VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

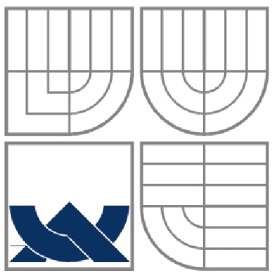
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE LIDSKÉ ŘEČI V AUDIONAHRÁVCE

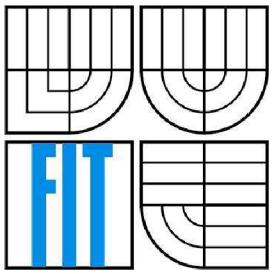
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE ROMAN BŘENEK
AUTHOR

BRNO 2011



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER SYSTEMS

DETEKCE LIDSKÉ ŘEČI V AUDIONAHRÁVCE

VOICE ACTIVITY DETECTION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — ROMAN BŘENEK
AUTHOR

VEDOUCÍ PRÁCE — MATĚJKA PAVEL, Ing., Ph.D.
SUPERVISOR

BRNO 2011

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2010/2011

Zadání bakalářské práce

Řešitel: **Břenek Roman**
Obor: Informační technologie
Téma: **Detekce lidské řeči v audio nahrávce**
Voice Activity Detection
Kategorie: Zpracování řeči a přirozeného jazyka

Pokyny:
Pro rychlejší fungování rozpoznávačů řeči či pro správnou funkci systémů na identifikaci mluvčího a jazyka je potřeba správně určit hranice kdy v nahrávce někdo mluví, nebo také vysekávání ticha mezi hovorem.

1. Seznamte se s metodami detekce řeči v audio nahrávkách
2. Naimplementujte jeden zvolený přístup
3. Svoje výsledky porovnejte na reálných datech s implementací třetí strany
4. Zhodnoťte výsledky

Literatura:
- Mluvíme s počítačem česky, Josef Psutka, Jindřich Matoušek, Luďek Muller, Vlasta Radová, ISBN 80-200-1309-1
- Dokumentace k toolkitu HTK na trénování řečových modelů - http://www.cs.tut.fi/courses/SGN-4507/htkbook.pdf

Při obhajobě semestrální části projektu je požadováno:
- bod 1

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Matějka Pavel, Ing., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2010
Datum odevzdání: 18. května 2011

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Tato práce se zabývá technikami detekce lidské řeči v nahrávkách. Je nutné při rozpoznávání správně klasifikovat všechny neřečové segmenty a naopak rozpoznat veškerou řeč i v hlučných a zašuměných prostředích. V práci je popsán celý proces rozpoznávání, tzn. digitalizace audio signálu, extrakce příznaků, trénování klasifikátoru, rozpoznávání a samotné vyhodnocení a úpravy před vyhodnocením. Pro rozpoznávání byly použity tři systémy, z nichž jeden je založen na fonémovém rozpoznávání pomocí neuronových sítí, další dva jsou založené na GMM, přičemž každý systém byl testován na třech datových sadách - Tactical Speaker Identification Speech Corpus (TSID), Ham Radio (HR) a Rich Transcription Evaluation (RT05-RT07). Nejlepší výsledky každého systému jsou pak zhodnoceny i s výsledky třetích stran.

Abstract

This thesis describes techniques for voice activity detection in audio recordings. It is necessary to correctly classify all non-speech segments and recognize speech with noisy background.
The whole process of voice activity detection (VAD) is described in this thesis, i.e. digitizing audio signal, feature extraction, training of the system, post-processing and final evaluation. There are three different systems compared within the thesis . The first one is based on phoneme recognition using neural network, the other two are variations of Gaussian Mixture Models (GMM). Each system was tested on three data sets - Tactical Speaker Identification Speech Corpus (TSID), Ham Radio (HR) and Rich Transcription Evaluation (RT05-RT07). The best results of each system are compared with the results of the third side.

Klíčová slova

extrakce příznaků, VAD, detekce řeči, GMM trénování, fonémový rozpoznávač, TSID, HR, RT

Keywords

feature extraction, VAD, voice activity detection, GMM, phoneme recognizer, TSID, HR, RT

Citace

Břenek Roman: Detekce lidské řeči v audionahrávce, bakalářská práce, Brno, FIT VUT v Brně, 2011

Detekce lidské řeči v audionahrávce

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením
Ing. Pavla Matějky Ph.D., a že jsem uvedl všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Roman Břenek
30.4.2011

Poděkování

Rád bych poděkoval mému vedoucímu Pavlu Matějkovi za trpělivou odbornou pomoc, kterou mi během mé práce poskytoval a svým rodičům za všeobecnou podporu při studiu.

Obsah

Obsah 1
1 Úvod 3
2 VAD - detekce řečové aktivity 5
2.1 Digitalizace 5
2.2 Extrakce příznaků 6
2.3 Rozpoznávání 6
2.4 Post processing 6
3 Předzpracování řečového signálu 7
3.1 Řečový signál 7
3.2 Digitalizace 7
3.2.1 Pulsní kódová modulace 7
4 Extrakce příznaků 10
4.1 Metody analýzy řečových signálů 10
4.1.1 Analýza v časové oblasti 10
4.1.2 Analýza ve frekvenční oblasti 11
4.1.3 Homomorfní analýza 11
4.2 Výběr příznaků 12
5 Databáze a pomocné nástroje 14
5.1 HTK 14
5.2 Data 14
5.2.1 RT05-07 - Rich Transcription Evaluation 14
5.2.2 TSID 15
5.2.3 HR - Ham Radio 15
5.2.4 SpeechDat-E 15
5.3 Evaluace 16
6 VAD systémy 17
6.1 Fonémový rozpoznávač 17
6.2 GMM 19
6.2.1 Trénování GMM 20
6.2.2 VAD založené na GMM 22
6.3 VAD LIA 23
7 Post-processing 24
7.1 Aplikované metody 24
7.2 Výsledky zlepšení 25

8 Experimenty a výsledky 28
8.1 Výsledky pro jednotlivé data sety 28
8.1.1 NIST-RT data 28
8.1.2 HR data 29
8.1.3 TSID data 30
9 Závěr 34
Glosář pojmů a zkratek 35
Seznam obrázků a tabulek 36
Seznam příloh 38
Literatura 39
Přílohy 41

1 Úvod

Touha člověka vyvíjet a objevovat nové věci, patří bezpochyby k lidské povaze. Naše doba nám však s příchodem informačních technologií, umožnila objevování a posunování našich možností dramaticky urychlit.

Téměř každým dnem rozšiřují výzkumní pracovníci po celém světě, hranice lidského poznání. Jedním z takových odvětví, jež si touží lidstvo přisvojit, je snaha o porozumění lidské řeči počítačem. Jak se lze dočíst v [1], trvá tato snaha již více, než čtyřicet let a nutno dodat, že úplné porozumění stroje člověku, je stále hudbou budoucnosti. Neznamená to však, že by odvětví zpracování řeči, za celá ta léta nikam nepokročilo. Spíše naopak. Je nutné si uvědomit obrovskou komplexnost mluvené řeči. Proto musíme úlohu zpracování a porozumění řeči rozdělit na několik samostatných problémů a věnovat pak pozornost každému zvlášť.

Jednou z částí, která zaznamenala obrovské pokroky, je syntéza řeči, neboli strojové čtení textových dat. Díky aplikaci prozodických charakteristik, zní dokonce člověku takto vzniklá řeč poměrně přirozeně. Využití pak lze nalézt například u lidí s poruchami hlasu, či podobnými omezeními. Nebo také pro předčítání různých textů, ať už knihy nahrané v mp3 přehrávači nebo sms zpráv nevidomým. Škála využití je opravdu velmi široká.

Do odvětví automatického zpracování řeči dále spadají úlohy, jako jsou rozpoznávání řečníka, emoční analýza, rozpoznávání slov nebo také samotná detekce lidské řeči v nahrávkách. Informace nejsou obsaženy jen ve významu pronesené řeči, ale hlas samotný vypovídá o mnoha věcech. Lze podle něj například jednoznačně identifikovat řečníka, který právě promluvil, protože hlas každého člověka je specifický, podobně jako např. otisky prstů. To může být užitečné v mnoha odvětvích, především pak pro různé bezpečností orgány, které jsou schopni z databáze nahrávek určit ty, v nichž mluvil daný člověk.

Samozřejmě lze z hlasu vyčíst také informace o pohlaví člověka, o jeho věku, ale například i o jeho momentálním psychickém rozpoložení, cíti-li strach, nervozitu, rozčílení, smutek nebo dokonce je-li opilý. Této analýzy, mimojiné, využívají např. mobilní operátoři při telefonních hovorech jejich zaměstnanců s klienty. Lze z toho určit, zda byl operátor slušný, jak vůbec rozhovor probíhal, jak se při něm zákazník cítil. Rozpoznáváním slov se pak dále kontrolují sprosté výrazy, které v hovoru případně mohou padnout.

Abychom ale nezůstali jen u telefonních hovorů, rozpoznáváním slov je dále možné, automaticky tvořit titulky při zapnutí filmu nebo vyhledávat v nahrávkách klíčová slova. Velkým přínosem pro studenty, by tato detekce klíčových slov ve videonahrávkách, mohla být při vyhledávání v záznamech z přednášek na vysokých školách (tento funkční systém lze již nalézt na http://www.prednasky.com). Dále by bylo možné pronést jakékoliv slovo v rodném jazyce a odpovědí by na to byl překlad slova do cizího jazyka, a to v psané i mluvené formě. A každý z nás by jistě dokázal vymyslet mnoho dalších využití pro tyto systémy.

Základním kamenem výše popsaných metod je však detekce lidské řeči. Každý z těchto systémů, vyjma syntézy řeči, potřebuje pro svoji další práci správně označit, kdy se v nahrávkách mluví. Dalo by se tedy říci, že se v těchto případech jedná o rozšíření a nadstavby VAD systémů (detektorů řečové aktivity), kterými se právě tato práce zabývá.

V následující kapitole se lze dočíst, z jakých částí se takový VAD systém skládá a dále pak jsou tyto části detailněji popsány. V kapitole 3 jsou to úpravy signálu před aplikací algoritmů na

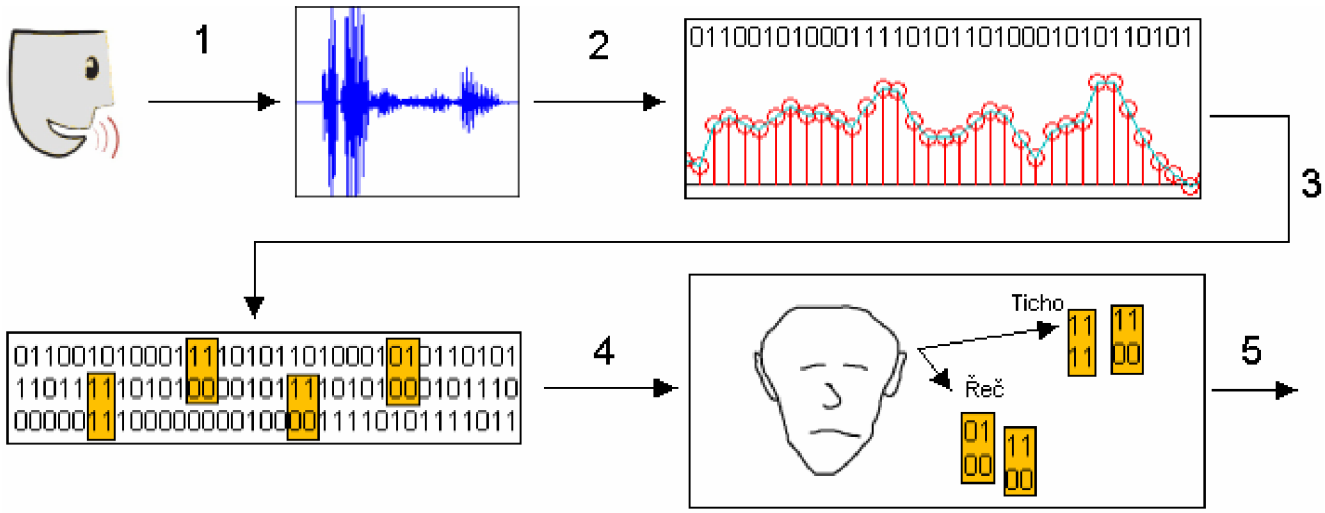
detekci řeči, v kapitole 4 je popsán výběr takových informací ze signálu, které jsou pro naší úlohu důležité. Dále jsou definovány databáze nahrávek, na kterých jsem prováděl experimenty, popis samotných VAD systémů a samozřejmě také výsledky rozpoznávání pro jednotlivé systémy na různých datech.

