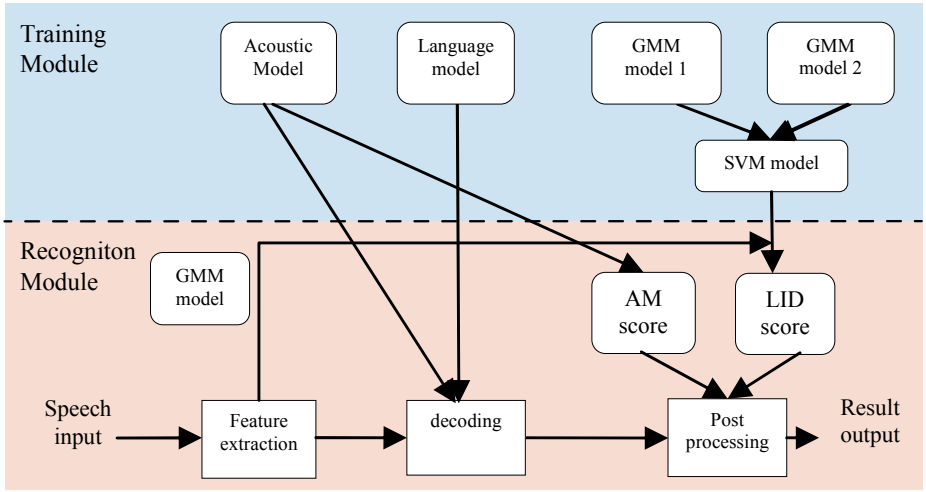
Tabuľka 2.1: Porovnanie výkonnosti jednotlivých podsystémov rozpoznávania reči s použitím rôzne veľkých slovných zásob.

2.3 Rozpoznávanie reči so striedaním kódov s využitím jazykových a akustických informácií

Zhang v roku 2012 navrhol systém [15], ktorý sa zameriava na identifikáciu jazyka v mandarínsko-anglickej reči. K tomu využíva spojenie akustického a jazykového modelu. Bloková schéma systému je zobrazená na obrázku 2.2.

Akustický model je riešený pomocou GMM pre každý jazyk a SVM klasifikátoru. Jazykový model využíva bi-gramy alebo tri-gramy slov a slovník vytvorený zjednotením slovníkov oboch jazykov. Váženou kombináciou skóre z oboch podsystémov je získané výsledné rozhodnutie.

Na trénovanie boli použité dáta z jednojazyčných databáz CASIA98-99 pre mandarínčinu a Resource Management pre angličtinu. Systém bol natrénovaný a testovaný na konverzačnom korpuse SEAME[1], ktorý obsahuje 63 hodín reči so striedaním mandarínčiny a angličtiny. Na vyhodnotenie výkonnosti systému bola zvolená metrika WER – Word Error Rate. Tá hodnotí kvalitu prepisu viacjazyčnej reči a je daná podielom $\frac{I+D+S}{N}$, kde I je počet vložených slov, D počet nerozpoznaných slov, S počet chybne rozpoznaných slov a N počet všetkých slov.



Obr. 2.2: Schéma systému využívajúceho akustické a jazykové informácie.

Metóda	LB detection rate [%]	WER [%]
Bi-phone probability (knowledge-based)	69,62	35,3
Bi-phone probability (data driven)	76,54	28,9
Combination of LID and AM information	82,1	22,7

Tabuľka 2.2: Porovnanie WER navrhovaného systému s dvomi variantami systému popísaného v podkapitole 2.1. Druhý stĺpec obsahuje podiel správne detekovaných hraníc jazyka.

2.4 Diarizácia jazyka v konverzácií so striedaním kódov

V roku 2013 Lyu ai. predstavujú metódu diarizácie jazyka, ktorá spája akustický a fonotaktický prístup identifikácie jazyka [6]. Sú využité ako akustické, tak fonotaktické príznaky z dlhodobého kontextu (z postupnosti niekoľkých foném). Rovnako ako v predchádzajúcej podkapitole, aj tento systém rieši diarizáciu mandarínčny a angličtiny.

V akustickom podsystéme zobrazenom na obrázku 2.3a je každá trieda (jazyk) modelovaná jedným GMM. Pre každú fonému sú vypočítané logaritmické vierohodnosti oboch GMM a klasifikované pomocou SVM.

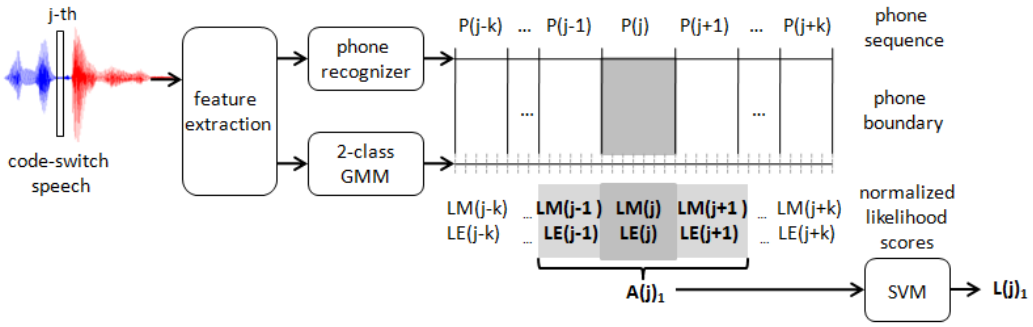
Fonotaktický systém zobrazený na obráku 2.3b využíva trifónový rozpoznávač založený na HMM. Postupnosti trifónov sú klasifikované pomocou podmienených náhodných polí (CRF – Conditional Random Fields).

[1]South-East Asia Mandarin-English: https://catalog.ldc.upenn.edu/LDC2015S04

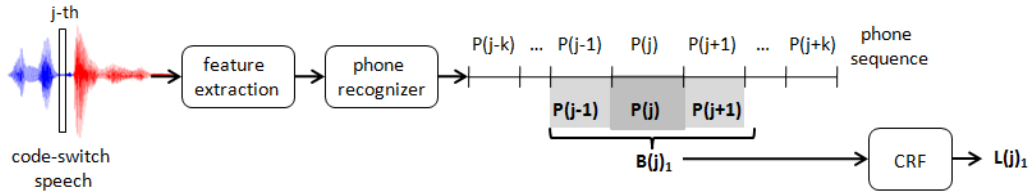
Dĺžka segmentu	0,1 - 0,5 s	0,5 - 1 s	1 - 3 s	3 - 9 s
Zlepšenie EER	5,2 %	13,8 %	15,1 %	17,9 %

Tabuľka 2.3: Zlepšenie EER[2] identifikácie jazyka na jednojazyčnej reči rôznej dĺžky použitím navrhovaného systému oproti jednojazyčným LID systémom.

Systém bol testovaný na korpuse SEAME popísanom v predchádzajúcej podkapitole. Na reči so striedaním jazykov systém správne identifikuje jazyk s chybou 14,7 % na rámec (FER – Frame Error Rate). Na jednojazyčných segmentoch extrahovaných z rovnakých dát dosahuje systém v identifikácií jazyka nasledujúce zlepšenia:



(a) Akustický systém pre každú identifikovanú fonému j vypočíta $LM(j)$ a $LE(j)$ označujúce logaritmickú vierohodnosť pre každú triedu (mandarínčinu a angličtinu). Z postupností týchto hodnôt $A(j)_k$ s dĺžkou $2k+1$ je pomocou SVM určená výsledná identita jazyka $L(j)_k$.



(b) Fonotaktický systém pomocou fonémového rozpoznávača transformuje vstupnú reč na postupnosť foném $P(j)$, $j \in \langle 1, N \rangle$. Fonémy sú zoskupené do $(2k+1)$-gramov $B(j)_k$ a klasifikované pomocou CRF.

Obr. 2.3: Akustický a fonotaktický systém z [6].

2.5 Zhrnutie

Vaščina vyššie uvedených metód detekcie zmeny jazyka v hovore využíva spojenie dvoch dopĺňajúcich sa prístupov: akustického a fonotaktického. Akustické informácie jazyka sú modelované pomocou GMM, ktoré je možné natrénovať použitím ľahko dostupných jednojazyčných dát. Vo fonotaktickej časti je dôležitý fonémový rozpoznávač, ktorý nielen prevádza reč na postupnosť foném, ale aj určuje potenciálne hranice jednojazyčných segmentov. Množina foném môže byť univerzálna pre všetky hovorené jazyky, a rozpoznávač môže byť taktiež natrénovaný na jednojazyčných dátach. Pre každý jazyk je vytvorený N-

[2]Equal Error Rate je hodnota určujúca presnosť systému pri rovnosti miery nesprávnych prijatí (FAR) a miery nesprávnych odmietnutí (FRR).

gramový model na úrovni foném. Použitie jazykových jednotiek vyššej úrovne nie je vhodné kvôli veľmi krátkym jednojazyčným segmentom.

S vyhodnotením výkonnosti týchto systémov je situácia zložitejšia. Každý z nich sa totiž zameriava na špecifické jazyky a spôsoby striedania jazykov. Zatiaľ neexistuje žiadny štandard pre objektívne vyhodnotenie a porovnanie systémov detekcie zmeny jazyka. Viacjazyčných dát je nedostatok a vyhodnocovacie metriky sa pre každú metódu líšia.

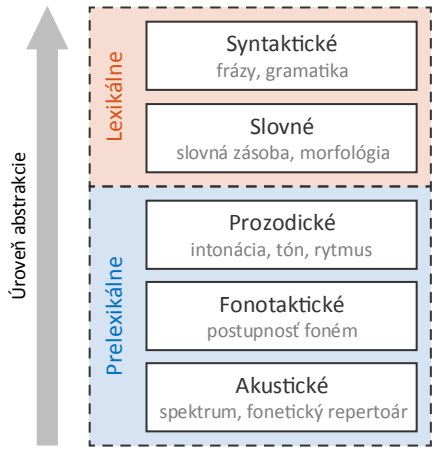
Kapitola 3

Základné princípy identifikácie jazyka

Techniky používané v úlohe detekcie zmeny jazyka sú takmer zhodné s technikami identifikácie jazyka (LID). V tejto kapitole je popísaný spôsob rozpoznávania jazyka človekom a základné metódy automatickej identifikácie jazyka, z ktorých vychádzajú všetky súčasné LID systémy.

3.1 Identifikácia jazyka človekom

Štúdia [10] ukazuje, že novorodenci v bilingválnych domácnostiach sú schopní rozlišovať jazyky bez akýchkoľvek *lexikálnych znalostí*, tj. znalosť významu slov. Rovnako sa aj dospelý človek pri identifikácií dvoch pre neho neznámych jazykov spolieha na tzv. *prelexikálne informácie* (fonetický repertoár jazyka, fonotaktika, rytmus a intonácia). Ak je však aspoň jeden z jazykov človeku známy, lexikálne znalosti začnú hrať rozhodujúcu úlohu v úspešnej identifikácií jazyka. Osvojiť si lexikálne znalosti úplne nového jazyka však vyžaduje od človeka väčšie úsilie, ako pri získavaní prelexikálnych informácií postačujúcich na účel identifikácie jazyka. Na obrázku 3.1 sú zobrazené rôzne úrovne informácií, ktoré môžu byť použité na identifikáciu jazyka.



Obr. 3.1: Zoznam informácií používaných pri identifikácií jazyka zoradený podľa úrovne abstrakcie.

