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 TELEFONNÍCH HOVORŮ V ROZHLASOVÝCH DATECH

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. PETER JAVORSKÝ
AUTHOR

BRNO 2010

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 TELEFONNÍCH HOVORŮ V ROZHLASOVÝCH DATECH

DETECTION OF PHONE CALLS IN BROADCAST DATA

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE Bc. PETER JAVORSKÝ
AUTHOR

VEDOUCÍ PRÁCE Ing. OLDŘICH PLCHOT
SUPERVISOR

BRNO 2010

Abstrakt

Táto diplomová práca sa zaoberá detekciou telefónnych hovorov v nahrávkach z rozhlasového vysielania. Bude tu uvedený prístup, pomocou ktorého môžeme nahrávku rozdeliť na segmenty náležiace telefónnemu kanálu a segmenty, v ktorých hovorí moderátor. Úspešnosť bude vyhodnocovaná na úrovni rámcov oproti ručne anotovanej nahrávke.

Abstract

This master's thesis deals with phone call detection in recordings of radio broadcast data. I will introduce an approach, that will diarize broadcast record to segments with phone calls and segment where moderator speaks. Efficiency will be evaluated at frame level against manually annotated record.

Klíčová slova

Detekce, Diarizace, Klasifikace, Fourierova transformace, Spektrální analýza, Telefonní kanál

Keywords

Detection, Diarization, Classification, Fourier transform, Spectral Analysis, Telephone canal

Citace

Peter Javorský: Detekce Telefonních Hovorů v Rozhlasových Datech, diplomová práce, Brno, FIT VUT v Brně, 2010

Detekcia telefónnych hovorov v rozhlasových dátach

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Ing. Oldřicha Plchota. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Peter Javorský
Datum: 19. 5. 2010

Poďakovanie

Chcel by som poďakovať vedúcemu mojej diplomovej práce Ing. Oldřichovi Plchotovi za všetku pomoc, ktorú mi poskytol.

© Peter Javorský, 2010
Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů..

Obsah

Obsah 1
1 Úvod 2
2 Vlastnosti signálu preneseného cez telefónne pásmo 4
3 Metóda detekcie I 6
4 Program využívajúci metódu I 9
5 Metóda detekcie II 12
5.1 Možné vylepšenia detekcie 14
5.1.1 Vylepšenie „A“ zoskupovania výstupných rámcov metód I a II do regiónov – mediánový filter 15
5.1.2 Metóda „B“ vytvárania regiónov z výstupu metódy II 16
5.1.3 Metóda „C“ vytvárania regiónov z výstupu metódy II 18
5.1.4 Predspracovanie nahrávky s použitím segmentátora a následnou diarizáciou hovoriaceho 20
5.1.5 Metóda „D“ vytvárania regiónov z výstupu metód I a II s použitím segmentátora 23
6 Finálny detektor – projekt FindCall 24
7 Výsledky merania presnosti prvej verzie detektoru 29
8 Automatický vyhodnocovací nástroj AutoScore 31
9 Výsledky merania úspešnosti finálnej verzie detektoru 33
9.1 Nahrávka amha_c.wav 33
9.2 Nahrávka hydepark.wav 35
9.3 Nahrávka soma_b.wav 37
9.4 Nahrávka poplatky.wav 39
10 Záver 41

1 Úvod

Táto diplomová práca sa zaoberá detekciou telefónnych hovorov v nahrávkach z rozhlasového vysielania. Cieľom je takúto nahrávku rozdeliť na časti, v ktorých zvuk pochádza priamo od moderátora a časti, ktoré pochádzajú z telefónneho hovoru. Odpoveď na otázku, prečo chceme z rozhlasovej nahrávky vyfiltrovať tieto pasáže je jednoduchá – hlas, ktorý prešiel telefónom je vhodný na extrakciu príznakov pre vstup programov, ktoré slúžia napríklad na detekciu hovoriaceho, alebo detekciu jazyka konkrétneho hovoriaceho. Vstupné dáta, v ktorých by hovoriaci rozprával nejakým vopred určeným jazykom sú drahé a je ich málo. Pre niektoré jazyky tieto dáta nie sú dostupné vôbec, ani z komerčných databáz. Ako lacná alternatíva sa ponúka možnosť získať tieto dáta z vysielaní voľne dostupných rádií a televízií, ktoré nie sú zaťažené licenčnými podmienkami. Ako náhle tieto dáta nahráme, môžeme ich použiť na čokoľvek. Zdrojov sa naskytuje viacero, okrem už v dnešnej dobe nedostačujúceho analógového vysielania aj kvalitnejšie DVB-T vysielanie a živé internetové vysielania a záznamy.