Oblast výzkumu a automatického zpracování řeči, má mnohem více potenciálu a uplatnění, než je popsáno výše. Zde jsou zmíněny jen některé základní, většinou již využívané, způsoby používaní. Ale mnohé z nich si však na své objevení, či zdokonalení, musí prozatím počkat.

2 VAD - detekce řečové aktivity

Detektor řečové aktivity je základní stavební kámen pro analýzu a automatické zpracování řeči. V každém systému, který s řečí pracuje, je nejdříve nutné ji rozeznat od ostatních zvuků a správně označit, kdy řečník začne a kdy skončí mluvit.[1] Pokud chceme například rozpoznávat jednotlivá slova, musíme správně rozeznat kdy se mluví a také správně rozeznat začátky a konce slov od ostatních zvuků v nahrávce. Z toho je zřejmé, že detekce řečové aktivity výrazně ovlivňuje chybovost dalších systémů, které jsou na tomto závislé.

Celý proces detekce řečové aktivity se skládá z několika částí, které lze vidět na obrázku 1.



Obr. 1 - Detekce řeči: 1 - akustický signál, 2 - zaznamenaný analogový signál, 3 - digitalizovaný signál, 4 - vektory příznaků, 5 - rozdělení do tříd `ticho` a `řeč`

2.1 Digitalizace

Číslicová reprezentace signálu je pro zpracování řeči mnohem efektivnější a pohodlnější. Analogové zpracování se v praxi téměř nepoužívá [2]. Proto se signál převádí z analogové podoby do digitální. Tento převod se dá chápat jako „sbírání“ hodnot spojitého signálu v čase, přičemž je nutné mít těchto vzorků dostatečné množství. Podrobnější popis lze nalézt v kapitole 3.2.

[1] Neplatí u řečové syntézy, kdy je postup opačný.

2.2 Extrakce příznaků

V každém řečovém signálu je ukryto velké množství informací, které tento signál charakterizují. Můžeme říci, že řečník otiskne do zaznamenaného signálu vlastnosti, které ho jednoznačně identifikují, které popisují jazyk, jakým mluvil, pohlaví řečníka a mnoho dalšího. Proto je třeba, především podle charakteru úlohy, vybrat správné vlastnosti, které by nám nejvíce vyhovovaly.

2.3 Rozpoznávání

Při rozpoznávání je nutné, aby systém obsáhl velkou proměnlivost řeči a dokázal se přizpůsobit změně prostředí. Je totiž obrovský rozdíl, pokud například mluví malé dítě v bezhlučném prostředí a dospělý člověk v jedoucím autě. Tato vlastnost, kdy se příliš nemění přesnost rozpoznávání při změně podmínek, se nazývá robustnost [1]. Robustních systémů můžeme dosáhnout vhodným a dostatečným množstvím trénovacích dat.

Pro samotné rozpoznávání lze pak volit z několika základních přístupů. Pro detekci řečové aktivity může být nejjednodušší z nich založen pouze na rozpoznávání podle energie, kdy se určí práh, od kdy je hodnota energie již řeč.

Nejčastěji využívané metody jsou založené na skrytých Markovových modelech (HMM) nebo směsi Gaussovských rozložení pravděpodobnosti (GMM). U těchto statistických přístupů se pracuje s pravděpodobností, s jakou daný mikrosegment signálu přísluší k jednotlivým třídám. Tato pravděpodobnost se určí na základě koeficientů získaných předchozím trénováním. Jiným přístupem pro trénování a rozpoznávání mohou být systémy založené na neuronových sítích.

Při rozpoznávání i trénování se ale vždy pracuje právě s příznaky, které je potřeba nejdříve extrahovat ze signálu.

2.4 Post processing

Jedná se o úpravy aplikované na výstup po rozpoznávání. Tyto úpravy se většinou také odvíjejí od smyslu zpracování, který zamýšlíme. Nejčastěji se používají různé metody vyhlazení výstupů nebo rozšiřování řečových segmentů, což však může být pro nějaké zpracování naopak nežádoucí (např. při rozpoznávání slov). Více o po-zpracování lze nalézt v kapitole 7.

3 Předzpracování řečového signálu

3.1 Řečový signál

Řečovým signálem vycházejícím od úst člověka, rozumíme soubor akustických vln ve slyšitelných frekvencích, které se tvoří v hlasivkách a dalších řečových orgánech. Celý proces vzniká v plicích, odkud se bránicí vytlačí proud vzduchu přes hlasivkovou štěrbinu, kde se rozkmitávají hlasivky. Tímto rozkmitem vznikají pravidelné vzduchové rázy, které tvoří budící signál [3]. Na konečné podobě akustického výstupu se ale podílí ještě mnoho dalších orgánů, jako jsou rty, zuby, jazyk, nostní dutina a další. Poměrně podrobný popis procesu tvorby řeči lze nalézt například v [1].

Slyšitelné frekvence člověkem jsou přibližně od 20 Hz do 20 KHz. Normální řeč, kterou se běžně komunikuje, má však kmitočtový rozsah mnohem nižší, a to od cca. 180 Hz
do 6 kHz [3].

3.2 Digitalizace

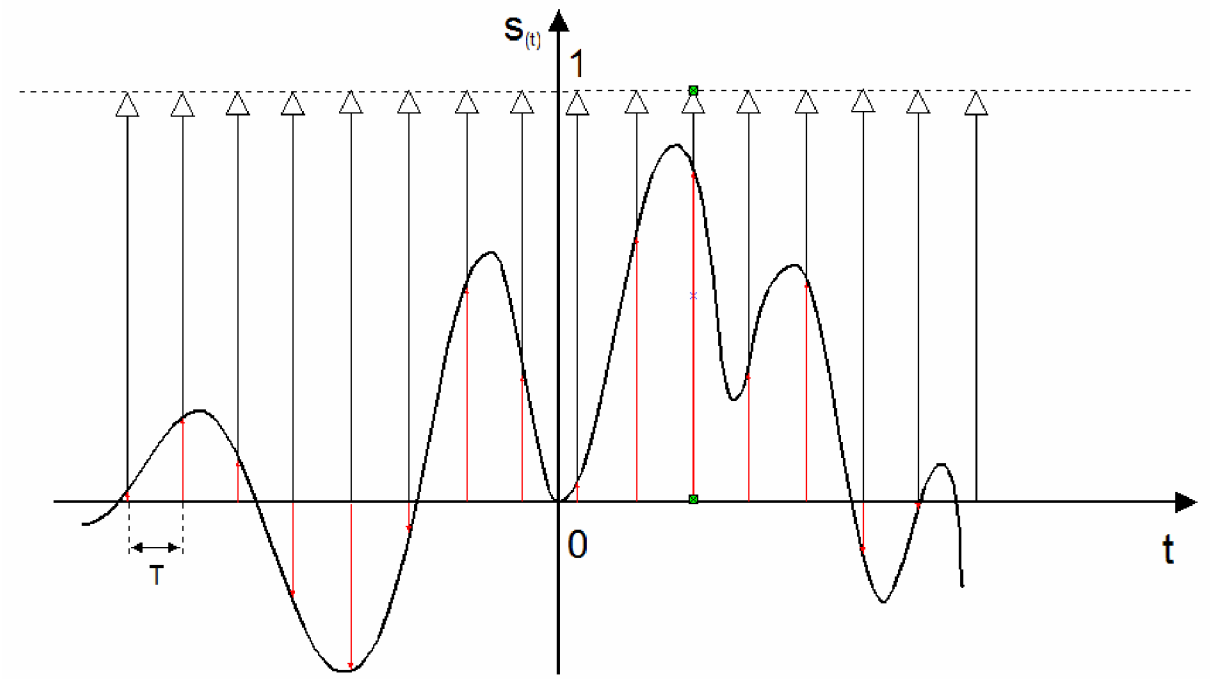
Jak jsme zmínili výše, řeč je tvořena pravidelnými kmity, které se tvoří v hlasovém traktu. Pro zaznamenání řeči je možné tyto kmity zachytit například mikrofonem, který signál uloží v analogové formě. Při zpracování řeči se však pracuje výhradně s diskrétní reprezentací signálu. Je tedy nutné signál vhodně převést na digitální podobu.

3.2.1 Pulsní kódová modulace

Pulsní kódová modulace nám zajišťuje analogově-digitální převod (digitalizaci), který se skládá ze dvou částí: a) vzorkování a
b) kvantizace.

3.2.1.1 Vzorkování

„Teoreticky vysvětlujeme vzorkování tak, že násobíme signál periodickým sledem Diracových impulzů (nekonečná výška, nulová šířka, plocha - „mocnost“ 1).
Po násobení dostaneme opět periodický sled Diracových impulzů, ale s mocnostmi danými hodnotami původního signálu v bodech *n*T (Obr. 2).“ [3]



Obr. 2 - Násobení signálu sledem Diracových impulzů.

Zjednodušeně můžeme tento proces popsat jako opakované snímání hodnot signálu v čase. Jak často budeme tyto hodnoty snímat, závisí na periodě vzorkování T. Výstupem pak bude množina hodnot analogového napětí, kterých nabývá signál v konkrétním čase. Perioda, s jakou se bude vzorkovat, musí být konstantní a musí splňovat Shannonùv-Kotelnikovùv-Nyquistùv vzorkovací teorém: *Vzorkovací frekvence musí být alespoň 2x vyšší, než max. frekvence vzorkovaného signálu.*

$$F_s > 2f_{\max} \tag{1}$$

Pokud bude frekvence nižší, než výše popsaný S-K-N teorém, bude docházet k tzv. *aliasingu*, kdy se jednotlivé kopie spektra budou překrývat. Takové spektrum bude vypadat jinak, než původní a nebude možné je zpětně rekonstruovat [4].

3.2.1.2 Kvantizace

Kvantizací aproximujeme analogové vzorky signálu konečnými číselnými hodnotami. Tyto hodnoty jsou dány úrovní kvantování, kterou máme zvolenou. Hodnoty vzorků většinou do těchto úrovní nezapadají, proto mluvíme pouze o aproximaci [1]. Tím samozřejmě dochází k jistým nepřesnostem zaokrouhlovací chybou. Tato ztráta informace se nazývá *kvantizační šum*.

Podle [3] je člověk schopen zachytit akustickou informaci o max. rychlosti 50 bitů za sekundu. Při pulsní kódové modulaci má signál značnou redundanci - např. při vzorkování

8 KHz a kódováním na 8 bitů, získáváme informační rychlost 64 000 bitů/s. Proto se v praxi spíše využívá např. *Diferenční pulsní kódová modulace*, při které se neukládají všechny vzorkované hodnoty, ale pouze některé. Zbytek se pak vypočte z předešlých uložených hodnot.

Tento postup lze aplikovat díky velké podobnosti sousedních vzorků a pomalu se měnícímu charakteru řeči. Tyto „pomalé“ změny jsou způsobeny fyzikálními omezeními, která na člověka působí. Je nutné si uvědomit, že např. jazyk je sval s nezanedbatelnou hmotností, jehož přesun v ústech zabere určitý čas. Nejkratší intervaly mezi změnami nastavení hlasového ústrojí jsou 10-25 ms [3]. Proto se při zpracování řečových signálů téměř výhradně využívá zpracování metodami krátkodobé analýzy, kdy se signál rozdělí na malé části, tzv. segmenty, které jsou dlouhé právě 10-25 ms.