3.2 Automatická identifikácia jazyka

Na základe poznatku, že prelexikálne informácie sú u človeka dostačujúce na úspešnú identifikáciu jazyka sú navrhnuté aj automatické LID systémy. Tie využívajú dva základné prístupy:

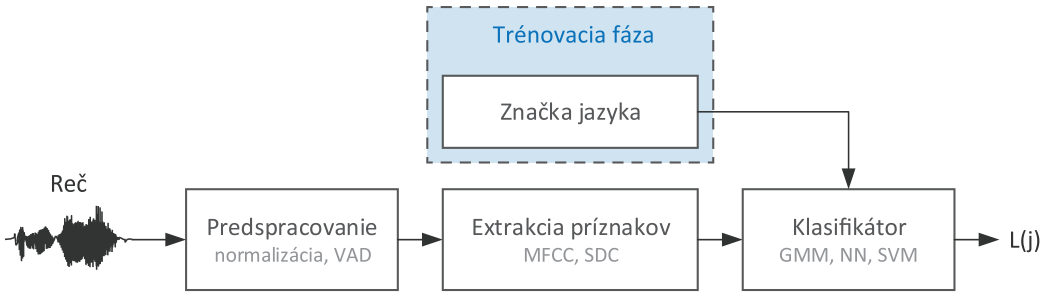
- akustický
- fonotaktický

Tie sú podrobne popísané v sekciách 3.3 a 3.4. V súčasných metódach je veľmi rozšírené spojenie oboch prístupov popísané v sekcií 3.5. Prozodické informácie ako intonácia, tón, trvanie, prízvuk alebo rytmus, označované tiež ako *supra-segmentové* sú podľa [10] málo informatívne a náročné na extrakciu z reči. Ich použitie je veľmi zriedkavé, napr. na určenie, či sa jedná o tónové jazyky, medzi ktoré patria všetky čínske dialekty, japončina ai.

3.3 Akustický prístup

Tento prístup predpokladá, že každý jazyk jazyk je možné rozlíšiť už na základe akustických informácií. Tie je možné modelovať použitím akustických príznakov získaných zo spektrálnej charakteristiky signálu. Akustické príznaky je možné získať z úsekov dlhých už niekoľko desiatok milisekúnd. V tomto časovom intervale považujeme rečový signál za stacionárny, pretože vokálny trakt má istú zotrvačnosť, ktorá mu nedovoľuje sa meniť rýchlejšie. Tento prístup je vhodnejší na identifikáciu krátkych úsekov reči, pretože sme schopní v krátkej dobe získať dostatok potrebných informácií.

Všeobecná schéma akustického systému je zobrazená na obrázku 3.2. Najpoužívanejšia technika na modelovanie akusticko-fonetických vlastností jazyka je *zmes Gaussových rozložení* (GMM). Pre každý jazyk môže byť natrénovaný jeden GMM a nahrávka je priradená k jazyku reprezentovanému GMM s najvyššou vierohodnosťou (likelihood). Klasifikátor môže využívať aj iné techniky, napr. *support vector machines* alebo *neurónové siete*. V súčasnosti sa najrozšírenejšou stala kombinácia GMM, i-vektorov a klasifikátoru.



Obr. 3.2: Všeobecná schéma akustického LID systému.

3.4 Fonotaktický prístup

V tomto prípade sú ako príznaky použité *fonémy*, najmenšie súčasti zvukovej stránky reči, ktoré rozlišujú význam slov. Fonémy sú získané z reči pomocou *fonémového rozpoznávača*, v ktorom sú modelované použitím *skrytého Markovovho modelu* (HMM). Na zachytenie

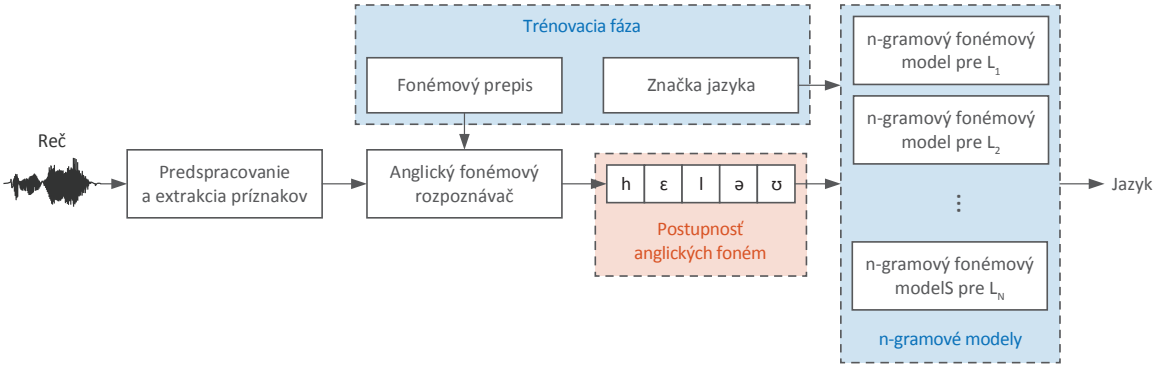
fonologickej informácie je nutné pracovať s dlhšími úsekmi reči ako pri akustickom prístupe. Naviac na trénovanie fonémového rozpoznávača musí byť k dispozícii fonémový prepis reči, ktorého vytvorenie je časovo náročné. Na výkonnosť fonotaktického systému má najväčší vplyv práve použitý fonémový rozpoznávač. Ten by mal byť čo najpresnejší, ale hlavne by mal poskytovať konzistentný výstup.

3.4.1 PRLM

Na obrázku 3.3 je zobrazený systém s fonémovým rozpoznávačom nasledovaným jazykovým modelom (PRLM), ktorý využíva rozpoznávač foném jediného jazyka, v tomto prípade angličtiny. Postupnosti anglických foném – n-gramy, najčastejšie *bi-gramy* (dvojice foném) alebo *tri-gramy* (trojice foném) sú použité na trénovanie štatistických jazykových modelov $\lambda_1, \lambda_2, \ldots, \lambda_N$ pre N cieľových jazykov $L_1, L_2, \ldots, L_N$. Pri testovaní je nahrávka rovnakým rozpoznávačom prevedená na postupnosť foném $\mathcal{Y} = w_1, w_2, \ldots, w_J$ s dĺžkou J. Logaritmická vierohodnosť pre jazyk l je potom daná nasledovne:

$$\log P(\mathcal{Y}|\lambda_l) = \sum_{j=1}^{J} \log P_{\lambda_l}\left(w_j|w_{j-1}\ldots w_{j-(n-1)}\right) \tag{3.1}$$

Testovacia sekvencia je následne priradená jazyku s najvyššou logaritmickou vierohodnosťou.



Obr. 3.3: Schéma PRLM systému, ktorý využíva rozpoznávač anglických foném.

3.4.2 PPR-LM

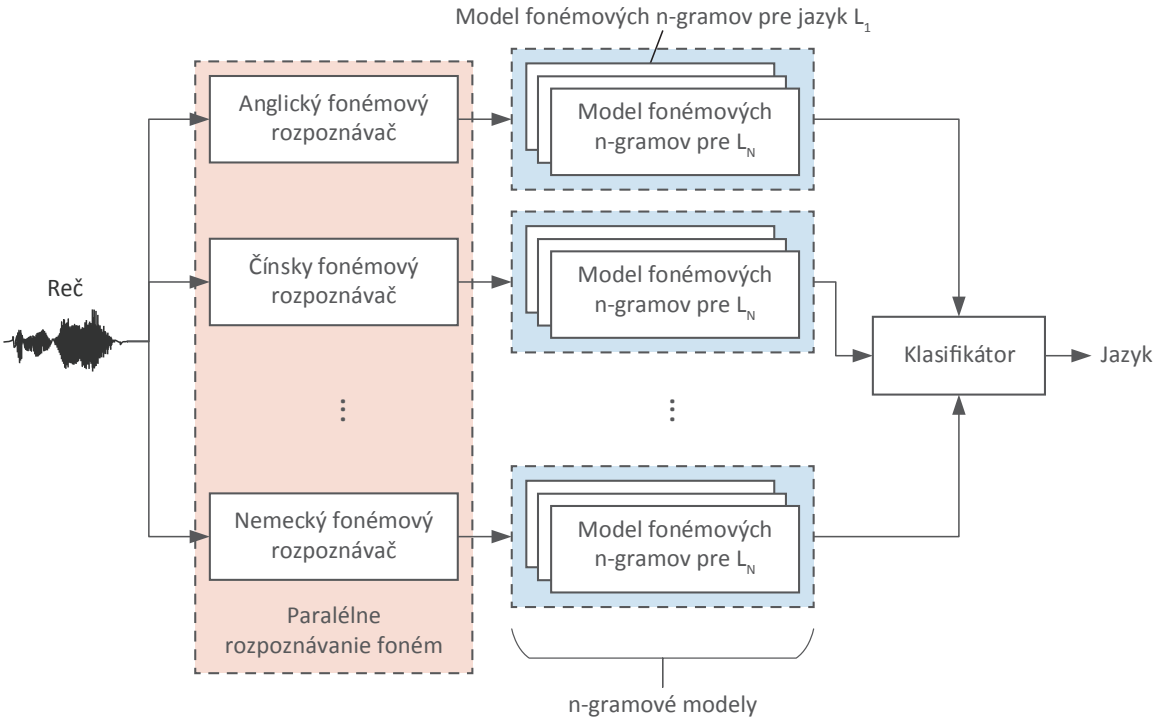
Rozšírením vyššie uvedeného systému je paralelné rozpoznávanie foném nasledované jazykovými modelmi (PPR-LM). V tomto prípade systém využíva niekoľko fonémových rozpoznávačov, každý natrénovaný na inom jazyku. Ako možno vidieť na obrázku 3.4, pre každý z F fonémových rozpoznávačov je natrénovaných N štatistických jazykových modelov pre N cieľových jazykov. Pre testovaciu nahrávku dostávame FN výsledných hodnôt z FN n-gramových modelov $\lambda_{f,l}$ pre $f = 1, 2, \ldots, F$ a $l = 1, 2, \ldots, N$.

Získané hodnoty je potom možné do výsledného rozhodnutia skombinovať rôznymi spôsobmi. Jednou z možností je zlúčiť posteriórne pravdepodobnosti z F paralélnych subsystémov nasledovne:

$$\log P(L_l|\mathcal{O}) = \sum_{f=1}^{F} \log \frac{P(\mathcal{Y}_f|\lambda_{f,l})}{\sum_{i=1}^{N} P(\mathcal{Y}_f|\lambda_{f,i})} \tag{3.2}$$

kde $\mathcal{Y}_f$ je postupnosť foném vygenerovaná z nahrávky $\mathcal{O}$ f-tým fonémovým rozpoznávačom a $P(\mathcal{Y}_f|\lambda_{f,l})$ je skóre pre l-tý jazyk.

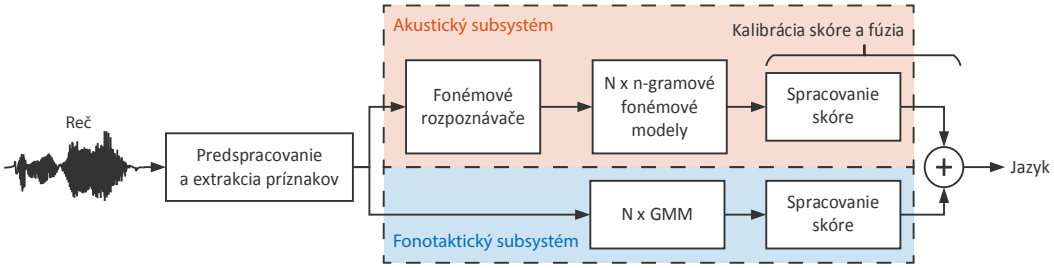
Použitím viacerých fonémových rozpoznávačov nemusí vždy znamenať lepšiu úspešnosť systému. V [13] je dokázané, že v niektorých prípadoch môže jeden dôkladne natrénovaný rozpoznávač prekonať fúziu niekoľkých rozpoznávačov s nedostatkom trénovacích dát. Ak ale tento rozpoznávač zakomponujeme do menej výkonnej fúzie, výsledná úspešnosť je takmer vždy lepšia.