V komunikáciách je pojem *wideband* relatívny výraz na popísanie širokého rozsahu frekvencií v spektre. V kontexte tejto práce to bude znamenať hlas z mikrofónu moderátora, alebo zvuk ktorý svojim frekvenčným rozsahom zasahuje aj do frekvencií, ktoré sú nad maximálnou frekvenciou, ktorú dokáže preniesť telefón. Podľa Shanonovho-Kotelnikovho-Nyquistovo vzorkovacieho teorému musí byť vzorkovacia frekvencia väčšia, ako dvojnásobok maximálnej frekvencie vo vstupnej nahrávke.

$$F_s > 2f_{max}$$

Vzorkovacia frekvencia telefónneho kanálu je 8 kHz, čo znamená, že telefónnym kanálom sa dá preniesť maximálne frekvencia 4 kHz. V praxi je táto maximálna frekvencia asi 3,475 kHz.

Po tom, čo nahrávku rozdelíme na segmenty z telefónneho kanálu a wideband kanálu (teda kanálu, v ktorom sa spektrálna hustota výkonu vyskytuje aj na frekvenciách nad 3,75 kHz a vyššie), môžeme na natrénovanie použiť buď len wideband dáta, alebo len dáta ktoré prešli telefónnym kanálom. Telefónne dáta sa môžu využiť na rozpoznávanie na tomto kanále. Príkladom môže byť call centrum tiesňového volania alebo triedenie dát z odposluchov telefónov v bezpečnostných zložkách.

Riešenie sa teda javí byť jasné – potrebujeme nástroj, ktorý dokáže nahrávku automaticky rozdeliť na segmenty pochádzajúce z mikrofónu moderátora a tie, ktoré pochádzajú z telefónneho hovoru. Spôsobov, ktoré sa nám na realizáciu tejto úlohy ponúkajú, existuje hneď niekoľko. V tejto práci popíšem tieto spôsoby a ukážem svoje riešenia na detekciu telefónnych hovorov.

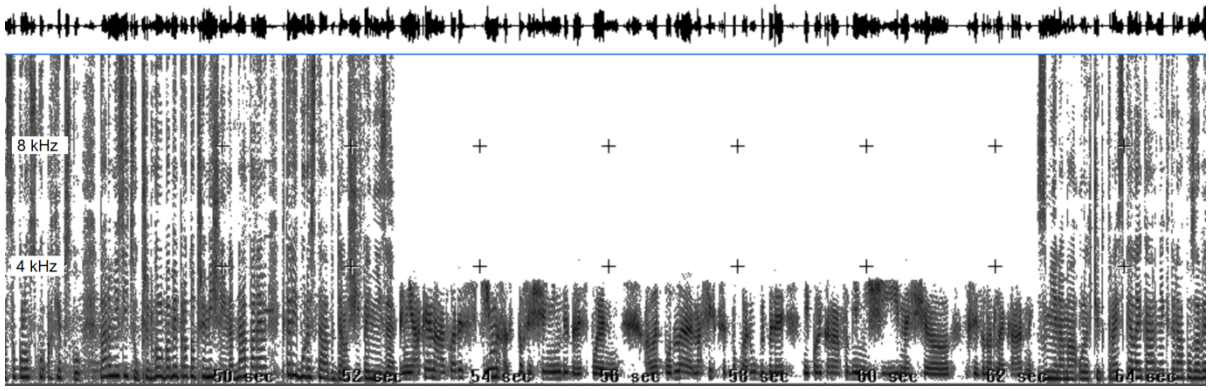
Ako som už spomenul, výstup z nástroja na detekciu telefónnych hovorov by sa mohol použiť na natrénovanie rozpoznávača jazyka, a pravdepodobne sa aj použije v NIST LRE evaluáciách. NIST (National Institute of Standards and Technology) je inštitút Americkej vlády, ktorý raz za 2 roky organizuje súťaž Language Recognition Evaluation (LRE). V roku 2008 bolo cieľom NIST evaluácií rozpoznať veľký počet jazykov a to aj takých, pre ktoré neexistujú dostupné komerčné databázy s trénovacími dátami a preto NIST poskytol účastníkom súťaže LRE ako vstupné dáta nahrávky z rádií, ktoré obsahovali detekované telefónne hovory. NIST chcel otestovať, ako sa budú takéto vstupné dáta chovať, či sa budú dať reálne použiť v praxi na natrénovanie rozpoznávačov jazyka.