4 Extrakce příznaků

Navzorkovaný digitalizovaný signál obsahuje velké množství informací, z nichž je však mnoho, pro detekci řeči, redundantních (viz. kapitola 3). Bylo by neefektivní pracovat s celou množinou navzorkovaných hodnot. Proto se využívá pro popis signálu tzv. posloupnosti příznaků nebo také vektory příznaků, které popisují jen takové vlastnosti signálu, které jsou pro nás důležité.

Pro rozpoznávání obecně, hrají příznaky velmi důležitou roli. To samozřejmě platí i pro rozpoznávání řeči. Nejdůležitějším kritériem při výběru, je vhodnost určení příznaků pro danou úlohu. Nezáleží zde zdaleka tolik na kvantitě, jako na kvalitě.

Níže si popíšeme několik základních metod získávání příznaků.

4.1 Metody analýzy řečových signálů

4.1.1 Analýza v časové oblasti

U těchto metod se nejčastěji sledují hodnoty energie signálů v čase. Mezi základní a zároveň nejjednodušší metody, které zmíníme, patří:
- Krátkodobá energie
- Počet průchodů nulou

4.1.1.1 Krátkodobá energie

Jedná se o nejjednodušší způsob detekce řeči v nahrávce. Energie řečového signálu *s(n)* na jednom segmentu o délce *N* vzorků by se dala spočítat vztahem: [3]

$$E = \sum_{n=1}^{N} s^2(n) \tag{2}$$

Tento způsob má ale značné problémy u nahrávek, které jsou zašuměné, protože pak nedokáže rozlišit šum od nízkoenergetických řečových úseků.

S těmito příznaky dále pracuje metoda vad lia, která je popsáná v kapitole 6.3.

4.1.1.2 Počet průchodů nulou

Pomocí tohoto algoritmu jsme schopni, na základě počtu průchodů signálu nulou, určit začátky a konce slov, tzn. řečové úseky. Jedná se o velmi jednoduchou základní metodu, o které se lze více dočíst např. v [3].

4.1.2 Analýza ve frekvenční oblasti

Metody založené na krátkodobé spektrální analýze, využívají nejčastěji různé modifikace *Fourierovi transformace*, díky níž jsme schopni převést časový průběh do frekvenční oblasti.
Tato úprava je definována následujícím vztahem:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2\pi jft} dt, \tag{3}$$

kde *X (f)* označujeme jako spektrum. Toto spektrum nelze analyticky přesně spočítat, proto se v praxi používá odhad Fourierovou transformací [5]. Pomocí spektrálních charakteristik se lze zaměřit na různé vlastnosti řeči, které bychom v časovém průběhu neviděli, např. různé fonetické detaily apod.

4.1.3 Homomorfní analýza

Jedná se o nelineární zpracování signálů, kdy se snažíme od sebe oddělit jednotlivé složky vzniklé konvolucí. Jak píše [3], hlasivky vytvářejí budící funkci *g(n)*, kterou dále moduluje hlasový trakt *h(n)*. Výsledný signál je pak konvolucí *g(n)* a *h(n)*. Pro rozpoznávání řeči bude ale výhodnější pracovat pouze se složkou hlasového traktu *h(n)*, protože buzení je příliš závislé na řečníkovi. [5]
Dekonvoluci dosáhneme pomocí tzv. kepstra (někdy se uvádí také cepstra).

Kepstrum

Pro získání *kepstra* z řečového signálu nejprve logaritmujeme součiny spektrálních funkcí, které získáme pomocí DFT (diskrétní Fourierovi transformace). Diky této úpravě se nám stane ze součinu součet. Dále pak pomocí inverzní Fourierovi transformace převedeme zpět do časové oblasti.
Matematicky lze zapsat následujícím vztahem: [3]

$$F^{-1}\{\log[|G(f)||H(f)|]\} = F^{-1}\{\log[|G(f)|]\} + F^{-1}\{\log[|H(f)|]\} \tag{4}$$

Nejčastěji používané příznaky v oblasti zpracování řeči, jsou v dnešní době MFCC, neboli Mel-Frekvenční-Cepstrální-Coeficienty. Jejich porovnání lze nalézt v kapitole 8.1.3.1 a podrobné výsledky pak v příloze C.

Méně používané jsou pak např. LPC či LFC koeficienty, o nichž se podrobněji píše v [1].

4.1.3.1 MFCC - MEL-frekvenční spektrální koeficienty

Jak píše [5], lidské ucho má větší rozlišovací schopnost při nižších frekvencích. Výše uvedený postup však má konstatní rozlišení v celé kmitočtové škále. Při rozpoznávání řeči se snažíme, aby bylo kepstrum co nejvíce blízké lidskému nelineárnímu slyšení. Toho se dosahuje filtry, které se lineárně rozmisťují po tzv. *Mel-ové* ose, čímž se dosáhne nelinearity na frekvenční ose. Převod mezi těmito osami je definován následujícím vztahem: [1]

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

Tyto MFCC koeficienty jsem dále využíval při experimentech na trénování systému rozpoznávání řeči i pro samotné testování.

4.1.3.2 Dynamické koeficienty

Dynamické koeficienty se spočítají z již vypočítaných statických koeficientů. Určují časové změny vektorů obsahující jednotlivé příznaky a spočítají se první derivací u `delta` koeficientů, druhou derivací u `delta-delta` koeficientů, těm se pak říká *akcelerační koeficienty*. Výsledný vektor pak obsahuje *M* statických koeficientů, *M* `delta` koeficientů a *M* *delta-delta* koeficientů, tzn. že se ztrojnásobí jeho velikost (v případě použití pouze `delta` koeficientů se velikost zdvojnásobí) [1].

Extrakce příznaků je poměrně rozsáhlá oblast, která není hlavní náplní této práce. Další metody a jejich podrobnější popis lze nalézt např. v [1].

4.2 Výběr příznaků

Čím větší bude počet dimenzí, tím větší budou nároky na výpočetní techniku; bude náročnější analýza i samotná klasifikace. Jak píše [6], v běžné praxi většinou vystačíme s 10 - 20 koeficienty.

Pro výběr se využívají dva hlavní přístupy:

a) selekce - kdy z existujících příznaků vybereme podmožinu, přičemž všechny příznaky v této podmožině zůstanou nezměněny, oproti původním

b) extrakce - kdy z množiny původních příznaků vytvoříme pomocí transformační funkce *f* podnomižinu nových příznaků

Obr. 3 - Selekce a extrakce příznaků

Mnohem častějším způsobem je metoda *extrakce* příznaků, kdy je ale nutné natrénovat vlastní transformační funkci *f*, většinou se pak jedná o LDA nebo PCA analýzy, o kterých se lze podrobněji dočíst v [7].

5 Databáze a pomocné nástroje

V této kapitole jsou především popsány databáze nahrávek, s kterými se experimentovalo, a dále pak nástroje pro extrakci příznaků a evaluaci.

5.1 HTK

Jedná se o nástroj, který je primárně určen pro rozpoznávání řeči založený na HMM (skrytých Markovových modelech). Nabízí však mnoho dalších funkcí, jako např. parametrizaci řečových signálů (extrakce příznaků) nebo samotné vyhodnocení výsledků rozpoznávání a jiné [8]. Tento software pochází z univerzity v Cambridge a pro nekomerční použití je volně k dispozici.[1] Nastavení HTK toolkitu pro mé experimenty lze nalézt v příloze B.

5.2 Data

Všechny tři databáze nahrávek, na kterých jsem testoval VAD systémy a které jsou popsány níže, měly shodný formát: raw soubory, 8 KHz, mono, 16 bitové lineární kódování, Little Endian.

Název	Počet nahrávek	Délka celkem (hod)
NIST-RT	27	19,5
TSID	640	16,8
HR Radio	12	2

Tabulka 1: Datové sady nahrávek

5.2.1 RT05-07 - Rich Transcription Evaluation

Datová sada mluvená v angličtině, která se převážně nahrávala na schůzích a poradách. Je to velmi obsáhlá databáze, která se používá především pro rozpoznávání řeči. Neobsahuje příliš šumu, je téměř čistá, ale je v ní mnoho neřečových událostí vzniklých od řečníka - smích, kašel atd.
Nahrávky a jejich evaluaci vytvořil institut standardů a technologií v USA, který užívá zkratku NIST. [2] Dále budeme tato data označovat zažitějším názvem NIST-RT.

---

[1] Domovské stránky HTK: http://www.htk.eng.cam.ac.uk/

[2] National Institute of Standards and Technology: http://www.itl.nist.gov/iad/mig/tests/rt/

5.2.2 TSID

Tactical Speaker Identification Speech Corpus (řečová databáze pro taktickou identifikaci mluvčího), je obsáhlá databáze nahrávek, v kterých se hovoří anglickým jazykem. Shromáždil je Douglas Reynolds a Gerald C. O'Leary (MIT Lincoln Laboratory).[3] V nahrávkách mluví celkem 35 lidí (čtyři ženy a 31 mužů).

Tato databáze, která obsahuje dohromady přes šest set nahrávek a téměř 17 hodin řeči, byla náhodně rozdělena na testovací (316) a trénovací data (324). Jen s podmínkou rovnoměrného rozdělení mužů a žen do každé části. Trénování VAD, které je podrobně popsáno v kapitole 5, jsem prováděl právě s touto trénovací množinou. Testovací část jsem pak zkoušel na všech systémech, s kterými jsem pracoval (viz výsledky v kapitole 8.1.3).

5.2.3 HR - Ham Radio

Jedná se o soubor velmi nekvalitních nahrávek, které jsou zašuměné a plné umělých zvuků. Označení Ham Radio se používá pro radioamatéry a i tyto nahrávky vznikly právě při amatérském radiovém spojení. Data byla stažena z internetového archivu [4] a připravil je Long Nguyen.[5]

5.2.4 SpeechDat-E

Na SpeechDat [6] databázi nahrávek byl trénovaný fonémový rozpoznávač. Velké písmeno ‘E’ v názvu znamená East - východní, což značí databázi pro východní Evropu. Obsahuje obsáhlé fonémově bohaté nahrávky telefonních hovorů pro pět různých jazyků (CZ, HU, RU, SK a PL). Soubor nahrávek pro každý jazyk je vyvážený na pohlaví, věk i dialekt, aby měl co nejširší a nejobjektivnější vypovídací hodnotu. Počet řečníků pro některé vybrané jazykové databáze ukazuje tabulka 2.

Jazyk	Mužů	Žen	Celkem
RU	1242	1258	2500
CZ	526	526	1052
HU	511	489	1000

Tabulka 2: Počet řečníků v databázích SpeechDat-E.

[3] Domovské stránky MIT lze nalézt na: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC99S83

[4] Internetový archiv: http://www.archive.org/details/HamRadioRecordings1

[5] Děkuji tímto Long Nguyen z Raytheon BBN Technologies za přípravu dat HR a také za definování dat pro Tactical SID databázi v rámci DARPA-RATS

[6] Domovské stránky SpeechDat: http://www.fee.vutbr.cz/SPEECHDAT-E/

5.3 Evaluace

Pro vyhodnocení výstupů systémů, je nutné tyto výstupy porovnat s referenčními daty. Referenční data vytváří skupina lidí, která klasifikuje úseky nahrávek do řečových nebo neřečových tříd, což by se dalo nazvat jako manuální rozpoznávání řeči.

Já jsem ve své práci k získání výsledků využíval script „md-eval-v21-.pl“ [7], což je nástroj na evaluaci vyvinutý institutem standardů a technologií NIST. Jeho výstupem je chybovost, jaké se systém dopouští, v několika kategoriích. Pro rozpoznávání řečové aktivity jsou nejdůležitější následující údaje:

- MISS speech - (missed speech) - nerozpoznaná řeč, řečové úseky, které systém označil jako ticho
- FA speech - (false alarm speech) - planý poplach, neřečové úseky označené jako řeč
- VADER - (voice activity detection error rate) - celková chyba rozpoznávání, která se spočítá jako součet ztracené řeči a planého poplachu: VADER = MISS + FA

Nastavení a spouštění evaluačního skriptu:
./md-eval-v21.pl -m -c 0.25 -r ./ref/name.rttm -s ./ output/name.rttm -u ./devset.uem

-m - zobrazení mapování
-c - nastavení tzv. límce reprezentujícího hranice referenčních segmentů, které jsou ještě v mezích [8]
-r - cesta k referenčním datům
-s - cesta k našim získaným datům
-u - cesta k souboru definujícímu oblasti, které se budou hodnotit

[7] Tento skript je možné stáhnout na: http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/code/md-eval-v21.pl

[8] Hodnota parametru -c 0.25 s jsem zvolil podle Pavla Tomáška [11].