Obr. 3.4: Schéma PPR-LM systému, ktorý využíva viacero fonémových rozpoznávačov natrénovaných na rozličné jazyky. Pre každý rozpoznávač je natrénovaných N n-gramových modelov. V bloku klasifikátor sú vygenerované skóre zlúčené do jedného.

3.5 Kombinácia akustického a fonotaktického prístupu

Pokročilé techniky identifikácie jazyka sa skladajú z rôznych typov akustických a fonotaktických subsystémov. Kalibráciou a vhodnou kombináciou ich výstupov je možné zlepšiť celkovú úspešnosť systému, väčšinou o viac ako 20 % oproti jednotlivému systému. Najjednoduchším spôsobom fúzie je vážený priemer výstupov. Bežne sú používané spôsoby založené na lineárnej logistickej regresií, unimodálnych Gaussovských rozloženiach alebo neurónových sieťach. Na obrázku 3.5 je príklad fúzie dvoch subsystémov.

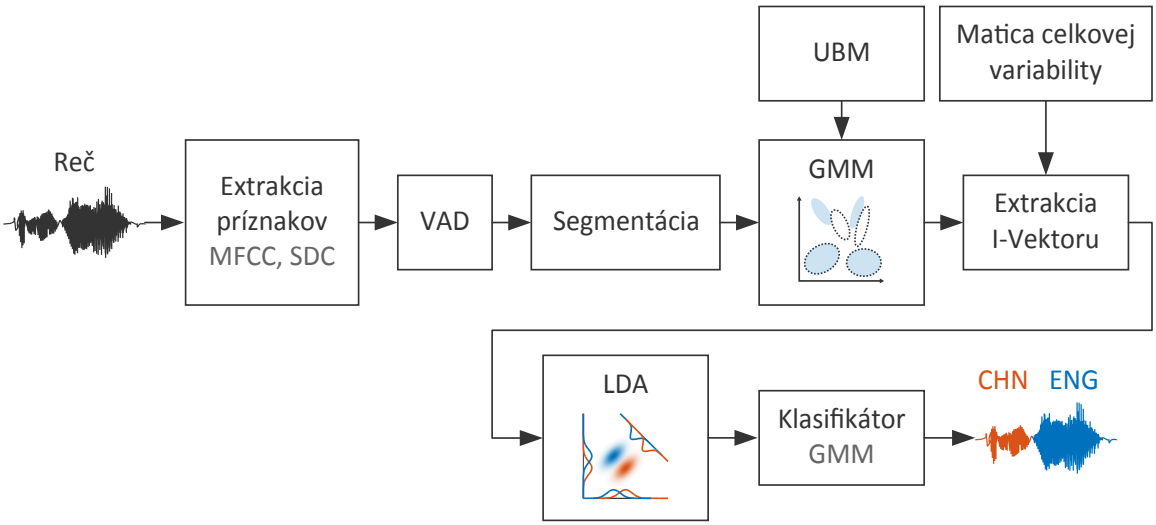


Obr. 3.5: Schéma systému, ktorý kombinuje akustický a fonotaktický podsystém.

Kapitola 4

Akustická metóda diarizácie jazykov

V predchádzajúcej kapitole sú stručne popísané techniky identifikácie jazyka. Táto kapitola podrobne popisuje systém diarizácie jazykov založený na akustickom prístupe identifikácie jazyka, ktorý bol použitý v tejto práci. Na obrázku 4.1 je zobrazená jeho bloková schéma. Jednotlivé bloky budú popísané v nasledujúcich podkapitolách.



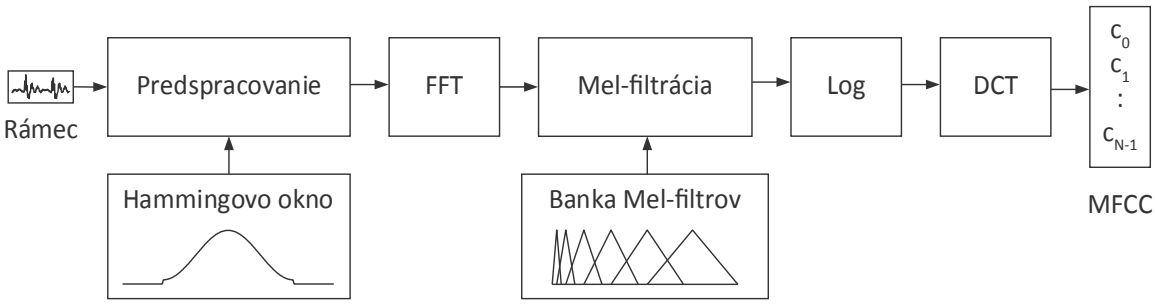
Obr. 4.1: Schéma implementovaného akustického systému diarizácie jazyka.

4.1 Extrakcia príznakov

Zvukové nahrávky sú pred spracovaním rozdelené na krátke úseky konštantnej dĺžky nazývané *rámce*. Bežne používané sú rámce s dĺžkou 20 ms a posunom 10 ms. Pre každý rámec sú extrahované akustické príznaky. Najefektívnejšími a najpoužívanejšími príznakmi v úlohách spojených so spracovaním reči sú *Mel-frekvenčné kepstrálne koeficienty* (MFCC) [16]. Pre účely identifikácie jazyka sú naviac použité *koeficienty SDC* (Shifted Delta Cepstra), ktoré zachytávajú dynamiku reči v rozsahu viacerých rámcov [9]. Koeficienty oboch typov sú zreťazené do výsledného príznakového vektoru. Proces ich extrakcie je detailne popísaný v nasledujúcich sekciách.

4.1.1 MFCC

Výpočet MFCC je znázornený na obrázku 4.2. Signál je najprv filtrovaný Hammingovým oknom[1] a následne je pomocou *krátkodobej rýchlej Fourierovej transformácie* prevedený do frekvenčnej oblasti, *spektra*. Spektrum je prevedené do Mel-ovej stupnice použitím *banky Mel-filtrov*. Filtre aproximujú odozvu ľudského sluchového systému, ktorý citlivejšie rozlišuje nižšie frekvencie. Vypočítané hodnoty sú zlogaritmované a pomocou *diskrétnej kosínovej transformácie* (DCT) prevedené na MFCC.



Obr. 4.2: Postup výpočtu Mel-frekvenčných kepstrálnych koeficientov z rečových rámcov.

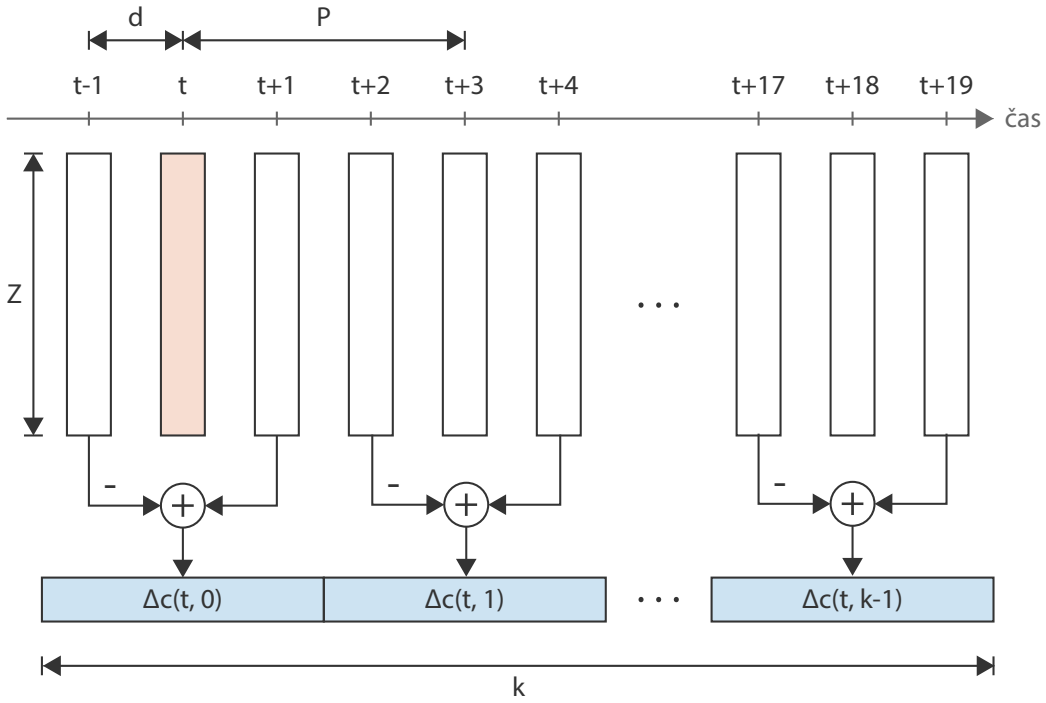
Na získané MFCC je v úlohách identifikácie jazyka aplikovaná *normalizácia dĺžky vokálneho traktu* (Vocal Tract Length Normalization – VTLN) a *normalizácia kepstrálnej strednej hodnoty a odchýlky* (Cepstral Mean and Variance Normalization – CMVN).

4.1.2 Shifted Delta Cepstra príznaky

Z MFCC sú následne vypočítané SDC príznaky, ktoré sú špecifikované štvoricou parametrov $\{Z, d, P, k\}$. Z reprezentuje počet MFCC na rámec, d udáva časový posun vpred a vzad pre výpočet Δc podľa vzťahu 4.1, k je počet delta-kepstrálnych blokov $\Delta c(t, i)$, ktorých zreťazené koeficienty tvoria SDC príznakový vektor a P reprezentuje časový posun medzi po sebe idúcimi blokmi. Postup výpočtu je znázornený na obrázku 4.3.

$$\Delta c(t, i) = c(t + iP + d) - c(t + iP - d) \tag{4.1}$$

[1] http://en.wikipedia.org/wiki/Window_function#Hamming_window



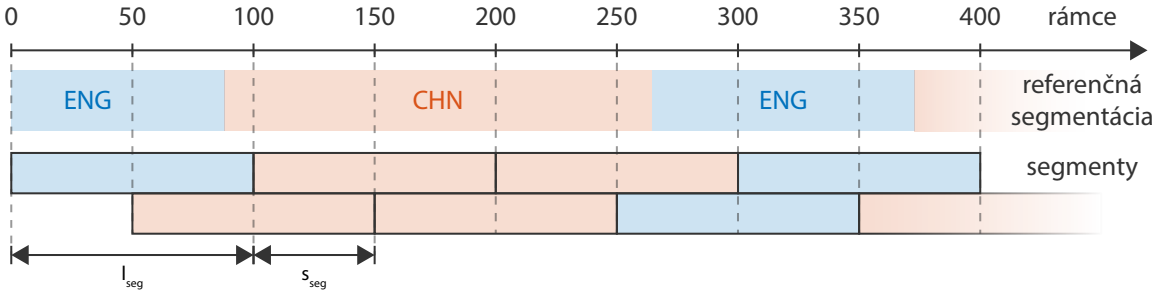
Obr. 4.3: Výpočet SDC príznakov v klasickej konfigurácii $Z-d-P-k = 7-1-3-7$ v čase t. Farebne vyznačené bloky sú nakoniec zreťazené do výsledného príznakového vektoru.