V minulosti sa podobný detektor na Speech@FIT už vyvíjal. V októbri roku 2007 ho vyvinuli páni Ing. Oldřich Plchot a Ing. Petr Schwarz, Ph.D. Primárnym účelom detektoru bolo jeho použitie na datamining. Nie je to hotová aplikácia, ale akási sada skriptov. Už v roku 2007 sa výstup z tohto detektoru použil na predspracovanie dát a tieto dáta boli použité na extrakciu príznakov pre rozpoznávač jazyka. Vstupné dáta boli nahrávky z rádia v thajštine. Úspešnosť tohto programu sa explicitne netestovala, ale podľa slov samotného autora detekcia nebola spoľahlivá.

Mojou úlohou bolo preto vytvoriť nový a dokonalejší detektor, ktorého vlastnosťami by boli dobrá škálovateľnosť a nastaviteľnosť aplikácie. Výsledkom mojej práce je prvá, jednoduchšia aplikácia a hotová finálna multiplatformová aplikácia použiteľná aj z príkazového riadku, s plne funkčným GUI, ktorá je popísaná v kapitole 6. V kapitole 2 uvediem základné vlastnosti signálu, ktorý prešiel telefónnym kanálom. V kapitole 3 popíšem metódu I detekcie telefónneho hovoru, v kapitole 4 uvediem môj prvý pokus o vytvorenie funkčného programu na detekciu, využívajúci metódu I. Kapitola 5 sa zaoberá metódou II detekcie a jej vylepšeniami. Program uvedený v kapitole 6 využíva metódy I a II súčasne, a aj vylepšenie oboch metód pomocou toolkitu na segmentáciu nahrávky. Kapitola 7 sa zaoberá výsledkami merania presnosti určenia regiónov pri použití prvej verzie detektoru. V kapitole 8 popíšem pomocný nástroj na presné vyhodnocovanie úspešnosti detekcie – frame accuracy. V kapitole 9 zhrniem výsledky meraní presnosti detekcie na finálnej verzii detektoru. Kapitola 10 je záver, v ktorom sú zhrnuté všetky hlavné zistenia.