6 VAD systémy

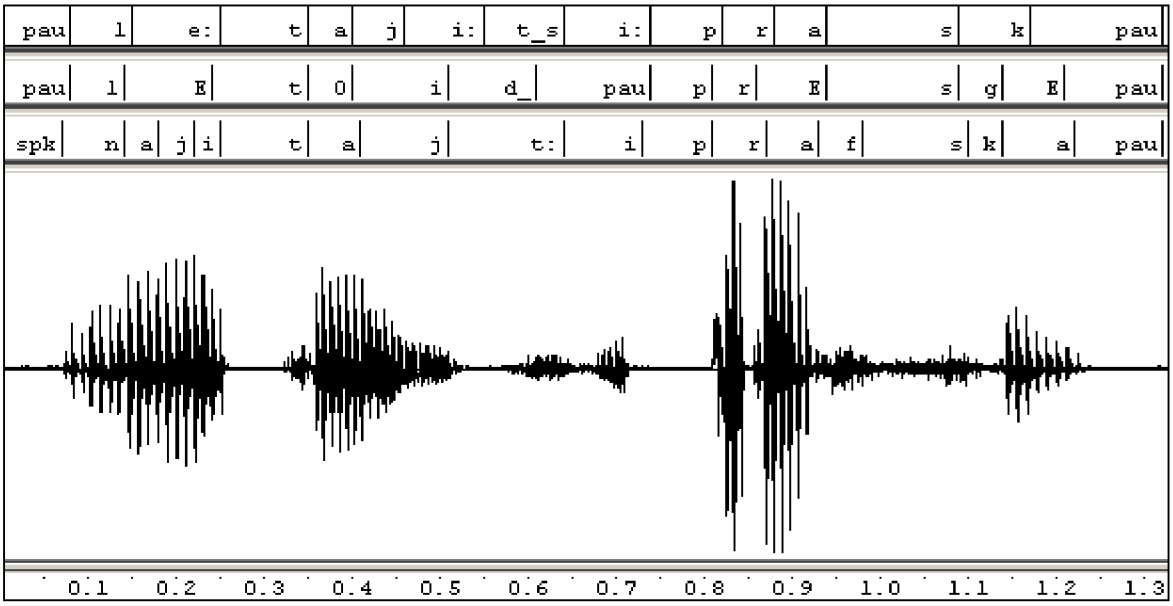
Ve své práci jsem pracoval s třemi systémy na rozpoznávání řečové aktivity. Dva byly založené na GMM a třetí z nich byl založen na fonémovém rozpoznávání. Ten se poměrně výrazně odlišuje od předchozích dvou, především svým přístupem k rozpoznávání. Je založen na neuronových sítích a výstupem jsou jednotlivé fonémy, nikoliv pouze označení časového úseku a ohodnocení řeč/ticho.

6.1 Fonémový rozpoznávač

Tento systém [1] je postaven na umělých neuronových sítích (vždy 1500 neuronů ve všech sítích [9]) a pracuje na principu rozpoznávání fonémů. Podle [1] je foném nejmenší lingvistickou jednotkou, která je schopná rozlišovat významové jednotky, jako jsou například slova. Sada fonémů se pro každý jazyk liší jak jednotlivými typy, tak jejich počtem. V češtině je například 40 fonémů, v angličtině 46, v ruštině 42 apod. [1]

Systém se tedy snaží namapovat k částem signálu nejvhodnější odpovídající foném, který k této části přiřadí (viz obrázek 4). Proto je nutné zavést ještě další fonémy, jako je například ticho nebo různé neřečové události od řečníka (smích, kašel apod.)

Celkový přehled fonémů, s kterými systém pracuje a popis některých, lze nalézt v příloze A.



Obr. 4 - Příklad výstupu fonémového rozpoznávače CZ (horní řádek), HU (prostřední) a RU (spodní řádek) pro větu „létající prase“

[1] Systém phnrec, neboli Phoneme Recognizer byl vyvinut na VUT v Brně na fakultě Informačních techologií a je možné jej pro výzkumné účely stáhnout z internetových stránek: http://speech.fit.vutbr.cz/en/software/phoneme-recognizer-based-long-temporal-context

Na Obr. 4 je zobrazen časový průběh krátké nahrávky, v které je českým jazykem proneseno spojení „létající prase“. V horní části obrázku jsou pak tři výstupy rozpoznávače. Horní řádek je výsledek rozpoznávání systému trénovaného na češtině, prostřední řádek je výstupem pro stejný systém, ale trénovaný na maďarštině a spodní řádek je pro ruštinu. Detekci řečové aktivity by měl systém zvládat bez závislosti na jazyku trénovacích nahrávek (spíše záleží na kvalitě a rozmanitosti).

Výstupem fonémového rozpoznávače je pak posloupnost fonémů, které by při správném rozpoznávání, měly tvořit pro člověka čitelný text. Na Obr. 4 můžeme vidět úspěšnost, s jakou systémy mapovaly jednotlivé fonémy. Pro český jazyk je výstup poměrně čitelný. Pro maďarštinu a ruštinu pochopitelně příliš čitelný není, protože pracuje s jinou fonémovou sadou (viz příloha A).

Transformace výstupů tohoto systému na VAD je pak poměrně snadná. Posloupnost časových úseků, kterým byl přidělen nějaký řečový foném (na Obr. 5 jsou to fonémy *n, ae, n*), se spojí do jednoho a označí se jako řeč. Fonémy typu int, pau nebo spk (viz. příloha A) se pak také spojí do jednoho celku, ale označí se jako ticho.

17397000000 17399500000 pau	……
17399500000 17401100000 pau	173,970 174,011 silence
17401100000 17401700000 n	……...
17401700000 17404100000 ae	……...
17404100000 17404800000 n	174,011 174,048 speech
17404800000 17406900000 pau	174,048 174,069 silence

Obr. 5 - Výstup fonémového rozpoznávače

Nastavení a spouštění fonémového rozpoznávače:
phnrec -c PHN_HU_SPDAT_LCRC_N1500 -l vstup.list -m vystup

-c - adresář s konfiguračními soubory
-l - seznam vstupních souborů
-m - název výstupního souboru typu MLF

6.2 GMM

GMM neboli Gaussian Model Mixtures se česky překládá jako směsi Gaussovských modelů. Jedná se o statistický způsob rozpoznávání, kdy se snažíme klasifikovat do jednotlivých tříd na základě funkcí rozložení hustoty pravděpodobnosti. Toto rozložení modelujeme pomocí Gaussova rozložení hustoty pravděpodobnosti, jehož parametry spočítáme jako:

$$\mu = \frac{1}{T}\sum_{t=1}^{T} x(t) \tag{6}$$

a

$$\sigma^2 = \frac{1}{T}\sum_{t=1}^{T} \left(x(t) - \mu\right)^2 , \tag{7}$$

kde μ je střední hodnota a σ^2 je druhá mocnina směrodatné odchylky, neboli rozptyl [10].

Hodnota funkce rozložení hustoty pravděpodobnosti se pak spočítá pro jednu gaussovku jako:

$$N\left(x;\mu,\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} , \tag{8}$$

kde x v tomto případě značí jednorozměrný příznak. My však budeme používat vícerozměrné příznaky.
Pro získání hodnoty rozložení pravděpodobnosti ve dvojdimenzionálním prostoru platí:

$$N\left(x;\mu,\Sigma\right) = \frac{1}{\sqrt{\left(2\pi\right)^P \left|\Sigma\right|}}\, e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{9}$$

kde $\sum$ je kovarianční matice, x je zde dvourozměrný vektor a μ určuje střední hodnoty [10]. Vyhodnocením tohoto výrazu dostaneme skalární hodnotu [2], která nám určuje rozdělení hustoty pravděpodobnosti pro daný vektor x [5].

---

[2] Násobíme kovarianční matici transponovaným vektorem x, čímž dostaneme řádkový vektor. Ten násobíme sloupcovou maticí středních hodnot μ a dostáváme skalární hodnotu.

Tyto případy jsou pro jednu jedinou Gaussovu křivku, kterou modelujeme výsledné rozložení. V našem případě však budeme vždy pracovat se směsí těchto křivek: [10]

$$p(x \mid \Theta) = \Sigma_c \; P_c N\left(x; \mu_c, \sigma_c^2\right) \tag{10}$$

Ke každé gaussovce jsme přidali navíc tzv. váhy P_c, které určují přechodové pravděpodobnosti, přičemž musí platit, že součet všech vah je roven jedné: [10].

$$\Sigma_c \; P_c = 1 \tag{11}$$

Výslednou hodnotu pro daný vektor získám součtem všech gaussovek násobených jednotlivými váhami.

6.2.1 Trénování GMM

Při trénování VAD systému založeném na GMM, předkládáme vzory řeči a ticha, přičemž vždy řekneme o jakou třídu se jedná (tzv. trénování s učitelem). Data, na kterých je systém trénovaný, mají významný vliv na kvalitu celého rozpoznávače. Ideálním případem by bylo rozpoznávání na podobných datech jako při trénování. Ve většině případů to tak však není. Je tedy nutné předkládat systému co nejrozmanitější nahrávky, v kterých je mnoho řečníků, různé druhy šumů a neřečových událostí.

6.2.1.1 Viterbiho trénování

Tento iterativní algoritmus upravuje (přetrénovává) parametry jednotlivých gaussovek podle vektorů, které k ní náleží. Přiřazení vektorů se dělá vynásobením jednotlivých vektorů postupně se všemi gaussovkami, čímž zjistíme pro kterou má největší hodnotu hustoty rozdělení pravděpodobnosti. Na základě této maximální hodnoty se pak rozhodne příslušnost vektoru.

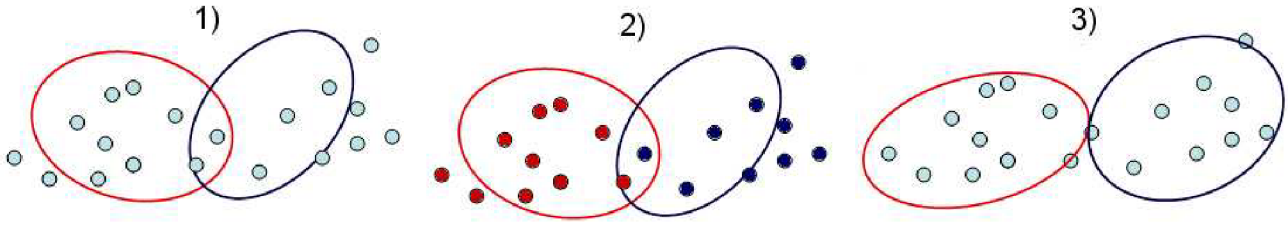
Celý model je tvořen „submodely“ z jednotlivých gaussovek, ale navenek se chová, jako by sám byl tvořen pouze jediným gaussovským rozdělením hustoty pravděpodobnosti.
Algoritmus trénování by se dal popsat následujícími kroky: [10]

1) Odhadnutí parametrů modelu
2) Přiřazení vektorů
3) Přetrénování parametrů modelu
4) Opakování kroků 2, 3

Princip tohoto algoritmu zobrazuje následující obrázek: [3]

---

[3] Inspirováno z [10].