4.2 Detekcia hlasovej aktivity

Získané príznaky sú spracované *detektorom hlasovej aktivity* (VAD). Tento krok odstráni z nahrávky ticho, prípadne nežiadúce zvuky a ponechá len reč. Detektor je založený na rozpoznávači maďarských foném [13]. Rozpoznané fonémy sú následne priradené do jedinej triedy – reč. Iba rámce spadajúce do tejto triedy budú ďalej použité.

4.3 Segmentácia

Postupnosť príznakových vektorov je v tomto kroku segmentovaná na krátke úseky pevnej dĺžky, ktoré budeme považovať za jednojazyčné. Spôsob segmentácie je zobrazený na obrázku 4.4 a vychádza zo segmentácie nahrávky na rámce. V tomto prípade je dĺžka segmentov l_{seg} väščia pre uchovanie akustických informácií potrebných na identifikáciu jazyka. Úseky však nemôžu byť ani príliš dlhé, pretože by bola stratená potrebná presnosť. Referenčný jazyk segmentu je daný jazykom väčšiny jednotlivých rámcov.



Obr. 4.4: Príklad segmentácie nahrávky na úseky dĺžky $l_{seg} = 100$ s posunom $s_{seg} = 50$. Pre rámce s dĺžkou 20 ms a posunom 10 ms je dĺžka segmentu 1 s.

4.4 Zmes Gaussových rozložení

V úlohách spojených so spracovaním reči je rozloženie akustických príznakových vektorov väčšinou modelované ako zmes Gaussových rozložení (Gaussian Mixture Model – GMM). Tento generatívny model sa skladá z kombinácie niekoľkých Gaussových komponentov. Funkcia hustoty pravdepodobnosti komponentu $\mathcal{N}(x|\mu, \Sigma)$ pre D-rozmernú náhodnú veličinu x je definovaná ako

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \tag{4.2}$$

kde μ je stredná hodnota a Σ je kovariačná matica. Pre príznakový vektor x modelovaný pomocou GMM s M komponentami je vypočítaná vierohodnosť príslušnosti do triedy c nasledovne:

$$P_\lambda(x|c) = \sum_{m=1}^{M} w_{cm}\mathcal{N}(x|\mu_{cm}, \Sigma_{cm}) \tag{4.3}$$

kde w_{cm} sú váhy komponentov, pre ktoré platí $\sum_{m=1}^{M} w_{cm} = 1$. Celý model je možné charakterizovať množinou parametrov $\lambda = \{w_m, \mu_m, \Sigma_m\}$, $m = 1, \ldots, M$. V praxi sa namiesto plných kovariačných matíc Σ_m používajú diagonálne σ_m^2. Podľa [11] výkonnostne prekonávajú plné kovariačné matice a sú naviac menej výpočetne náročné.

Ak uvažujeme postupnosť príznakových vektorov $X = x_1, \ldots, x_T$, potom logaritmická vierohodnosť vzhľadom k triede c je

$$\log P_\lambda(X|c) = \sum_{t=1}^{T} \log P_\lambda(x_t|c) \tag{4.4}$$

4.4.1 EM algoritmus

Parametre λ rozloženia $P(X|\lambda)$ je možné pri dostatku trénovacích dát určiť metódou maximálnej vierohodnosti (Maximum Likelihood – ML), kde sú neznáme odhadované parametre považované za pevné:

$$\tilde{\lambda}_{ML} = \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda) \tag{4.5}$$

V prípade unimodálneho Gaussového rozloženia[2] je možné ML parametre určiť priamo analyticky. Pri viacmodálnom modeli, kde nie je známa príslušnosť trénovacích dát k jednotlivým komponentom, je nutné ML parametre odhadnúť. Najčastejšie využívaným je *EM algoritmus*, ktorý iteratívne odhaduje parametre modelu v závislosti na skrytých premenných. V prípade GMM to sú informácie o príslušnosti trénovacích dát k jednotlivým komponentom. Po inicializácií parametrov modelu λ sa v každej iterácií opakujú dva kroky:

1. **E krok** (expectation) vypočíta pre každý trénovací vektor x_t a komponent i tzv. *okupačnú pravdepodobnosť* $\gamma_{t,i}$. Táto hodnota vyjadruje mieru príslušnosti príznakového vektoru x_t k i-tému komponentu stanovenú na základe aktuálnych parametrov modelu $\tilde{\lambda}$:

$$\gamma_{t,i} = \frac{w_i \mathcal{N}(x_t|\mu_i, \Sigma_i)}{\sum_{m=1}^{M} w_m \mathcal{N}(x_t|\mu_m, \Sigma_m)} \tag{4.6}$$

   Počas výpočtu okupačných pravdepodobností je výhodné akumulovať premenné N_i, F_i a S_i nazývané *postačujúce štatistiky*. Tie budú v nasledujúcom kroku použité na výpočet nových hodnôt parametrov modelu.

$$N_i = \sum_{t=1}^{T} \gamma_{t,i} \tag{4.7}$$

$$F_i = \sum_{t=1}^{T} \gamma_{t,i} x_t \tag{4.8}$$

$$S_i = \sum_{t=1}^{T} \gamma_{t,i} x_t x_t^T \tag{4.9}$$

2. **M krok** (maximization) hľadá nové parametre λ^{ML} modelu metódou maximálnej vierohodnosti. Pre i-ty komponent sú vypočítané nasledovne:

$$w_i^{ML} = \frac{N_i}{T} \tag{4.10}$$

$$\mu_i^{ML} = \frac{F_i}{N_i} \tag{4.11}$$

$$\Sigma_i^{ML} = \frac{S_i}{N_i} - \mu_i^{ML}(\mu_i^{ML})^T \tag{4.12}$$

Algoritmus vždy konverguje a zvyšuje vierohodnosť modelu pre trénovacie dáta, ale nie je zaručené, že nájde globálne maximum. Uviaznutie v lokálnom maxime môže nastať pri nevhodnej náhodnej inicializácií stredných hodnôt komponentov. Riešením je použitie algoritmu k-means, kde sú počiatočné stredné hodnoty dané stredmi zhlukov a kovariačné matice sú vypočítané na základe dát priradených k jednotlivým zhlukom.

[2] tvorené jediným Gaussovým komponentom

4.4.2 Univerzálny hlasový model

Vyššie popísaná metóda ML odhadu sa v úlohách identifikácie jazyka a rečníka používa predovšetkým na natrénovanie *univerzálneho hlasového modelu* (Universal Background Model – UBM). Tento model reprezentuje rozloženie príznakových vektorov nezávisle na jazyku a rečníkovi. Na trénovanie je použité čo najväčšie množstvo dát v rozličných jazykoch, aby bolo dostatočne pokryté široké spektrum akustických informácií. Počet komponentov UBM je typicky vysoký (v súčasných systémoch 1024 alebo 2048 komponentov).

4.4.3 Adaptácia UBM

V GMM-UBM systéme je jazykový model získaný adaptáciou parametrov UBM metódou maximálnej apostériornej pravdepodobnosti (MAP). V tomto prípade sú parametre λ považované za náhodné veličiny, pre ktoré je známe ich apriórne rozloženie $P(\lambda)$.

$$\tilde{\lambda}_{MAP} = \underset{\lambda}{\operatorname{argmax}}\, P(\lambda|X) \tag{4.13}$$

$$= \underset{\lambda}{\operatorname{argmax}}\, P(X|\lambda)P(\lambda) \tag{4.14}$$

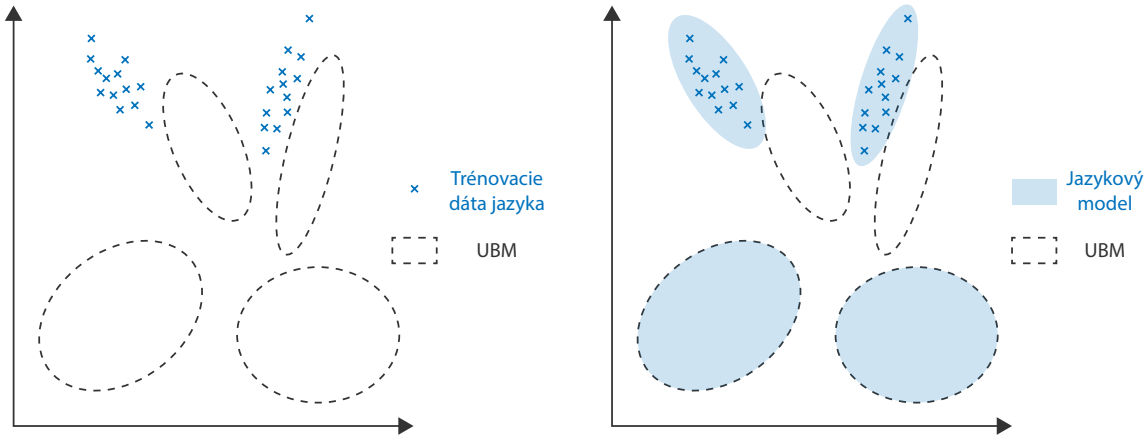
Rovnako ako pri ML odhade, ani pri MAP odhade nie je známa príslušnosť trénovacích dát jazyka k jednotlivým komponentám. Preto je opäť použitý iteratívny výpočet pozostávajúci z dvoch krokov. Prvý krok ilustrovaný na obrázku 4.5a je identický s E krokom z EM algoritmu, kde sú vypočítané hodnoty okupačných pravdepodobností a akumulované postačujúce štatistiky trénovacích dát.

M krok je v prípade MAP odhadu odlišný od vyššie popísaného M kroku. Postačujúce štatistiky jazykového modelu sú použité na adaptáciu parametrov UBM, čo je možné vidieť na obrázku 4.5b. V tejto práci sú adaptované iba stredné hodnoty μ_i nasledovne:

$$\mu_i^{MAP} = \frac{F_i}{N_i} - \mu_i \tag{4.15}$$

Váhy jazykového modelu w_i sú získané ML odhadom z trénovacích dát daného jazyka a kovariačné matice Σ_i zostávajú nezmenené.

Použitie jazykových modelov odvodených z UBM má oproti klasickému modelovaniu nezávislom na UBM niekoľko výhod. Dáta nevidené počas trénovania dosahujú pri klasifikácii približne nulové skóre pre natrénované jazykové modely. Skóre pre UBM je naopak vysoké a dáta teda nebudú priradené do žiadnej z pozorovaných tried. Ďalšou výhodou je zníženie nároku na uloženie modelu jazykov, pretože uložené sú iba rozdiely parametrov adaptovaných komponentov od UBM (viac v [14]).



(a) Priradenie trénovacích dát ku komponentom UBM a akumulácia postačujúcich štatistík.

(b) MAP odhad nových parametrov jazykového modelu na základe postačujúcich štatistík a pôvodných parametrov UBM.

Obr. 4.5: Príklad adaptácie prametrov UBM na jazykový model v dvoch krokoch.