2 Vlastnosti signálu preneseného cez telefónne pásmo

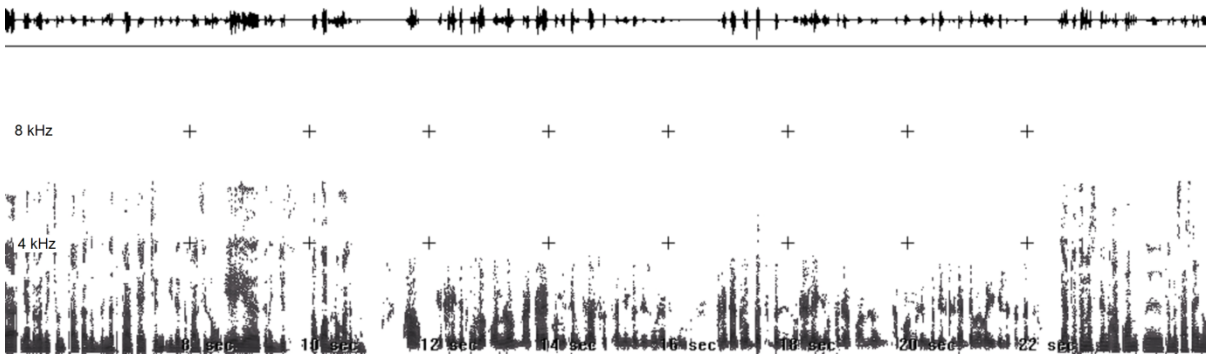
Ako rozprávame, vznikajú akustické vlny, ktoré sa v telefóne menia na elektrické signály. Tieto sa prenášajú po vybudovanom telefónnom kanáli k volanému účastníkovi a po ceste sú v dôsledku rôznych rušivých vplyvov skresľované. Preto sa signál, prichádzajúci k volanému účastníkovi, líši od signálu, ktorý generuje volajúci. Elektrický signál sa spätne premení na akustické vlny, ktoré ľudské ucho vníma ako reč. Technický systém, zabezpečujúci telefónnu službu je skonštruovaný tak, aby umožňoval prenos čo najväčšieho počtu volaní aj za cenu mierneho zhoršenia kvality hovoru. Systém zaručuje telefonujúcim uspokojivú zrozumiteľnosť, ale v žiadnom prípade nie kvalitu komunikácie tvárou v tvár. To sa prejaví tým, že hlas človeka nám cez telefón znie ináč, ako pri priamom rozhovore – chýba mu zafarbenie, charakteristické pre každého jednotlivca.[1] Analýzou nahrávok nahratých z televízie alebo rozhlasového vysielania zistíme, že hlas prenesený cez telefón je frekvenčne orezaný. Ako som už spomínal v úvode, vzhľadom na fakt, že telefónne hovory sú vzorkované na 8 kHz, telefónny kanál neprenesie žiadne frekvencie nad 3,475 kHz. Tento fakt je veľmi jasne viditeľný na spektrograme nahrávky spravodajskej relácie z DVB-T vysielania jednej českej televízie, obsahujúcej telefónny hovor, ktorý je na obrázku 1.



Obrázok 1. : Spektrálna hustota výkonu (PSD) nahrávky spravodajskej relácie z DVB-T vysielania s telefónnym hovorom. Os *x* značí čas, os *y* frekvenciu. Je jasne viditeľné, že frekvencie nad 3,475 kHz sa v nahrávke pri telefónnom hovore nevyskytujú.

Táto nahrávka je veľmi kvalitná, čo ešte viac zvýrazňuje kontrast priamo hovorenej reči a telefónneho hovoru. Vidíme, že spektrum telefónneho hovoru je frekvenčne orezané tesne pod 4 kHz. Telefónny hovor z takto kvalitnej nahrávky nie je až také ťažké detekovať, bez problémov ho zvládne aj metóda I, ktorá je popísaná v nasledujúcej kapitole.

Na kvalite vstupnej nahrávky záleží. Vo všeobecnosti platí, že čím kvalitnejšiu nahrávku máme k dispozícii, tým presnejšia detekcia bude. My ale takto kvalitné dáta mnohokrát k dispozícii nemáme. Niekedy nám musí stačiť nahrávka, ktorá je z pôvodného wideband signálu orezaná na 8 kHz, v ktorej je navyše telefónny hovor prerušovaný rôznymi náhodnými zvukmi, ktoré spadajú aj do wideband spektra. Nástroj na detekciu musí byť dostatočne robustný na to, aby si poradil aj s takýmito vstupnými dátami. Na obrázku 2 je spektrogram nahrávky z afrického rádia, ktorá nie je veľmi kvalitná. Už po prvom pohľade na tento spektrogram je jasné, že detekcia telefónneho hovoru na takejto nahrávke bude obtiažna. Bohužiaľ práve pre takéto jazyky, ku ktorým nie sú k dispozícii klasické trénovacie databázy, je kvalita získaných dát zlá. Moja výsledná aplikácia kombinujúca metódu I a II s ich vylepšením, ktoré je popísané v kapitole 5.1.4, si vie poradiť aj s takouto nahrávkou.