Obr. 6 - Trénování GMM

Část 1) na obrázku 6 ukazuje náhodný odhad parametrů pro jednotlivé Gaussovo rozložení. Světle modré jsou vektory, které se snažíme rozdělit do tříd. V části 2) spočítáme maximální hodnotu funkce rozložení hustoty pravděpodobnosti, a tím určíme příslušnost jednotlivých vektorů k daným třídám. Ve třetí fázi pak podle všech vektorů, které jsme do každé třídy přiřadili, přepočítáme parametry modelu.

Na začátku trénování musíme mít všechny parametry modelu inicializované. Žádné hodnoty ale zatím neznáme, proto se je většinou snažíme odhadnout. Nepřesnost v tomto odhadování nás naštěstí příliš neovlivňuje, protože se jednotlivé parametry velmi rychle přenastaví podle vstupních dat. Každou další iterací pak modelujeme jemnější detaily výsledného rozložení, čímž zvyšujeme přesnost rozpoznávání. Jak píše [3], tyto iterace je vhodné ukončit v momentě, kdy se přestane měnit součet hodnot funkcí rozložení hustoty pravděpodobnosti (často se pro toto používá zažitý anglický termín „*likelihood*“). Ukázka výsledků, jaký vliv má počet iterací na celkovou chybu při rozpoznávání, lze nalézt v kapitole 8.1.3.3.

6.2.1.2 EM trénování

EM (Expectitation Maximization) trénování je podobné jako Viterbiho trénování, jen s rozdílem, že se jednotlivé vektory nepřiřazují jednotlivým třídám (jednotlivým rozložením), ale počítá se posteriorní pravděpodobnost, s jakou vektory do dané třídy náleží [12].

$$\gamma_c(t) = \frac{P_c \; N\left(x(t); \mu_c^{(old)}, \Sigma_c^{(old)}\right)}{\Sigma_c \; P_c \; N\left(x(t); \mu_c^{(old)}, \Sigma_c^{(old)}\right)} \tag{12}$$

Pro každý vektor a každé rozložení se tedy spočítá pravděpodobnost γ, s jakou vektor do dané třídy patří, přičemž součet pravděpodobností pro každý vektor musí být roven 1.

$$\sum \gamma(t) = 1 \tag{13}$$

Poté se parametry jednotlivých rozložení spočítají stejně jako u Viterbiho, čili průměrem vektorů, které k danému rozložení náleží, navíc ale budeme každý tento vektor násobit naší spočítanou posteriorní pravděpodobností γ, jak můžeme vidět ve vztahu (14) a (15):

$$\mu_l^{(new)} = \frac{\sum_{t=1}^{n} \gamma(t) x_t}{\sum_{t=1}^{n} \gamma(t)} \tag{14}$$

$$\Sigma_l^{(new)} = \frac{\sum_{t=1}^{n} \gamma(t)(x_t - \mu_l^{(new)})(x_t - \mu_l^{(new)})^T}{\sum_{t=1}^{n} \gamma(t)} \tag{15}$$

6.2.2 VAD založené na GMM

Níže popíši zjednodušený postup při trénování a rozpoznávání, který jsem pro tento systém aplikoval.

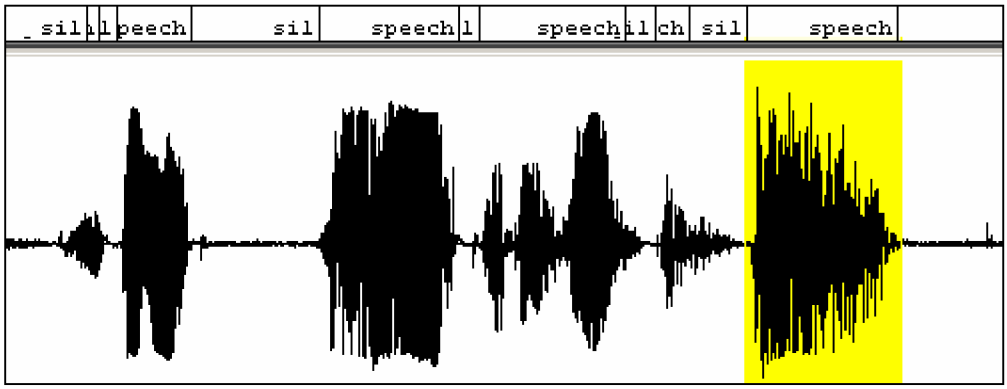
Při trénování GMM systému jsem nejprve podle referenčních dat rozdělil příznaky nahrávek do dvou skupin. Do jedné jsem zahrnul veškeré neřečové události a do druhé řeč. Každou tuto skupinu jsem pak předkládal systému zvlášť, z čehož mi vznikly dva naučené modely - jeden trénovaný pro řeč a druhý pro ticho a jiné zvuky.

Při rozpoznávání jsem pak podle extrahovaných vektorů nových nahrávek hodnotil jejich příslušnost k daným natrénovaným modelům, tzn. který vektor spadá do kterého rozložení hustoty pravděpodobnosti. Podle jeho příslušnosti jsem pak daný mikrosegment (1 ms) signálu klasifikoval jako řeč nebo jako ticho.

6.3 VAD LIA

VAD systém založený na GMM (Gaussian Mixture Models), pracující pouze s jednorozměrným příznakem - krátkodobou energií. Obsahuje tři Gaussovi křivky trénované pro různé hodnoty energie. Jedna modeluje ticho a je trénovaná na minimum, další dvě jsou trénované pro řeč a zachytávají střední hodnoty a špičky, neboli maximum.

Tento systém je inspirovaný systémem vyvinutým v LIA ve Francii (Laboratoire Informatique d'Avignon). Jeho zvláštností je, že není nutné jej předem trénovat, protože trénování proběhne vždy na samotné nahrávce, v které chceme rozpoznávat.



Obr. 7 - Rozpoznávání řeči systémem vad_lia

Obecně systémy detekující řečovou aktivitu založené na krátkodobé energii, mají problém správně určit neřečové události s vysokou energií. Na obrázku 7 je vidět špatné označení právě takového úseku, kdy v nahrávce řečník zafouká do mikrofonu (žlutě označené) a systém to označí jako řeč, ačkoliv se v nahrávce v tomto úseku nemluví. Velmi dobře lze toto pozorovat ve výsledcích testování v kapitole 8 na datech HR. I přes tento nedostatek, dosahuje systém poměrně dobrých výsledků (viz kapitola 8).

7 Post-processing

Česky by se daly tyto operace nazývat jako „po-zpracovaní“. Stejně jako se audio signál zpracovává před samotným rozpoznáváním, aby co nejvíce vyhovoval našim požadavkům, lze aplikovat některé postupy i po jeho průchodu VAD systémem, a to při vytváření výsledného souboru (v našem případě vždy .lab) nebo až na tento samotný výstupní soubor. Všechny tyto metody slouží především ke zlepšení výsledků rozpoznávání.

7.1 Aplikované metody

1. **Slučování stejných segmentů**

   Používá se především u fonémového rozpoznávače, kde výstupem je posloupnost fonémů, přičemž např. deset fonémů po sobě tvoří jedno slovo, čili řeč. Spojením těchto malých úseků do jednoho se ulehčí jednak časové náročnosti při evaluaci, a také velikosti paměti, kterou soubory zabírají.

2. **Rušení příliš malých řečových segmentů**

   Typický příklad ukazuje následující obrázek, kdy systém označil úsek v nahrávce v čase 0,35 až 0,61 jako ticho, poté na pouhou jednu setinu sekundy klasifikoval řeč, a dále pak určil několik rámců opět jako ticho.

1.	0.35 0.61 sil
2.	0.61 0.62 sp
3.	0.62 0.71 sil

Obr. 8 - Výstup VAD systému

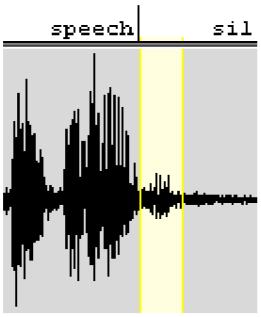
   Z kapitoly 3.2.1.2 víme, že nejkratší změny v nastavení řečového ústrojí zaberou 10-25 ms. To znamená, že pokud systém správně detekoval ticho, musel udělat chybu při detekci řeči. Proto tento řečový segment nahradíme tichem, které spojíme do jednoho celku.

3. **Rušení příliš malých segmentů ticha**

   Velmi podobná záležitost jako rušení příliš malých řečových segmentů. Pokud mezi dvěma řečovými úseky bude malý úsek ticha, můžeme jej přepsat na řeč. Tato úprava je ovšem mnohem citlivější, než výše popsaná úprava 2, proto ji lze aplikovat jen pro opravdu malé segmenty ticha, typicky do 5 ms.

4. **Rozšiřování řečových segmentů**

   Bezesporu nejdůležitější a nejefektivnější úprava po rozpoznávání, která výrazně zlepšuje úspěšnost VAD systémů.



Obr. 9 - Oříznutí řečového signálu

Na obrázku 9 je řečový signál a výstup systému vad_lia. Je vidět, že systém určí hlavní část řečového úseku, pak jej ale ustřihne pro nižší složky a dál mapuje ticho. Část vyznačená žlutě je právě takto odříznutý úsek řeči. Úpravou rozšiřování se toto snažíme eliminovat.

V následujích tabulkách lze nalézt zlepšení úspěšnosti systémů po aplikaci rozšíření řečových segmentů. Pro každý systém a každá data je toto rozšíření jiné a je tedy většinou nutné, najít správnou hodnotu rozšíření experimentálně. Tato úprava výrazně snižuje chybu MISS, neboli nedetekování řeči. Pokud by tato chyba byla malá a např. FA velké, nebude mít rozšiřování příliš velký vliv na zlepšení. Tento případ je popsán v následující podkapitole.

Rozšiřováním řečových segmentů ale zasahujeme do výsledků poměrně významným a nepružným způsobem. Hodnota rozšíření se totiž nebude měnit a bude se aplikovat na všechny výsledky na všech datech. Pro moji práci jsem pro každá data nejlepší hodnotu rozšíření experimentálně hledal. V praxi však tento postup většinou aplikovat nepůjde. Pro optimální nastavení systému je vhodnější způsob nejdříve manipulovat s prahem pro klasifikaci mezi řečí a neřečovými událostmi (resp. tichem), a až poté takto pevně nastavit rozšíření. Jeho velikost pak nejlépe zvolit jako nějaký průměr z různých dat.

7.2 Výsledky zlepšení

Následující tabulky obsahují nejlepší výsledky zlepšení, jakých jsem dosáhl pro daná data s daným systémem, a to použitím výše uvedených metod. Všechna čísla v nich uvedená jsou procentuální chybovost systémů při rozpoznávání, neboli VADER (viz. kapitola 5.3).

HR			
Systém	**Bez po-zpracování [%]**	**Po po-zpracování [%]**	**Relativní zlepšení [%]**
vad_lia	42,72	37,21	12,90
phnrec CZ	90,67	64,50	28,87
phnrec HU	82,29	51,69	37,19
phnrec RU	57,90	33,08	42,87
GMM [1]	69,33	69,33	0

Tab. 3: Výsledky zlepšení před a po aplikaci post-processingu pro data HR

[1] Celková chyba se u systému GMM po jakémkoliv rozšíření vždy zhoršila. Proto zůstala hodnota jako bez po-zpracování.

Aplikace post-processingu pro HR data má hned několik specifik. Prvním z nich je, již výše popsaný problém s vysokým FA. Když se podíváme do tabulky 3 na řádek vad_lia, vidíme pro tento systém zlepšení necelých 13 %, což je v porovnání s ostatními systémy poměrně málo. Důvod, proč tomu tak je, lze nalézt v kapitole 8.1.2 v tabulce 7, kde jsou vidět konkrétní hodnoty FA a MISS. Třetinu řečových segmentů, které systém označil, spadají do FA, proto bychom rozšiřováním „řečových segmentů“ rozšiřovali také výrazně toto chybné označení řeči.