4.5 I-vektor

Zreťazením vektorov stredných hodnôt komponentov GMM je vytvorený tzv. *supervektor*. V prípade D-rozmerného príznakového vektoru a GMM s M komponentami má GMM supervektor rozmer MD. Pre redukciu tohto rozmeru je použitý jednoduchý model faktorovej analýzy nazývaný *i-vektor*. I-vektor je nízkorozmerný vektor pevnej dĺžky, ktorý je extrahovaný z nahrávky s ľubovolnou dĺžkou. Okrem redukcie dimenzioinality i-vektor berie do úvahy aj variablility rečníka a kanálu (parametre jazyka sa môžu pri rôznych rečníkoch alebo nahrávkach líšiť), ktoré spája do tzv. celkovej variability (angl. total variability). Pre supervektor $m_{r,l}$ odpovedajúci nahrávke r a jazyku l je predpokladaný model

$$m_{r,l} = \mu + T x_{r,l} \tag{4.16}$$

kde μ je supervektor nezávislý na jazyku a nahrávke (zreťazené stredné hodnoty UBM-GMM), T je matica rozmeru $MD \times D_{ivec}$ definujúca *priestor celkovej variability* (Total Variability Space) a $x_{r,l}$ je D_{ivec}-rozmerný vektor s rozložením $\mathcal{N}(0, I)$ označovaný ako i-vektor. Rozmer i-vektoru je mnohonásobne menší (v rádoch 100) než rozmer supervektoru (rádovo 10000).

4.6 Lineárna diskriminačná analýza

Cieľom lineárnej diskriminačnej analýzy (LDA) je násť taký podpriestor, v ktorom sú triedy (jazyky) najlepšie rozlíšiteľné. V tomto podpriestore sú dáta dekorelované s malým rozpltylom v rámci triedy a s veľkým rozptylom medzi triedami, čo je možné vidieť na obrázku 4.6. Projekcia n-rozmerných príznakových vektorov do priestoru s nižším rozmerom m má tvar $x' = A^T x$. Transformačná matica A má rozmery $n \times m$ a je daná vlastnými vektormi matice $\Sigma_W^{-1}\Sigma_B$. Pre C tried (jazykov) a N dostupných pozorovaní (i-vektorov) sú jej jednotlivé zložky dané nasledovne:

$$\Sigma_W = \frac{1}{N}\sum_{c=1}^{C} N_c \Sigma_c \tag{4.17}$$

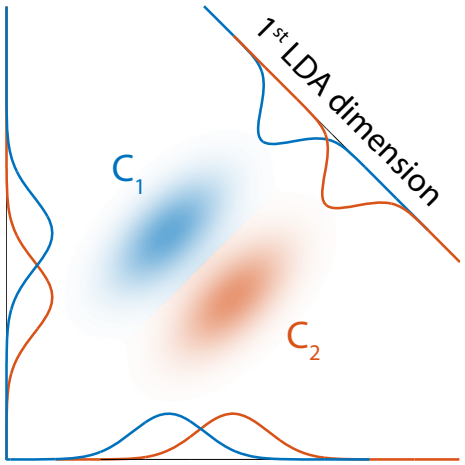
$$\Sigma_B = \frac{1}{N}\sum_{c=1}^{C} N_c (\mu_c - \mu)(\mu_c - \mu)^T \tag{4.18}$$

kde Σ_W je priemerná kovariácia vnútri tried a Σ_B označuje priemernú kovariáciu medzi triedami. V prípade i-vektorov s rozložením $\mathcal{N}(0, I)$ je $\mu = 0$. Pre triedu c reprezentuje N_c počet pozorovaní, μ_c strednú hodnotu a Σ_c kovariačnú maticu:

$$\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} x_{i,c} \tag{4.19}$$

$$\Sigma_c = \frac{1}{N_c}\sum_{i=1}^{N_c} (x_{i,c} - \mu_c)(x_{i,c} - \mu_c)^T \tag{4.20}$$

V prípade použitia LDA na transformáciu dát, ktoré patria do M tried môže mať hľadaný podpriestor rozmer maximálne $M - 1$, pretože transformačná matica má najviac $(M - 1)$ vlastných čísel nenulových.



Obr. 4.6: Príklad transformácie dvojrozmerných dát jednorozmerného podpriestoru pomocou LDA. V tomto podpriestore sú triedy C_1 a C_2 najlepšie separabilné.

4.7 Klasifikátor

Klasifikátor priradí vstupným dátam $x \in X$ značku príslušnej triedy $c \in C$. Existujú dve hlavné triedy klasifikátorov:

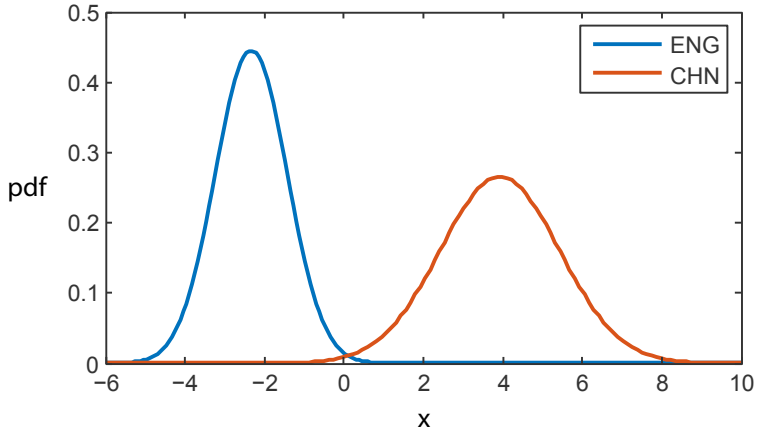
- *generatívne* – implicitne alebo explicitne modelujú rozloženie vstupných a výstupných dát a je možné nimi vygenerovať umelé vstupné dáta
- *diskriminatívne* – založené na učení diskriminačej funkcie, ktorá vstupné dáta priamo zobrazuje na označenie tried

V tejto práci je použitý generatívny klasifikátor, ktorý modeluje každú triedu unimodálnym Gaussovým rozložením. Funkcia hustoty podmienenej pravdepodobnosti $P(x|c)$ pre každú triedu $c \in C$ je reprezentovaná funkciou hustoty pravdepodobnosti normálneho rozloženia daná vzťahom 4.2. Posteriórna pravdepodobnosť $P(c_i|x)$ pre triedu c_i je určená použitím Bayesovej vety:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)} \tag{4.21}$$

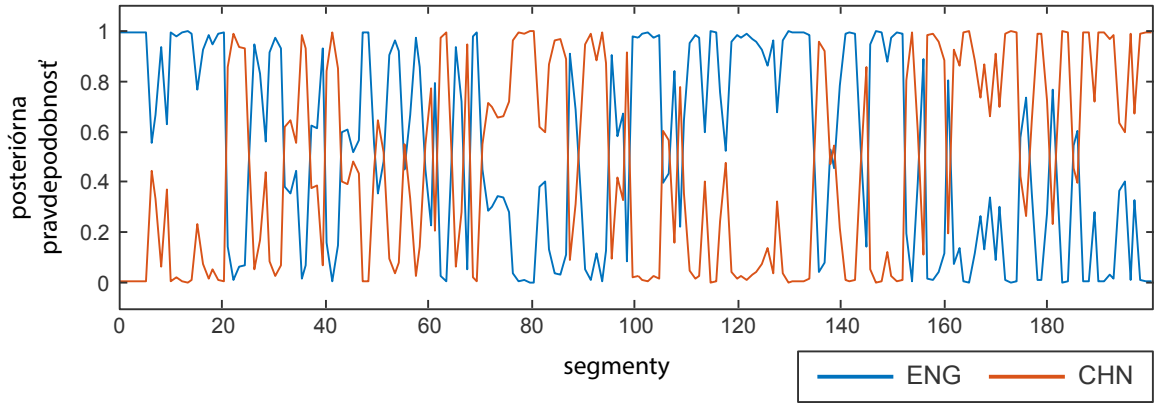
$$P(x) = \sum_{j=1}^{C} P(x|c_j)P(c_j) \tag{4.22}$$

kde $P(c_i)$ je apriórna pravdepodobnosť triedy c_i a $P(x)$ je normalizačný člen. Na základe posteriórnych pravdepodobností je rozhodnuté o príslušnosti vstupných dát k jednej z tried. Na obrázku 4.7 je zobrazený príklad binárneho klasifikátoru, ktorý modeluje triedy (jazyky) jednorozmernými normálnymi rozloženiami.



Obr. 4.7: Funkcie hustoty pravdepodobnosti reprezentujúce dve triedy – angličtinu a mandarínčinu.

Každý i-vektor reprezentujúci krátky segment reči je klasifikovaný samostatne. Obrázok 4.8 znázorňuje posterórne pravdepodobnosti segmentov v celej nahrávke.

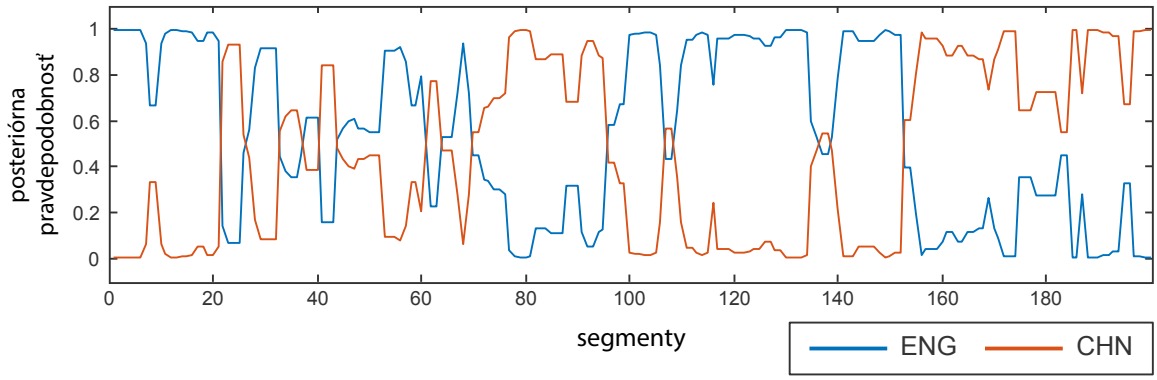


Obr. 4.8: Posteriórne pravdepodobnosti pre jednotlivé segmenty v nahrávke reči so striedaním angličtiny a mandarínčiny.

4.8 Dodatočné spracovanie mediánovým filtrom

Segmentácia nahrávky je dodatočne vyhladená aplikáciou jednorozmerného mediánového filtru na posteriórne pravdepodobnosti jednotlivých segmentov. Hodnota vzorku x_i je v prípade použitia mediánového filtru k-teho rádu daná mediánom vzoriek $\{x_j \mid i - (k-1)/2 \leq j \leq i + (k-1)/2\}$.

Filtráciou docielime odfiltrovanie krátkych úsekov reči, kde nastane zmena jazyka, ktoré môžu byť nesprávne klasifikované a tým narušiť spojitý úsek jednojazyčnej reči. Rád filtra je však nutné voliť čo najkratší. V opačnom prípade strácame presnosť segmentácie, ktorá je v úlohe detekcie zmeny jazyka veľmi dôležitá. Na obrázku 4.9 je možné vidieť segmentáciu vyššie zobrazenej nahrávky po aplikovaní mediánového filtra.



Obr. 4.9: Posteriórne pravdepodobnosti pre jednotlivé segmenty v nahrávke reči so striedaním angličtiny a mandarínčiny po aplikácií mediánového filtra 5. rádu.

Kapitola 5

Experimenty a výsledky

Táto kapitola obsahuje popis experimentov s akustickým systémom na detekciu zmeny jazyka. V podkapitole 5.2 sú detailne popísané použité dáta, nasledované popisom nastavenia systému v sekcií 5.3. Samotné experimenty s výsledkami sú popísané v podkapitole 5.4. Nakoniec sú výsledky experimentov stručne zhrnuté v sekcií 5.5.