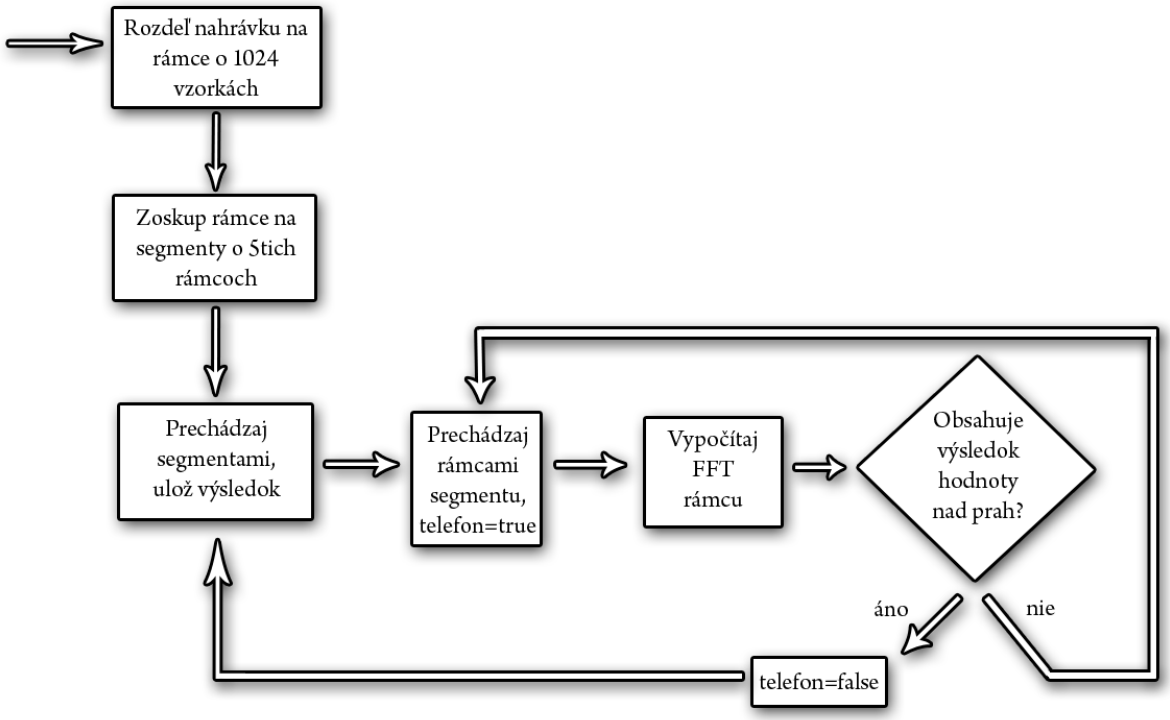


Obrázok 2. : Spektrálna hustota výkonu (PSD) nahrávky z afrického rádia s telefónnym hovorom. Os *x* značí čas, os *y* frekvenciu. Rozdiely medzi wideband signálom a telefónnym hovorom nie sú veľmi výrazné a je ťažké určiť presnú pozíciu začiatku a konca telefonátu v tejto nahrávke.