Dalším specifikem pro tato data, byly výsledky phnrec HU a především phnrec CZ. Oba systémy měly velmi špatné výsledky. Například phnrec CZ má bez úprav 90 % chybovost při rozpoznávání. V tabulce je vidět pro tento systém necelých 30 % zlepšení, které se dosáhlo především rozšířením řečových úseků. Pro dosažení takového zlepšení bylo nutné rozšířit každý řečový segment na každou stranu o 2 sec. To samozřejmě nelze považovat za validní postup a proto musím konstatovat, že tyto dva systémy s chybovostí k 90 % pro HR data selhávají.

Příčinu je nutné hledat již při trénování, protože databáze Speech Dat-E (kap. 5.2.4) obsahuje poměrně čisté nahrávky, v kterých není mnoho šumu. Ve velmi zašuměných nahrávkách, jako jsou právě ty z databáze HR, pak označuje tento systém většinu segmentů jako ticho. Pro tento typ dat by bylo třeba systém přetrénovat.

Dále pak u systému GMM dokonce při rozšiřování řečových segmentů, nedocházelo vůbec k žádnému zlepšení, což je dánou velkou hodnotou chyby FA (viz. Tab. 3).

TSID			
Systém	**Bez po-zpracování [%]**	**Po po-zpracování [%]**	**Relativní zlepšení [%]**
vad_lia	34,88	14,74	57,74
phnrec CZ	52,11	29,01	44,33
phnrec HU	42,87	22,09	48,47
phnrec RU	32,47	14,39	55,68
GMM [2]	42,03	15,01	64,29

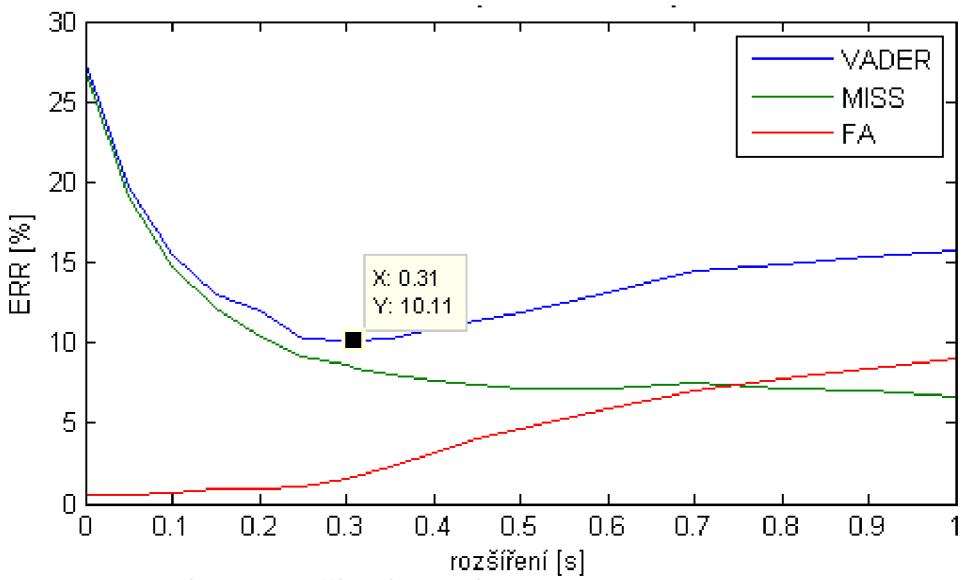
Tab. 4: Výsledky zlepšení před a po post-processingu pro data TSID

NIST			
Systém	**Bez po-zpracování [%]**	**Po po-zpracování [%]**	**Relativní zlepšení [%]**
vad_lia	38,29	9,38	75,50
phnrec CZ	35,87	14,86	58,57
phnrec HU	27,50	10,15	63.09
phnrec RU	24,92	8,75	64,89
GMM	41,96	10,96	73,88

Tab. 5: Výsledky zlepšení před a po post-processingu pro data NIST

[2] Tyto výsledky jsou pro MFCC, 4 gauss, 8 iterací, diagonální model

Největší hodnota zlepšení, které se mi v mé práci podařilo dosáhnout, je neuvěřitelných 75,5 % (viz tabulka 5, řádek vad_lia). Průměrná hodnota zlepšení napříč všemi daty a všemi systémy je sice o něco nižší, ale stále velmi velké procento: 51,07 %.



Obr. 10 - Závislost rozšíření řečových segmentů na chybě (phnrec HU - NIST)

Na obrázku 10 je znázorněna závislost rozšíření řečových úseků na celkové chybě rozpoznávání a na jejich jednotlivých složkách. V tomto demonstrativním příkladu byl použit systém phnrec HU a testovací data NIST. Hodnota 0,31 *s* pro rozšíření zde dosahuje nejlepšího výsledku, neboli nejmenší chybovosti při rozpoznávání.

U složky MISS je vidět poměrně strmé snižování chyby do hodnoty rozšíření 0,25, přičemž se složka FA nijak výrazně nezvyšuje. To je právě bod, který hledáme, přičemž je nutné podotknout, že ne vždy nastává takovýto případ, kdy hodnota FA začne dramaticky stoupat až v momentě, kdy se zlepšování MISS chyby začne ustalovat. To je pak ale většinou způsobené špatným rozpoznáváním.

8 Experimenty a výsledky

8.1 Výsledky pro jednotlivé data sety

V tabulkách níže jsou uvedené nejlepší výsledky, kterých se mi podařilo při mé práci dosáhnout. Na všechny byly aplikovány metody pro zlepšení výsledků, které jsou popsány v kapitole 7.

8.1.1 NIST-RT data

Pro tento soubor dat získaly všechny systémy nejlepší hodnocení. Jak je popsáno v kapitole 5.2.1, promluvy v nahrávkách mají poměrně dobrou kvalitu a nejsou zašuměné.

NIST			
Trénovací množina	**Miss speech**	**FA - False alarm**	**VADER**
phnrec RU	7,30	1,50	8,75
phnrec HU	8,60	1,50	10,15
phnrec CZ	12,50	2,30	14,87
vad_lia	5,60	3,80	9,38
GMM	4,00	7,00	10,96
Tomasek	3,10	2,00	5,13
SHoUT	1,50	1,70	3,20

Tab. 6: Vysledky systémů pro data NIST [1]

Abychom měli srovnání a vlastně i kontrolu, jsou-li naše výsledky vůbec v mezích nějaké validity, je vhodné porovnat své výsledky s výstupy jiných lidí, kteří se podobným problémem zabývali. Porovnání jsem udělal na datech NIST s Pavlem Tomáškem a Marijnem Huijbregtsem, kterým tímto děkuji za poskytnutí jejich výsledků.

Pavel Tomášek , vyvíjel svůj systém v rámci diplomové práce na VUT v Brně a jeho postupy a podrobné výsledky lze nalézt v [11].

Marijn Huijbregts (v tabulkách jako SHoUT) pak dosáhl svých velmi dobrých výsledků při výzkumu na univerzitě Twente v Nizozemí v rámci doktorského studia. Více o jeho systému, který nazval „SHoUT“, se lze dočíst v publikaci, kterou vydal jako svoji disertační práci [13].

---

[1] Program používaný na evaluaci zaokrouhluje na výstupu na jedno desetinné místo položky MISS a FA. VADER ale nezaokrouhluje, proto občas vznikají drobné nepřesnosti, pokud bychom sčítali z tabulek hodnoty pro získání celkové chyby (viz. tabulka 6 řádek phnrec RU).

Jak můžeme vidět v tabulce 6, oba systémy dosahují lepších výsledků, než systémy, s kterými jsem pracoval. To lze zdůvodnit např. tím, že se oba při své práci zaměřili pouze na jeden systém, který dále zlepšovali. Nutno dodat, že se touto problematikou i samotným vylepšováním rozpoznávače, zabývali mnohem více času (2 a více let).

Pro tuto práci bylo záměrem spíše zhodnotit několik existujících přístupů a pokusit se ukázat vliv různých „vylepšení“ na celkovou úspěšnost systémů. Proto můžeme konstatovat, že tyto získané výsledky lze bez jakýchkoliv pochybností považovat za validní.

8.1.2 HR data

Na těchto datech měly všechny VAD jednoznačně nejhorší výsledky. Žádný systém se nedostal s chybovostí pod 30 %, naopak některé dokonce rozpoznaly správně jen něco málo přes 30 %.

HR			
Trénovací množina	**Miss speech**	**FA - False alarm**	**VADER**
phnrec RU	28,00	5,10	33,08
phnrec HU	39,50	12,20	51,69
phnrec CZ	47,30	15,60	62,96
vad_lia	3,9	33,3	37,21
GMM [2]	18,10	51,20	69,33

Tab. 7: Výsledky systémů pro HR

Nejlépe dopadl fonémový rozpoznávač trénovaný na ruštině s 33 % chybně označených segmentů. Nejhůře pak fonémový rozpoznávač CZ, který bez post-processingu správně označil pouhých 10 % veškerých rámců. Proto můžeme tvrdit, že tento systém pro HR data naprosto selhává.

Zde je velmi dobře vidět vliv trénovací množiny dat na celkové úspěšnosti systému. Dva stejné systémy (phnrec CZ a phnrec RU) zde dosahují velmi rozdílných výsledků, přičemž jediný rozdíl mezi nimi je právě v datovém setu, na kterém se trénovaly.

Z tabulky by mohlo čtenáře na první pohled zaujmout nízké číslo Miss speech u systému vad_lia. Protože jsou data velmi zašuměné a obsahují mnoho hlasitých neřečových zvuků, tento systém, který rozpoznává pouze podle krátkodobé energie signálu, označuje většinu nahrávky jako řeč. Proto dosáhl tak dobrých výsledků v Miss speech. Tím ale samozřejmě zvyšuje FA, které je nepoměrně vyšší.

[2] Dosaženo s nastavením: 256 gauss, 8 iter, příznaky MFCC_D_A, DIAG

8.1.3 TSID data

Tato obsáhlá databáze nahrávek byla rozdělena Long Nguyenem z BBN na trénovací a testovací. Všechny níže uvedené výsledky jsou pro testovací sadu, kterou jsem používal na všechny systémy. Trénovací sadu jsem používal pro trénování GMM systému.

V následujících podkapitolách 8.1.3.1 je pak porovnání různých příznaků v závislosti na chybě a vliv nastavení počtu gaussovek a iterací v závislosti na celkovou chybovost systému.

8.1.3.1 Srovnání příznaků

Výsledky uvedené v tabulce 8 jsem získal systémem GMM (256 gauss, 8 iter) na testovacích datech TSID a pro názornost na nich nebyly aplikovány žádné další úpravy, které by mohly výsledky ovlivnit.

Přidáváním dynamických koeficientů delta a koeficientů delta-delta, můžeme pozorovat celkové zlepšení výsledků při rozpoznávání oproti statickým mel frekvenčním kepstrálním příznakům. S použitím dynamických příznaků je však třeba počítat se zvýšením velikosti vektorů 2x a při použití akceleračních koeficientů delta-delta dokonce 3x, oproti statickým.