5.1 Implementácia

Systém bol z implementovaný v programovacích jazykoch Python a MATLAB. Jazyk Python je v distribúcií Enthought Canopy[1] založenej na verzií Python 2.7.x. V nej sú obsiahnuté všetky dôležité moduly pre matematické výpočty, prácu s maticami a vykreslovanie grafov (NumPy, SciPy, Matplotlib,...). Experimentovanie so systémom prebiehalo v prostredí Matlab 2013a, ktorý je vhodný na rýchle prototypovanie. Naviac máme k dispozícií funkcie pre pohodlnú prácu s maticami, vizualizáciu výsledkov alebo meranie rýchlosti algoritmov.

5.2 Databázy

Jednou z nepríjemností pri úlohe rozpoznania jazyka v reči so striedaním kódov je nedostatok dát a ich obtiažne získavanie. Výskumu v tejto oblasti bola donedávna venovaná veľmi malá pozornosť, preto problém detekcie zmeny jazyka v reči zatiaľ chýba v NIST evaluáciách. V čase začiatku písania tejto práce nebola verejne dostupná žiadna databáza reči so striedaním jazykov. Dáta bolo teda nutné zozbierať a anotovať ručne. Ďalšou možnosťou je umelo spojiť existujúce jednojazyčné nahrávky. Tie by bolo možné použiť na trénovanie a otestovanie systém, ale pri vyhodnotení autentických viacjazyčných dát by systém pravdepodobne zlyhal.

V experimentoch s implementovaným systémom boli použité dve databázy: jedna s jednojazyčnými nahrávkami a druhá so striedaním jazykov v rámci nahrávky. Obe databázy budú popísané v nasledujúcich sekciách. Pôvodný zámer bol systém vyhodnotiť na anglicko-mandarínskom korpuse so striedaním jazykov SEAME[2]. Ten bol ale uvoľnený veľmi krátky čas pred termínom odovzdania práce (Apríl 2015) a naviac neobsahoval časové značky jednotlivých jednojazyčných úsekov. Použitie tohto korpusu je však možné v rámci budúcej práce pop9sanej v sekcií 6.0.1.

[1] https://www.enthought.com/products/canopy/

[2] South-East Asia Mandarin-English: https://catalog.ldc.upenn.edu/LDC2015S04

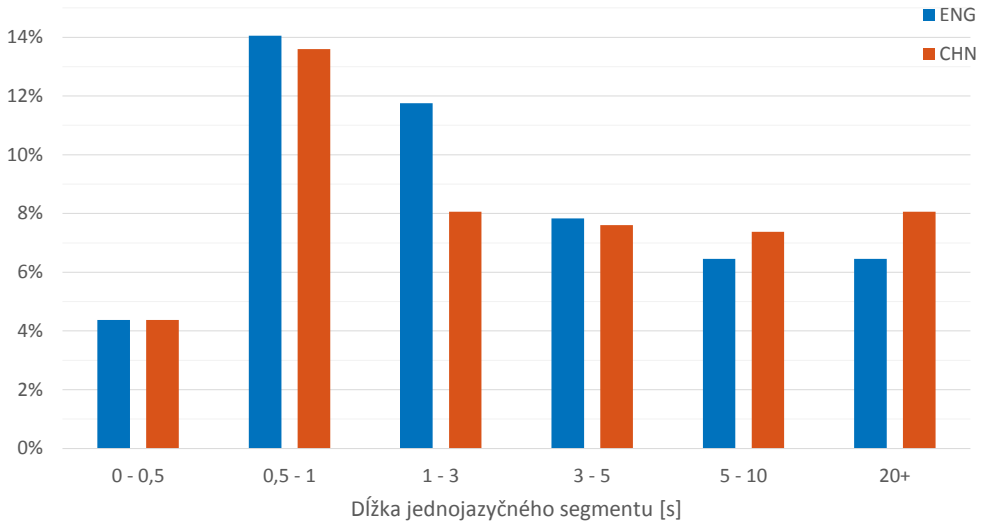
5.2.1 Databáza so striedaním angličtiny a mandarínčiny

Pre účely trénovania, validácie a testovania systému bola vytvorená anglicko-mandarínska databáza nahrávok so striedaním kódov (ďalej texte označená ako **DB-ME**). Nahrávky boli získané z podcastov taiwanského rádia ICRT[3], kde sa vo výukových reláciách strieda angličtina a mandarínska čínština. Niekoľko nahrávok je vo forme dialógu dvoch osôb, kde každá strieda jazyk. Väčšina je však vo forme monológu ženy, v ktorom po mandarínsky vysvetľuje význam anglických fráz. V niektorých nahrávkach sa vyskytuje hudba, ktorá je označená ako ticho a je pri vyhodnotení ignorovaná.

Na nahrávkach bola ručne vykonaná referenčná anotácia na úrovni jazyka. Celková dĺžka reči je približne 40 minút, oba jazyky sú zastúpené takmer v rovnakom pomere a priemerná dĺžka jednojazyčného segmentu sú 3 sekundy. Detailné štatistiky sú uvedené v tabuľke 5.1 a na obrázku 5.1 možno vidieť rozloženie dĺžky jednojazyčných segmentov v tejto databáze.

	Angličtina	Mandarínčina	Spolu
Dĺžka	19,88 min	20,44 min	40,32 min
Pomer	49,3 %	50,7 %	100 %
Priem. dĺžka segmentu	2,74 s	3,27 s	3 s
Počet segmentov	436	375	811

Tabuľka 5.1: Základné štatistiky databázy so striedaním čínštiny a angličtiny.



Obr. 5.1: Rozloženie priemernej dĺžky jednojazyčných segmentov v anglicko-mandarínskej databáze. Väčšina segmentov v oboch jazykoch má dĺžku 0–0,5 s.

5.2.2 Databáza s jednojazyčnými dátami

LDA a GMM klasifikátor sú natrénované na podmnožine dát, ktoré boli použité na trénovanie a vyhodnotenie systému identifikácie jazyka v [8]. Podmnožina obsahuje jednojazyčné

[3]International Community Radio Taipei: http://www.icrt.com.tw/newsroom_podcasts.php?pageOneId=2&pageTwoId=11&poId=3&ptId=46

nahrávky v angličtine, kantónčine, farsi, francúzštine, hindčine, kórejčine, mandarínčine, portugalčine, ruštine, španielčine a vietnamčine z databáz Callfriend, Fisher English časť 1 a 2, HKUST Mandarin, Mixer (dáta z NIST SRE 2004, 2005, 2006, 2008), dáta z NIST LRE, OGI-multilingual, OGI 22 languages, Foreign Accented English a dáta z rádiového vysielania Voice of America. V ďalšom texte bude táto databáza označená ako **DB-Mix**.

Jazyk	Počet nahrávok						
	Spolu	CallFriend	SRE	OGI-mltlng	OGI 22	VOA	Iné
angličtina	15197	240	1940	1069	297	3963	7688
mandarínčina	2370	240	249	266	290	1049	276
kantónčina	462		179		283		
farsi	656	120		255	281		
francúzština	403	120			283		
hindčina	755	120	158	198	279		
kórejčina	691	120	63	216	292		
portugalčina	294				294		
ruština	3714		334		289	3071	20
španielčina	2624	238	181	305	277	1623	
vietnamčina	856	120	147	237	239		

Tabuľka 5.2: Počet trénovacích nahrávok v pre jednotlivé jazyky a databázy.

5.3 Nastavenie systému

Vo všetkých experimentoch sú predpokladané len dva neprekrývajúce sa jazyky: angličtina a mandarínčina. Ich apriórna pravdepodobnosť je rovnaká. Ak sa v nahrávke vyskytujú iné jazyky, tak sú chybne klasifikované. Systém je však možné ľahko rozšíriť o modely iných jazykov, a využiť ho k diarizácií niekoľkých jazykov (viac v sekcií 6.0.1).

5.3.1 Extrakcia príznakov

Pre každý rámec dĺžky 20 ms s posunom 10 ms je vypočítaných 7 MFCC (vrátane c_0) + 49 SDC koeficientov v štandardnej konfigurácií $Z - d - P - k = 7 - 1 - 3 - 7$. Koeficienty sú zreťazené do 56-rozmerného príznakového vektoru. Pre každý úsek dĺžky 1 s (100 rámcov) s posunom 0,5 s sú vypočítané postačujúce štatistiky z univerzálneho hlasového modelu (UBM). Ten je tvorený 512 Gaussovskými komponentami a je natrénovaný na dátach z 54 jazykov použitých v [8]. Z UBM parametrov sú použité iba stredné hodnoty a váhy komponentov. Kovariačné matice zostávajú nezmenené. Z týchto štatistík je extrahovaný 400-rozmerný i-vektor. ktorý je priamo použitý na klasifikáciu. V prípade dát z DB-Mix je postup extrakcie príznakov zhodný, až na výpočet postačujúcich štatistík, ktoré sú akumulované z celej jednojazyčnej nahrávky.

5.3.2 Klasifikácia

V prvej fáze návrhu systému bola klasifikácia riešená použitím zhlukovacieho algoritmu *k-means*. I-vektory zo všetkých segmentov boli priamo použité ako body pre k-means a priradené do zhlukov. Výsledné zhluky boli následne namapované na triedy reprezentujúce príslušné jazyky. Nevýhodou tohto prístupu je snaha algoritmu rozdeliť dáta do zhlukov,

ktorých počet je dopredu daný. Problém vzniká pri klasifikácií jednojazyčnej reči, ktorá by mala byť priradená do jediného zhluku.

K-means klasifikátor bol neskôr nahradený tívnym klasifikátorom, ktorý modeluje každú triedu unimodálnym Gaussovým rozložením, popísaným v sekcií 4.7.

5.3.3 Vyhodnocovacia metrika

Úspešnosť systému je vyhodnotená metrikou SER (Segment Error Rate), ktorá je daná mierou chybne klasifikovaných jednojazyčných segmentov x_i v nahrávke $X = \{x_1, x_2, \ldots, x_N\}$:

$$SER(X) = \frac{1}{N} \sum_{c=1}^{C} F_c \tag{5.1}$$

kde F_c je počet nesprávne klasifikovaných segmetov do triedy c a N je celkový počet segmentov v nahrávke. Klasifikácia úsekov bez reči sa neberie do úvahy, pretože dáta boli referenčne anotované až po vykonaní automatickej detekcie hlasovej aktivity. Pri segmentácií nahrávky na úseky s dĺžkou 1 s a posunom 0,5 s je systém schopný určiť hranice jednojazyčného úseku s odchýlkou od skutočnej hranice maximálne 0,25 s.

5.4 Výsledky

Systém bol vyhodnotený na databáze so striedaním kódov popísanej v sekcií 5.2.1. Kvôli malému objemu dát v tejto databáze bola na testovanie použitá validačná technika *„vynechanie jedného“* (leave-one-out). Testovaciu množinu vždy predstavuje jediná vzorka, v našom prípade jedna nahrávka. Ostatné nahrávky sú použité na trénovanie a následne prebehne vyhodnotenie. Tento postup je opakovaný pre každú nahrávku a v závere je celková úspešnosť systému daná priemerom úspešnosti jednotlivých opakovaní. Tým je dodržaná podmienka, že testovacie dáta nie sú nikdy použité vo fáze tréningu.

V nasledujúcich sekciách sú popísané výsledky jednotlivých experimentov, ktorých cieľom je nájsť optimálne parametre systému diarizácie jazyka.

5.4.1 Porovnanie klasifikátorov

V tomto experimente sú porovnávané klasifikátory natrénované na rôznych dátach s cieľom vybrať ten, ktorý dosahuje priemerne najnižšiu chybu klasifikácie segmentov vo všetkých testovacích nahrávkach. Natrénované boli celkom tri klasifikátory:

- **Gauss-ME** je natrénovaný na dátach z DB-ME popísanej v sekcií 5.2.1
- **Gauss-Mix** využíva podmnožinu DB-Mix z kapitoly 5.2.2 obsahujúcu angličtinu a mandarínčinu
- **Gauss-All** je natrénovaný na oboch vyššie uvedených databázach

Do porovnania bol zahrnutý aj pôvodný klasifikátor založený na k-means, ktorý nebol v konečnom systéme použitý. Klasifikované boli priamo i-vektory reprezentujúce jednojazyčné segmenty. Z tabuľky 5.3 je zrejmé, že najúspešnejším klasifikátorom je Gauss-ME s mierou chybne klasifikovaného segmentu 18,4 %. Tento výsledok sa dal očakávať, pretože trénovacie a testovacie dáta pochádzajú z rovnakej databázy. Gauss-Mix je síce natrénovaný na väčšom množstve dát, ale tieto dáta sú od testovacích značne odlišné (rôzne dialekty,

iné kanály,...). V nasledujúcich experimentoch však bude ukázané, že ak sa rozhodneme použiť techniku LDA, výber klasifikátoru vplýva na výkonnosť systému len minimálne.

SER [%]	Klasifikátor			
	K-means	Gauss-Mix	**Gauss-ME**	Gauss-All
min	21,6	11,7	12,6	20,5
max	37,1	34,1	30,4	36,3
avg	28,84	27,7	**18,4**	25,5

Tabuľka 5.3: Porovnanie SER použitím rôznych klasifikátorov. Riadkok min udáva najmenšiu a riadok max zas najväčšiu chybu dosiahnutú v rámci jednej nahrávky. V riadku avg je priemerná chyba pre všetky nahrávky.

5.4.2 Experiment s rôznymi LDA transformáciami

Na 400-rozmerné i-vektory je aplikovaná technika redukcie dimenzionality LDA. Keďže pracujeme s $C = 2$ triedami, i-vektory sú pomocou LDA transformované vždy do $(C - 1)$-rozmerného podpriestoru, kde sú triedy najlepšie rozlíšiteľné. V rámci tohto experimentu sú na výpočet transformačnej matice použité dáta rozšírené o jazyky z databázy DB-Mix, ktoré sa nenachádzajú v testovacích dátach. Použitie týchto jazykov nám ukáže, či je model možné dobre generalizovať. Celkovo boli testované štyri varianty LDA transformácie:

- **LDA-Mix:** natrénovaná na angličtine a mandarínčine z DB-Mix
- **LDA-ME:** vypočítaná z dát DB-ME
- **LDA-All:** na výpočet boli použité všetky dáta z DB-Mix, celkom 11 jazykov uvedených v tabuľke 5.2. Dáta majú po transformácií 10 rozmerov.
- **LDA-Other:** natrénovaná na všetkých dátach z DB-Mix okrem cieľových jazykov, čo je celkom 9 jazykov a teda rozmer podpriestoru je 8. Cieľom tejto varianty je overiť robustnosť systému.

SER [%]	Klasifikátor		
Typ LDA	Gauss-Mix	**Gauss-ME**	Gauss-All
LDA-Mix	25,6	25,6	25,5
LDA-ME	16.3	**15.6**	15.9
LDA-All	25.5	24.3	25.5
LDA-Other	34.5	31.1	32.7

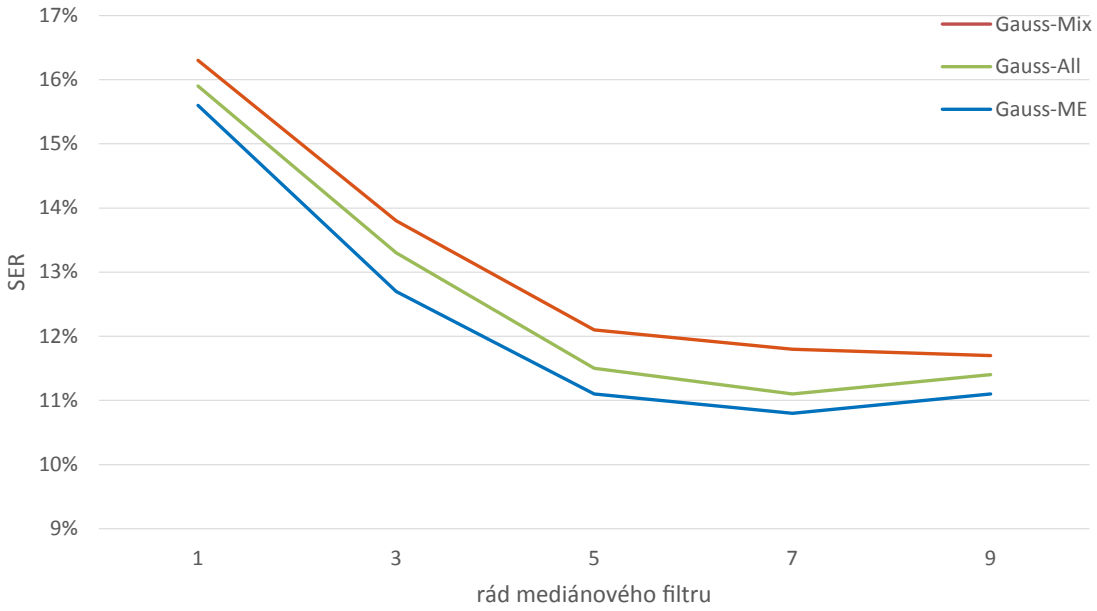
Tabuľka 5.4: Porovnanie SER štyroch variant LDA použitých s každým z klasifikátov.

Tabuľka 5.4 porovnáva úspešnosť jednotlivých klasifikátorov a variánt LDA. Najlepšie výsledky sú opäť dosiahnuté použitím dát so striedaním jazykov v LDA-ME a Gauss-ME. Najnižšiu úspešnosť podľa očakávania dosahuje transformácia LDA-Other, kde v trénovacích dátach nebol ani jeden z cieľových jazykov. Varianta LDA-All napriek nutnosti modelovať pridané triedy nedosahuje výrazne nižšiu úspešnosť ako LDA-Mix, ktorá modeluje len dva jazyky. Na základe výsledkov tohto experimentu možno urobiť záver, že pre minimalizáciu chyby systému je vhodné použiť na výpočet LDA cieľové jazyky. Zaujímavým

zistením je, že rôzne klasifikátory využívajúce rovnakú LDA variantu dosahujú takmer zhodné úspešnosti. Celkovú výkonnosť systému značne ovplyvňuje práve aplikácia LDA, pričom na voľbe klasifikátoru závisí minimálne.

5.4.3 Experiment s mediánovým filtrom

Ďalšie zvýšenie úspešnosti systému je možné dosiahnuť vyhladedím výslednej segmentácie nahrávky. Preto bol na postupnosť posteriórnych pravdepodobností segmentov aplikovaný jednorozmerný mediánový filter. Ten je možné použiť na filtráciu krátkych nesprávne klasifikovaných úsekov. Vo väčšine prípadov je výsledkom zníženie celkovej SER nahrávky. Na druhej strane dochádza k strate presnosti. ktorá je pri spracovaní reči so striedaním jazykov dôležitá, pretože reč sa skladá prevažne z krátkych jednojazyčných úsekov. V testovacej databáze má väčšina jednojazyčných úsekov dĺžku 0,5–1 s. Pri použití filtra piateho rádu by tieto segmenty vôbec nemuseli byť detekované. Preto musíme voliť rád filtra s ohľadom na požadovanú presnosť. Mediánový filter by bolo vhodné použiť na reč s málo častým striedaním jazykov. Graf 5.2 zobrazuje závislosť chyby systému na ráde mediánového filtru, ktorého hodnota priamo udáva presnosť.



Obr. 5.2: Závislosť SER na ráde aplikovaného mediánového filtru detekcie hranice pre rôzne varianty systémov.

Rád mediánového filtra	SER [%]	relatívne zlepšenie [%]
3	12,7	18,6
5	11,7	28,8
7	**10,8**	**30,8**
9	11,1	28,8

Tabuľka 5.5: Zlelepšenie SER aplikáciou mediánového filtra pre systém Gauss-ME.

Dosiahnuté relatívne zlepšenie SER v závislosti na ráde filtra pre systém Gauss-ME sú uvedené v tabuľke 5.5. Najlepšie výsledky boli dosiahnuté použitím filtra siedmeho rádu, kde SER klesla z 15,6 % na hodnotu 10,8 %, čo je relatívne zlepšenie o 4,8 %.

5.4.4 Fúzia dvoch systémov

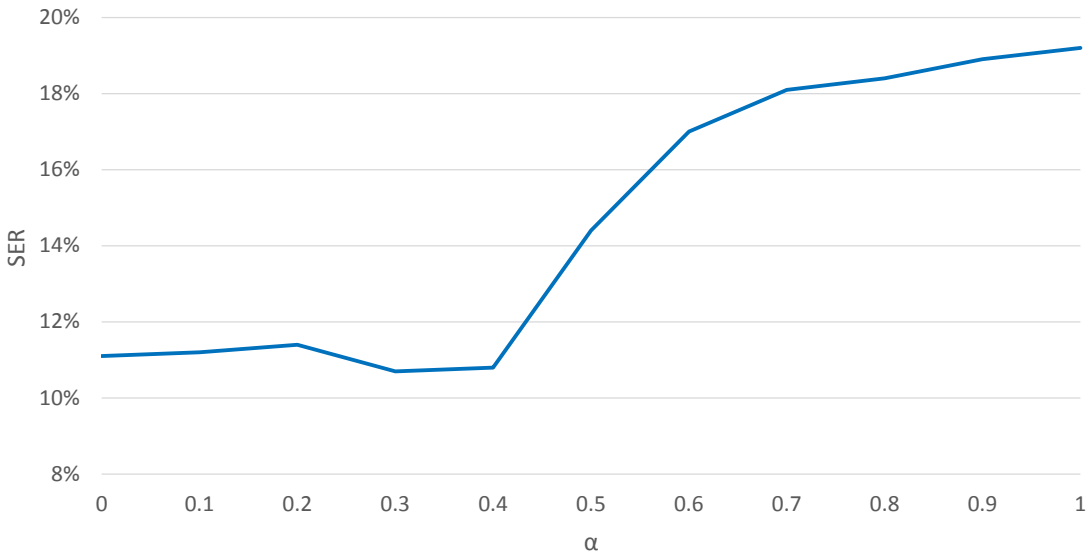
V tomto experimente boli skombinované dve varianty testovaného systému:

1. **LDA-ME s Gauss-ME klasifikátorom**, ktorý dosahoval v experimentoch najlepšie výsledky
2. **LDA-Mix s Gauss-Mix**, ktorý je vhodným komplementom k prvému podsystému, pretože je natrénovaný na iných dátach

Výsledná posteriórna pravdepodobnosť $P(c|x)$ príslušnosti segmentu x_i do triedy c je daná ako vážený priemer posteriórnych pravdepodobností oboch podsystémov:

$$P(c|x_i) = \alpha P_{Mix}(c|x_i) + (1-\alpha)P_{ME}(c|x_i) \tag{5.2}$$

kde P_{Mix} a P_{ME} sú posteriórne pravdepodobnosti systémov Gauss-Mix a Gauss-ME a koeficient α je určený experimentálne na testovacích dátach. V grafe 5.3 je zobrazená závislosť SER na hodnote koeficientu α. Pre $\alpha = 0,3$ je dosiahnutá maximálna úspešnosť systému s chybou SER = 10,7 %, s relatívnym zlepšením len 3,6 %. Zlepšenie celkovej úspešnosti systému je pomerne malé (zníženie SER o 0,4%), pretože podsystému Gauss-ME bol natrénovaný na dátach pochádzajúcich z rovnakej databázy ako testovacie dáta. Tým pádom jeho úspešnosť nemôže byť ďalej zlepšená použitím systému Gauss-Mix, natrénovanom na dátach úplne iných.