	FEA	Miss speech	FA - false alarm	VADER
GMM s diagonální kovarianční maticí	MFCC_0	34,80	3,10	37,94
	MFCC_0_D	29,10	3,10	32,27
	MFCC_0_D_A	27,90	4,40	32,34
	MFCC_E_A_D	28,00	4,50	32,52

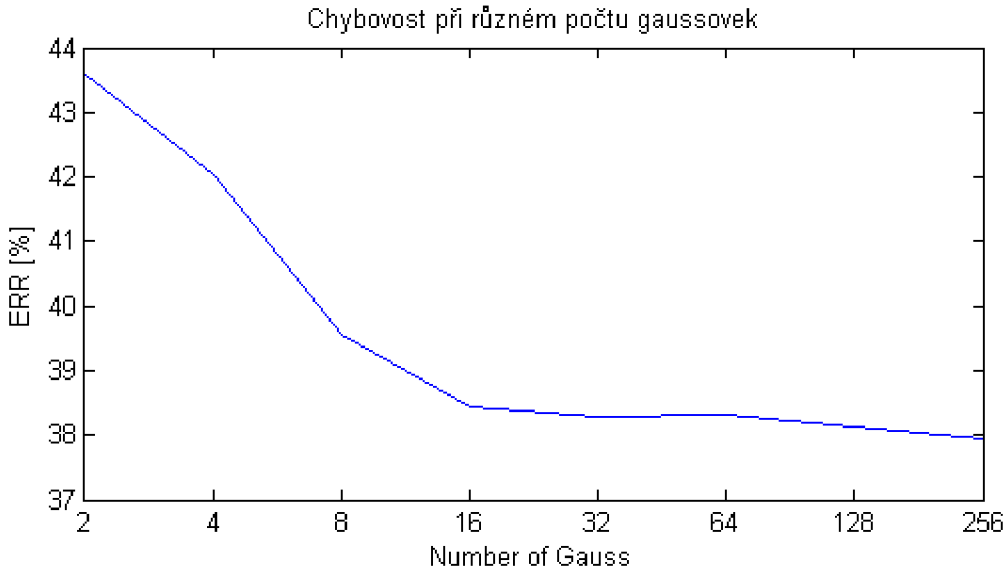
	FEA	Miss speech	FA - false alarm	VADER
GMM s plnou kovarianční maticí	MFCC_0	34,50	2,90	37,39
	MFCC_0_D	28,50	3,20	31,71
	MFCC_0_D_A	24,50	3,50	28,04

Tab 8: Srovnání chybovosti systému pro různé příznaky

Pro rozpoznávání jsem dále vybral koeficienty MFCC_0_D_A s plnou kovarianční maticí, které dosahovaly nejlepších výsledků.

8.1.3.2 Počet gaussovek v závislosti na chybě

Níže uvedené výsledky jsou výstupem systému po osmi iteracích s použitím MFCC_0 koeficientů. Celková chybovost rozpoznávání v závislosti na počtu gaussovek lze pozorovat v grafu na obr. 11. Z počátku se s přidáváním dalších Gaussovských rozložení chyba poměrně ostře snižuje. Od 16 křivek však snižování chyby ustane a je vidět, že zvyšování jejich počtu již nemá příliš velký vliv na zlepšení výsledků.



Obr. 11 - VADER při různém počtu Gaussových křivek

GMM - TSID - MFCC_0			
Gauss	**Miss speech**	**FA - False alarm**	**VADER**
2	41,30	2,3	43,60
4	39,70	2,3	42,03
8	36,50	3,1	39,55
16	35,40	3	38,45
32	35,00	3,3	38,27
64	35,10	3,3	38,33
128	35,00	3,2	38,14
256	34,80	3,1	37,94

Tab 9: Závislost chyby rozpoznávání na množství Gaussovek

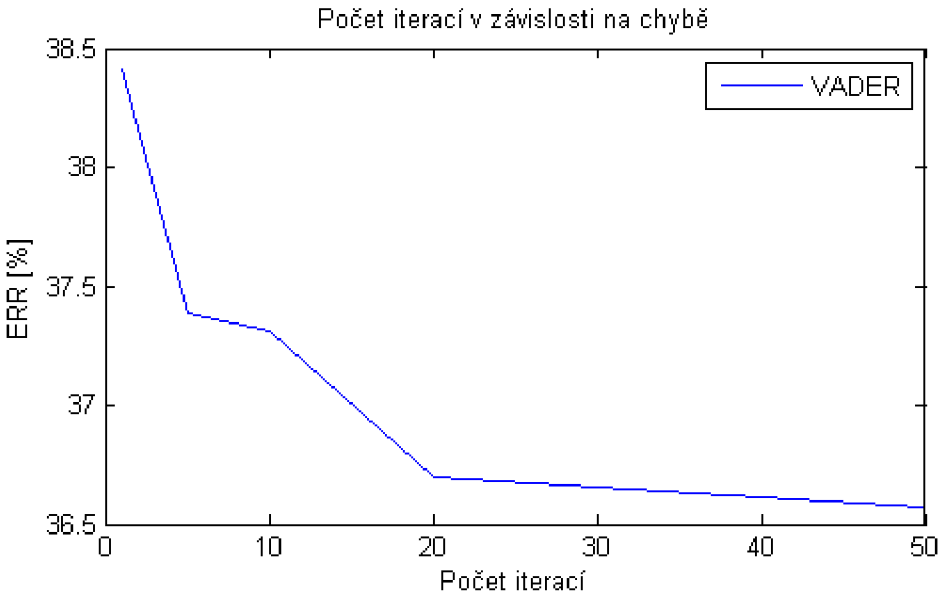
8.1.3.3 Počet iterací v závislosti na chybě

Jak je popsáno v kapitole 6.2.1, každou iterací se parametry výsledného rozložení hustoty pravděpodobnosti přesněji nastaví na předkládané vektory. Tento proces však má svůj bod nasycení, kdy se již křivky víceméně nemění a nemá proto smysl v iteracích pokračovat (více ve zmíněné kapitole).

Kromě výsledků v tabulce 10 jsem trénoval GMM systém vždy jen pro 8 iterací, a to především kvůli časové náročnosti. Jak lze vidět z grafu 12, chybovost systému se do počtu dvaceti iterací zlepší o necelé dvě procenta. Dále se ale vliv na chybu výrazně nemění.

Iterace	MISS	FA	VADER
1	35,60	2,90	38,41
5	34,50	3,00	37,39
10	34,20	3,10	37,31
20	33,30	3,40	36,70
50	33,10	3,50	36,57

Tab. 10: Výsledky GMM pro různý počet iterací [3]



Obr. 12 - VADER při různém počtu iterací

[3] Nastavení GMM: 64 gauss, diag, příznaky MFCC_0, data TSID

8.1.3.4 TSID data - výsledky systémů

Databáze TSID obsahuje velmi rozmanitý soubor nahrávek, kde některé z nich jsou i zašuměné. Nejlepší výsledky jsem získal na těchto datech s přibližně 15 % chybovostí se všemi systémy, přičemž systém GMM zde byl jako jediný trénovaný na podobných datech, jako jsou ta cílová.

TSID			
Trénovací množina	**Miss speech**	**FA - False alarm**	**VADER**
phnrec RU	12,10	2,20	14,31
phnrec HU	19,80	2,20	22,09
phnrec CZ	27,30	1,70	29,01
vad_lia	1,80	12,90	14,74
GMM[4]	6,10	8,90	15,01
BBN GMM	9,10	1,50	10,62

Tab. 11: Výsledky systémů pro TSID

Systém BBN GMM je založený na GMM a je podobný systému s kterým jsem pracoval. Navíc je zde ale aplikováno vyhlazování likelihood pomocí průměrů z okolních 50 hodnot. Výsledky pochází od společnosti BBN, která se zabývá výzkumem a vývojem v USA.

[4] Výsledek byl dosažen s následujícím nastavením: 64 gauss, 8 iter, MFCC120_D_A, Fullcov

9 Závěr

Cílem bakalářské práce bylo seznámit se s metodami rozpoznávání lidské řeči v audionahrávkách a implementace jedné z metod.

Ve své práci jsem krátce popsal některé z přístupů pro rozpoznávání řeči v časové oblasti, jako detekci na základě krátkodobé energie nebo podle počtu průchodů nulou. Podrobněji je již popsána statistická metoda pro rozpoznávání založená na směsi Gaussových rozložení hustot pravděpodobnosti - GMM, s kterými jsem dále pracoval. Systém založený na tomto přístupu jsem trénoval a upravoval pro získání nejlepších výsledků pro jednotlivá data. Toto GMM VAD využívá dvanácti-rozměrné vektory příznaků. Typ těchto příznaků, stejně jako počet gaussovek nebo počet iterací použitých v algoritmu, jsem dále zkoumal a porovnával výstupy při různých nastavení, s cílem dosáhnout co nejlepších výsledků. Tyto nejlepší výsledky, kterých jsem dosáhl během své práce, lze nalézt v kapitole 8. Pro data NIST-RT a TSID jsou získané hodnoty porovnány s výsledky třetích stran. V případě NIST-RT databáze mi pro srovnání poskytl výsledky Pavel Tomášek, který se zabývá touto problematikou v rámci své diplomové práce na FIT VUT a Marijn Hujbregts z Nizozemí, který na těchto datech testoval svůj systém SHoUT během tvorby disertační práce na univerzitě v Twente.

Se svým systémem jsem sice nedosáhl lepších výsledků, ale nutno dodat, že se nejen samotnou problematikou, ale i zdokonalováním jejich systémů, zabývali oba mnohem delší dobu (dva a více let). Pro TSID data jsem pak porovnával své výsledky s výsledky získanými v BBN, což je společnost pro výzkum a vývoj v USA.

Dalšími systémy, s kterými jsem pracoval jsou VAD LIA, který je sice také založený na GMM, ale pracuje pouze s jednorozměrným příznakem krátkodobé energie a s fonémovým rozpoznávačem vyvinutým skupinou SPEECH FIT na VUT v Brně. Na výstupy obou systémů, jsem aplikoval různé metody po-zpracování (popsané v kapitole 7), kterými jsem se snažil dosáhnout zlepšení výsledků rozpoznávání, přičemž se mi často dařilo zlepšit výsledky o více než 50 %.

Celkově nejlepších výsledků při rozpoznávání řeči, jsem dosáhl s fonémovým rozpoznávačem trénovaným na ruštině, a to pro všechna data. Pro systém GMM, se nejvíce osvědčily příznaky MFCC_D_A (mel-frekvenční kepstrální koeficienty + delta delta koeficienty) s plnou kovarianční maticí.

Glosář pojmů a zkratek

GMM	Gaussian Mixture Models
TSID	Tactical Speaker Identification Speech Corpus
HR	Ham Radio
RT	Rich Transcription Evaluation
VAD	Voice activity detection (detekce řečové aktivity)
HMM	Hidden Markov Models (skryté Markovovi modely)
LIA	Laboratoire Informatique d'Avignon
HTK	HMM Toolkit
phnrec	Phoneme Recognizer (fonémový rozpoznávač)
NIST	National Institute of Standards and Technology
MIT	Massachusetts Institute of Technology
BBN	Bolt, Beranek and Newman (společnost zabývající se výzkumem a vývojem, USA)
DARPA	Defense Advanced Research Projects Agency
RATS	Robust Automatic Transcription of Speech
MISS	Miss speech (nerozpoznaná řeč)
FA	False Alarm (neřečová událost označená jako řeč)
VADER	Voice Activity Detection Error Rate (chyba FA+MISS)
DFT	Diskrétní Fourierova Transformace
MFCC	Mel-frequency cepstral coefficient
LPC	Linear Predictive Coding
vad_lia	VAD systém založený na GMM, inspirovaný z LIA
FEA	Features (příznaky)
HU	Nahrávky v maďarském jazyce (SpeechDat) [5]
CZ	Nahrávky v českém jazyce (SpeechDat)
RU	Nahrávky v ruském jazyce (SpeechDat)
PCA	Principal Component Analysis (Analýza hlavních komponent)
LDA	Linear Discriminant Analysis (Lineární diskriminační analýza)
MLF	Master Label File

[5] Spojením zkratek HU, CZ nebo RU se systémem phnrec, se myslí fonémový rozpoznávač, který byl vytrénovaný na datech z databáze SpeechDat v daném jazyce.

Seznam obrázků a tabulek

Obr. 1 Znázornění obecného postupu při automatické detekci řeči.

Obr. 2 Proces vzorkování násobením signálu sledem Diracových impulzů.

Obr. 3 Rozdíl mezi selekcí a extrakcí příznaků.

Obr. 4 Ukázka řečového signálu v časové oblasti s výsledky fonémového rozpoznávače vytrénovaného na češtině, maďarštině a ruštině pro stejnou nahrávku.

Obr. 5 Ukázka výstupu fonémového rozpoznávače.

Obr. 6 Grafické znázornění trénování GMM systému.

Obr. 7 Výstup systému vad_lia pro ukázkový signál.

Obr. 8 Ukázka výstupního .lab souboru (ne vždy a ze všech systémů má lab soubor právě tuto syntaxi. Pro evaluaci je nutné soubory převést ještě do formátu .rttm, který pak již má jednotnou syntaxi).