Obr. 5.3: Závislosť SER na hodnote koeficientu α. Pre $\alpha = 0$ je použitý iba systém Gauss-ME, pre $\alpha = 1$ zas len systém Gauss-Mix.

5.5 Zhrnutie experimentov a budúca práca

S implementovaným systémom bolo vykonaných niekoľko experimentov na vytvorenej databáze so striedaním mandarínčiny a angličtiny. Tieto experimenty viedli k značnému zníženiu SER. Najväšší podiel na tom má technika LDA, ktorá naviac redukciou dimenzionality i-vektorov znížila výpočetné nároky systému. Dobré zlepšenie bolo dosiahnuté tiež vyhladením segmentácie nahrávky pomocou mediánového filtru. Posledný experiment, v ktorom bola vytvorená fúzia dvoch podsystémov nepriniesol očakávané zlepšenie úspešnosti.

V tabuľke 5.6 sú porovnané úspešnosti rôznych konfigurácií systémov. Najlepšie výsledky (mimo fúzie) dosiahla podľa očakávania varianta systému, ktorý využíval klasifikátor a LDA natrénované na dátach so striedaním jazykov, ktoré boli použité aj pri vyhodnotení.

Konfigurácia systému			**SER** [%]		
Klasifikátor	**LDA**	**medfilt**	**min**	**max**	**avg**
Gauss-Mix	-	-	11,7	34,1	27,7
	LDA-Mix	-	15	32,2	25,6
		7	9,7	22,6	19,2
Gauss-ME	-	-	12,6	30,4	18,4
	LDA-ME	-	11	26,7	15,6
		7	7,4	24,1	11,1
Fúzia	LDA-Mix+LDA-ME	5	6,7	22,4	10,7

Tabuľka 5.6: Porovnanie SER pre rôzne konfigurácie systému detekcie zmeny jazyka. Stĺpce min a max obsahujú najmenšiu a najväčšiu chybu v rámci jednej nahrávky. Stĺpec avg udáva priemerný SER pre všetky nahrávky.

Nevýhodou vykonaných testov bol nedostatok kvalitných trénovacích a testovacích dát so striedaním jazykov pre objektívne vyhodnotenie systému. Ďalej chýba možnosť porovnať výsledky systému s iným súčasným systémom. Každý systém používa svoju vlastnú metriku vyhodnocovania na svojích vlastných dátach, ktoré sú často nedostupné. Jediná dostupná databáza so striedaním jazykov SEAME nebola uvoľnená dostatočne skoro pred termínom odovzdania tejto práce.

Kapitola 6

Záver a ďalšia práca

Problému detekcie zmeny jazyka v hovore je v posledných rokoch venovaná stále väčšia pozornosť. Dôvodom je častejšie sa vyskytujúca multiligválna komunikácia. S rozpoznávaním reči so striedaním kódov však majú automatické rozpoznávače reči problém. Poskytnutie informácií o diarizácií reči má výrazný vplyv na celkovú úspešnosť rozpoznávania reči so striedaním kódov.

V úvode tejto práce je uvedený prehľad metód zaoberajúcich sa problematikou diarizácie jazyka. V tomto prehľade je možné pozorovať veľký pokrok v spôsobe riešenia tohto problému od základných prístupov až po pokročilé metódy používané v súčasnosti. Ďalej sú v práci popísané základné prístupy k identifikácií jazyka, ktorými je inšpirovaná väčšina systémoch diarizácie jazyka, vrátane systému implementovaného v tejto práci.

Pre implementáciu a testovanie bol zvolený akustický systém diarizácie jazyka využívajúci kombináciu GMM a i-vektorov pre modelovanie akustických informácií v krátkych úsekoch reči. Využíva tiež techniku redukcie dimenzionality LDA a vyhladenie výslednej segmentácie.

K účelu vývoja a vyhodnotenia systému bola z rádiových nahrávok vytvorená mandarínsko-anglická databáza so striedaním kódov. Na trénovanie boli použité tiež jednojazyčné dáta získané z rôznych databáz.

So systémom bolo vykonaných niekoľko experimentov. V nich sa osvedčila technika LDA ktorá priniesla relatívne zlepšenie miery chybne klasifikovaných segmentov približne o 40 %. Následná aplikácia mediánového filtru na výslednú segmentáciu relatívne zlepšuje úspešnosť systému o 30 %. Fúzia dvoch podsystémov priniesla zlepšenie iba 3,6 % z dôvodu veľkého rozptylu úspešnosti dvoch skombinovaných systémov. Systém dosahuje na vytvorenej databáze veľmi dobré výsledky s mierou chybne identifikovaných segmentov 10,7 %.

6.0.1 Ďalšia práca

V rámci budúcej práce na vytvorenom systéme by bolo vhodné vykonať nasledujúce úpravy a rozšírenia:

- Vyhodnotiť systém na rozsiahlejšej databáze so striedaním kódov (napr. SEAME). Systém je v súčasnom stave testovaný na databázi s malým počtom dát, pomocou ktorých nie je možné jazyky dostatočne dobre modelovať.
- Spresniť pozíciu detekovaných hraníc jazyka použitím resegmentácie. Po počiatočnej segmentácií nahrávky na jednojazyčné úseky je možné vykonať jemnejšiu resegmentáciu v miestach detekovanej zmeny jazyka a segmentáciu spresniť.

- Upravenie metriky vyhodnocujúcej presnosť detekcie hranice jazyka. V súčasnom stave sú za hranice jazyka považované hranice po sebe idúcich segmentov klasifikovaných do rozdielnych tried. Vhodnou voľbou by bola metrika LBE (Language Boundary Error) z [2].
- Spojenie implementovaného akustického systému s fonotaktickým. Túto architektúru využíva väčšina súčasných systémov diarizácie jazyka a implementovaný systém by sa stal konkurencieschopným.

Literatúra

[1] Brümmer, N.; Cumani, S.; Glembek, O.; aj.: Description and analysis of the Brno276 system for LRE2011. In *2012 IEEE Odyssey - The Speaker and Language Recognition Workshop*, Singapore, Jún 2012.
URL http://www.fit.vutbr.cz/research/groups/speech/publi/2012/brummer_odyssey2012_216-223-40.pdf

[2] Chan, J. Y. C.; Ching, P. C.; Lee, T.; aj.: Detection of language boundary in code-switching utterances by bi-phone probabilities. In *SympoTIC '04. Joint 1st Workshop on Mobile Future*, Hong Kong, China, 2004, s. 327–328.
URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1409644

[3] Chen, S.-Y. R.: Code-switching between English and Mandarin Chinese. In *Lancaster University Postgraduate Conference in Linguistics & Language Teaching*, ročník 1, Lancaster, United Kingdom, 2007.
URL http://www.ling.lancs.ac.uk/pgconference/v01/Chen.pdf

[4] Li, D. C. S.; yan Chan, H.; Hok-Shing, B. C.; aj.: Cantonese-English code-switching research in Hong Kong: a Y2K review. *World Englishes*, ročník 19, č. 3, 2000: s. 189–202.
URL
http://202.116.197.15/cadalcanton/Fulltext/20908_2014317_103645_24.pdf

[5] Liu, T.; Liu, X.; Yan, Y.: Speaker Diarization System Based on GMM and BIC. December 2006.
URL http:
//www.isca-speech.org/archive_open/archive_papers/iscslp2006/B11.pdf

[6] Lyu, D.-C.; Chng, E.-S.; Li, H.: Language diarization for code-switch conversational speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

[7] Lyu, D.-C.; Lyu, R.-Y.; chin Chiang, Y.; aj.: Speech Recognition on Code-Switching Among the Chinese Dialects. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, Máj 2006, ISSN 1520-6149.

[8] Martínez, D. G.; Plchot, O.; Burget, L.; aj.: Language Recognition in iVectors Space. In *Proceedings of Interspeech 2011*, ročník 2011, International Speech Communication Association, 2011, ISBN 978-1-61839-270-1, ISSN 1990-9772, s. 861–864.
URL http://www.fit.vutbr.cz/research/view_pub.php?id=9754

[9] Matějka, P.; Burget, L.; Schwarz, P.; aj.: Brno University of Technology System for NIST 2005 Language Recognition Evaluation. In *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, 2006.
URL http://www.fit.vutbr.cz/~matejkap/publi/2006/odyssey2006.pdf

[10] Ramus, F.; Mehler, J.: Language identification with suprasegmental cues: A study based on speech resynthesis. *Acoustical Society of America Journal*, ročník 105, Január 1999: s. 512–521.
URL http://cogprints.org/801/3/sasasa98.pdf

[11] Reynolds, D. A.; Quatieri, T. F.; Dunn, R. B.: Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, ročník 10, č. 1–3, 2000: s. 19–41, ISSN 1051-2004.
URL
http://web.cs.swarthmore.edu/~turnbull/cs97/f09/paper/reynolds00.pdf

[12] Sailaja, P.: Hinglish: code-switching in Indian English. *ELT Journal*, ročník 65, č. 4, 2011: s. 473–480.

[13] Schwarz, P.; Matějka, P.; Černocký, J.; aj.: Hierarchical Structures of Neural Networks for Phoneme Recognition. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Toulouse, France, 2006, s. 61–101.
URL http:
//www.fit.vutbr.cz/~matejkap/publi/2006/ICASSP2006_Schwarz_PhnRec.pdf

[14] Silovský, J.: *Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích*. Dizertační práce, Technická univerzita v Liberci, November 2011.
URL http://www.fm.tul.cz/files/autoreferat_Silovsky.pdf

[15] Zhang, H.: Code-switching Speech Detection Method by Combination of Language and Acoustic Information. In *Proceedings of the 2012 2nd International Conference on Computer and Information Applications (ICCIA 2012)*, Taiyuan, China, 2012, s. 3622–3627.
URL www.atlantis-press.com/php/download_paper.php?id=4038

[16] Černocký, J.: Předspracování řeči, tvorba řeči, cepstrum. Prednáška FIT VUT v Brně.
URL http://www.fit.vutbr.cz/study/courses/ZRE/public/pred/03_prepro_model_ceps/03_prepro_model_ceps.pdf

Dodatok A

Obsah DVD

Priložené DVD obsahuje:

- **src** – Kompletné zdrojové kódy systému detekcie zmeny jazyka v hovore, ktorý bol implementovaný v jazykoch Python a Matlab
- **data** – Mandarínsko-Anglická databáza so striedaním jazykov spolu s krátkym popisom a štatistikami
- **doc** – Programová dokumentácia implementovaného systému
- **report** – Adresár obsahujúci zdrojové kódy tejto správy v programe LaTeX
- **report.pdf** – Táto správa vo formáte PDF
- **README.txt** – Súbor s popisom obsahu DVD a krátkym popisom, ako používať systém detekcie zmeny jazyka