Obr. 9 Ukázka rozšiřování řečových segmentů.

Obr. 10 Graf závislosti rozšíření řečových segmentů na chybě. Ukázka pro systém phnrec HU na datech NIST.

Obr. 11 Graf závislosti počtu gaussovek na celkové chybě.

Obr. 12 Graf závislosti počtu iterací na celkové chybě.

Tab. 1 Databáze nahrávek s kterými jsem pracoval.

Tab. 2 Počty řečníků v databázi Speech Dat-E.

Tab. 3 Výsledky zlepšení před a po aplikaci metod post-processingu na datech HR.

Tab. 4 Výsledky zlepšení před a po aplikaci metod post-processingu na datech TSID.

Tab. 5 Výsledky zlepšení před a po aplikaci metod post-processingu na datech NIST.

Tab. 6 Přehled nejlepších výsledků všech systémů pro data NIST + výsledky třetí strany.

Tab. 7 Přehled nejlepších výsledků všech systémů pro data HR.

Tab. 8 Srovnání úspěšnosti systémů pro různé příznaky při stejném nastavení (256 gaussovek, 8 iterací).

Tab. 9 Výsledky závislosti počtu gaussovek na celkové chybě.

Tab. 10 Ukázka závislosti počtu iterací na chybě při rozpoznávání.

Tab. 11 Přehled nejlepších výsledků všech systémů pro data TSID + výsledky třetí strany.

Tab. A1 Fonémová sada, kterou využívá systém phnrec pro český jazyk.

Tab. A2 Fonémová sada, kterou využívá systém phnrec pro maďarský jazyk.

Tab. A3 Fonémová sada, kterou využívá systém phnrec pro ruský jazyk.

Tab. C1: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0 (mel-frekvenční kepstrální foeficienty) s diagonální kovarianční maticí.

Tab. C2: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D (mel-frekvenční kepstrální koeficienty + delta koeficienty) s diagonální kovarianční maticí.

Tab. C3: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D_A (mel-frekvenční kepstrální koeficienty + akcelerační koeficienty) s diagonální kovarianční maticí.

Tab. C4: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D_A (mel-frekvenční kepstrální koeficienty + akcelerační koeficienty) s plnou kovarianční maticí.

Seznam příloh

Příloha A. Přehled fonémů pro CZ, HU a RU fonémový rozpoznávač

Příloha B. Nastavení konfiguračního souboru pro HTK

Příloha C. Vybrané detailní výsledky

Literatura

[1] PSUTKA, Josef, MŰLLER, Luděk, MATOUŠEK, Jindřich, RADOVÁ, Vlasta. *Mluvíme s počítačem česky.* 1 vyd. Praha: Academia, 2006. 752 s. ISBN 80-200-1309-1.

[2] PSUTKA, Josef. *Komunikace s počítačem mluvenou řečí.*1 vyd. Praha: Academia, 1995. 287 s. ISBN 80-200-0203-0.

[3] SIGMUND, Milan. *Analýza řečových signálů : Přednášky*. 1. vyd. Brno: Mj servis, 2000. 86 s. ISBN 80-214-1783-8.

[4] ČERNOCKÝ, Jan. *ZRE Zpracování řečových signálů: 2 přednáška, 18.2.2009*. Videozáznam z 2. přednášky předmětu ZRE na FIT VUT v Brně. 20.2.2009, [cit. 2011-5-4]. Dostupné z: https://video4.fit.vutbr.cz/av/record-download.php?id=8986

[5] ČERNOCKÝ, Jan. *Zpracování řečových signálů.* Studijní opora na fakultě Informačních technologií Vysokého učení technického v Brně, 2006. 128 s.

[6] OLAJEC, Ján. *Analýza a rozpoznávanie reči a audiosignálov.* Písemná část k disertační zkoušce na Elektrotechnické fakultě Žilinské univerzity v Žilině na katedře telekomunikací, 2006. 81 s.

[7] *Klasifikace a rozpoznávání: Extrakce příznaků*. Prezentace k 3. přednášce předmětu IKR na FIT VUT v Brně. [cit. 2011-12-4]. Dostupné z: http://www.fit.vutbr.cz/study/courses/IKR/public/prednasky/03_extrakce%20priznaku/IKR3.pdf

[8] YOUNG, Steve., et al. *The HTK Book: for HTK Version 3.3*. c 2001-2005, [cit. 2011-15-4]. Dostupné z: http://www.cs.tut.fi/courses/SGN-4507/htkbook.pdf.

[9] ŽIŽKA, Josef. *Phoneme recognizer based on long temporal kontext.* 14.1.2008, [cit. 2011-5-5]. Dostupné z: http://speech.fit.vutbr.cz/en/software/phoneme-recognizer-based-long-temporal-context

[10] BURGET, Lukáš. *Statistical Models for Automatic Speech Recognition.* Prezentace k 9. přednášce předmětu ZRE na FIT VUT v Brně. [cit. 2011-20-4]. Dostupné z: http://www.fit.vutbr.cz/study/courses/ZRE/public/pred/09_hmm_lukas/statistical_models_for_automatic_speech_recognition.ptt

[11] TOMÁŠEK, Pavel. *Speaker Diarization*. Diplomová práce na FIT VUT v Brně, 2011. 51 s.

[12] ČERNOCKÝ, Jan. *ZRE Zpracování řečových signálů: 9 přednáška, 8.4.2009*. Videozáznam z 9. přednášky předmětu ZRE na FIT VUT v Brně. 9.4.2009, [cit. 2011-6-4]. Dostupné z: https://video4.fit.vutbr.cz/av/record-download.php?id=9926

[13] HUIJBREGTS, Marijn. *Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled*. Disertační práce na Univerzitě Twente v Nizozemí, 2008. 172 s. ISBN 978-90-365-2712-5.

Přílohy

Příloha A

PŘEHLED FONÉMOVÝCH SAD V SYSTÉMU PHNREC

Fonémová sada pro český jazyk (45)

a	a:	a_u	b	c	d	d_Z	d_z	e	e:	e_u	F
f	g	h_	i	i:	int	J	J_	j	k	l	m
N	n	o	o:	o_u	P_	p	pau	r	S	s	spk
t	t_S	t_s	u	u:	v	x	Z	z			

Tab. A1: Fonémová sada systému phnrec pro český jazyk

Fonémová sada pro maďarský jazyk (61)

A:	b	b:	d	d_	d_:	dz	E	e	F	g	h
h1	i	i:	int	J	J:	j	j:	k	k:	l	l:
m	m:	N	n	n:	O	o	o:	p	pau	r	r:
S	S:	s	s:	spk	t	t:	tS	tS_	ts	ts_	t1
t1:	u	u:	v	x	y	y:	Z	z	z:	:2	_2

Tab. A2: Fonémová sada systému phnrec pro maďarský jazyk

Fonémová sada pro ruský jazyk (52)

a	a:	b	b:	d	d:	e	e:	f	f:	g	g:
i	i:	int	j	k	k:	l	l:	m	m:	n	n:
o:	p	p:	pau	r	r:	S	s	s:	spk	Ss	t
t:	tS	t_S	ts	t_s	u	u:	v	v:	x	x:	Z
z	z:	_1	1:								

Tab. A3: Fonémová sada systému phnrec pro ruský jazyk

Vysvětlivky:

int - neřečové události nezpůsobené řečníkem (např. bouchnutí dveří, troubení auta atd.)

pau - neřečové události způsobené řečníkem (např. kašel, smích, dýchání atd.)

spk - ticho

: - značí dlouhé samohlásky

Příloha B

NASTAVENÍ KONFIGURAČNÍHO SOUBORU HTK

```
SOURCEKIND      = WAVEFORM
SOURCEFORMAT    = NOHEAD        # řečové soubory jsou bez hlavičky - raw data
SOURCERATE      = 1250          # vzorkovací frekvence bude 8kHz
BYTEORDER       = VAX
TARGETFORMAT    = HTK
TARGETKIND      = MFCC_0        # typ příznaků, které se budou extrahovat
#TARGETKIND     = E

LOFREQ          = 300           # dolní pasmo
HIFREQ          = 3400          # horní pasmo
NUMCHANS        = 25            # počet troujúhelníkových filtrů
USEPOWER        = T             # použití energetického spektra
USEHAMMING      = T             # použití Hammingova okna pro řečové rámce
ENORMALISE      = F             # bez normování energie

PREEMCOEF       = 0             # bez preemfáze
TARGETRATE      = 100000        # vzorkovací perioda výstupních vektorů - 10 ms
WINDOWSIZE      = 200000        # velikost okénka - 20 ms
SAVEWITHCRC     = F
ZMEANSOURCE     = F

#CEPLIFTER      = 22
NUMCEPS         = 12             # počet kepstrálních koeficientů

ADDDITHER       = 4

WARPFREQ        = 1
WARPLCUTOFF     = 3000
WARPUCUTOFF     = 3000
```

Pro experimentování jsem extrahoval různé příznaky vždy se stejným nastavením, které je popsáno výše, pouze jsem měnil typ koeficientů v poli TARGETKIND, a to následovně:

```
TARGETKIND = MFCC_0             # mel frekvenční kepstrální koeficinety
TARGETKIND = MFCC_0_D           # MFCC + delta koeficienty
TARGETKIND = MFCC_0_D_A         # MFCC + delta + akcelerační koeficienty
TARGETKIND = MFCC_E_D_A         # MFCC + delta + akcelerační + energie
```

Příloha C

VYBRANÉ DETAILNÍ VÝSLEDKY

1. Výsledky GMM pro MFCC_0 z TSID s diagonální kovarianční maticí

Počet gauss	Rozšíření [s]	MISS speech [%]	FA [%]	VADER [%]
2	0,30	8,90	7,10	16,00
4	0,32	7,80	7,40	15,20
8	0,25	7,70	8,90	16,54
16	0,28	7,10	9,10	16,12
32	0,25	7,30	9,30	16,66
64	0,25	7,30	9,20	16,56
128	0,28	6,80	9,30	16,17
256	0,28	6,90	9,40	16,25

Tab. C.1: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0

2. Výsledky GMM pro MFCC_0_D z TSID s diagonální kovarianční maticí

Počet gauss	Rozšíření [s]	MISS speech [%]	FA [%]	VADER [%]
2	0,25	5,80	12,90	18,71
4	0,25	5,90	12,00	17,96
8	0,25	5,70	12,60	18,36
16	0,25	5,30	12,60	17,91
32	0,25	5,90	10,40	16,33
64	0,25	5,70	10,90	16,57
128	0,25	5,70	10,80	16,42
256	0,25	5,70	10,30	16,00

Tab. C.2: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D

3. Výsledky GMM pro MFCC_0_D_A z TSID s diagonální kovarianční maticí

Počet gauss	Rozšíření [s]	MISS speech [%]	FA [%]	VADER [%]
2	0,25	5,70	13,10	18,81
4	0,25	5,30	13,50	18,86
8	0,20	6,20	13,10	19,26
16	0,20	6,20	13,10	19,29
32	0,25	5,00	13,80	18,76
64	0,20	6,20	12,20	18,39
128	0,20	6,10	12,20	18,29
256	0,20	5,90	12,70	18,60

Tab. C.3: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D_A

4. Výsledky GMM pro MFCC_0_D_A z TSID s plnou kovarianční maticí

Počet gauss	Rozšíření [s]	MISS speech [%]	FA [%]	VADER [%]
2	0,21	7,40	8,70	16,12
4	0,21	7,00	10,30	17,29
8	0,21	5,90	10,00	15,91
16	0,20	6,70	9,10	15,73
32	0,22	6,40	8,80	15,26
64	0,21	6,10	8,90	15,01
128	0,20	6,00	9,50	15,50
256	0,21	5,70	10,10	15,80

Tab. C.4: Nejlepší výsledky GMM pro jednotlivé gaussovky s koeficienty MFCC_0_D_A

Všechny výsledky i pro jednotlivé nahrávky lze nalézt na přiloženém DVD.