3 Metóda detekcie I

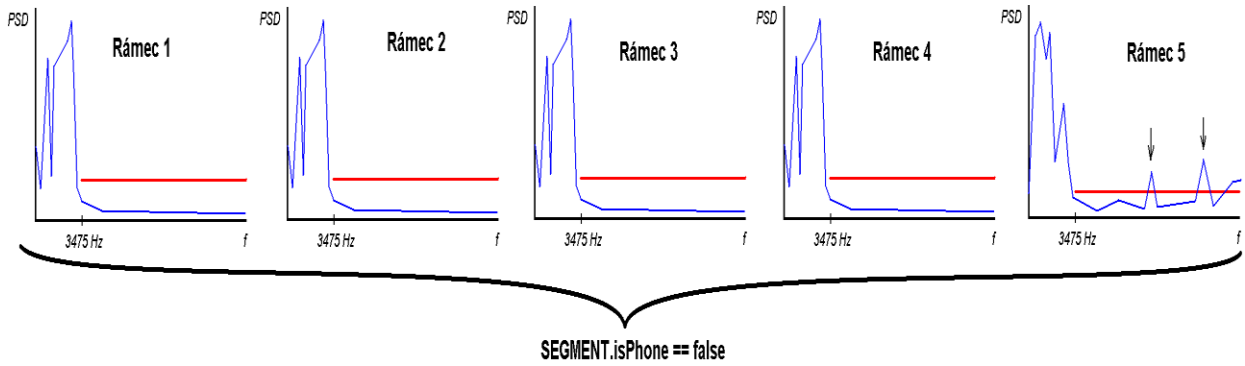
Základom väčšiny metód na detekciu telefónnych hovorov je spektrálna analýza. Nie je tomu inak ani v tejto metóde. Využíva sa poznatku, že časť nahrávky, v ktorej sa nachádza telefónny hovor neobsahuje frekvencie vyššie ako 4 kHz. Nahrávka je rozdelená na krátke segmenty, zložené z niekoľkých (v prvej verzii programu piatich) rámcov, v ktorých sa zisťuje, či obsahujú frekvencie nad 4 kHz (túto frekvenciu určíme parametrom, v 1. verzii programu (viď kapitola 4) sa dá táto frekvencia meniť posuvníkom). Najlepšie ale metóda funguje pri nastavení tejto frekvencie na 3475 Hz. Z každého rámca daného segmentu sa spočíta Fourierova transformácia a je určený prah vo výslednom PSD, ktorý sa uvažuje len od 3475 Hz vyššie. Hľadajú sa rámce, v ktorých v tejto frekvenčnej oblasti prekročí hodnota spektrálnej hustoty výkonu nejakej frekvencie tento prah. Pri analýze spektra môže hodnota výsledku Fourierovej transformácie v oblasti nad 3475 Hz pri telefónnom hovore dosahovať isté nízke hodnoty (asi 3-7% z hodnoty pri wideband reči), preto „výška“ prahu nie je nulová, ale je dostatočne nízka na to, aby pri wideband reči hodnoty výsledku FFT na nejakej frekvencii prekročila tento prah, ale aby ho neprekročili pri telefónnom hovore. Na tomto princípe je založená detekcia pomocou tejto metódy. Predpokladá sa, že spektrálna hustota výkonu vo frekvenčnej oblasti od 3475 Hz dosahuje iba zanedbateľné hodnoty. V prípade, že v ktoromkoľvek rámci segmentu prekročia hodnoty výsledku FFT v danom frekvenčnom rozsahu prah, je celý segment považovaný za wideband signál a pokračuje sa ďalším segmentom. Ak ho neprekročia, a pri tom v nižšom frekvenčnom pásme (do 3475 Hz) je určitá nenulová spektrálna hustota výkonu (musí byť vyššia ako vopred určená hodnota príslušného parametru), je tento segment považovaný za telefónny hovor. Energiu spektra v nízkych frekvenciách sledujeme pre to, aby sme eliminovali tiché časti nahrávky.

Na obrázku 3 je stavový diagram približne znázorňujúci princíp tejto metódy. Nahrávka sa rozdelí na rámce o 1024 vzorkách, bez prekrývania. (Je možné použiť aj kratšie rámce.) Tie sú následne zoskupené (iba pomyslene) na segmenty o 5-tich rámcoch. Sekvenčne sa prechádza segmentmi a ich rámcami. Na začiatku prechádzania každého segmentu sa predpokladá, že ide o telefónny hovor. Spočíta sa Fourierova transformácia každého rámca daného segmentu a v prípade, že hodnoty výsledku FFT na nejakých frekvenciách prekročia prah, (uvažuje sa iba časť spektra nad 3475 Hz, ale táto frekvencia môže byť upravená parametrom) – je celý segment považovaný za wideband. Hodnota, ktorá určuje či v danom segmente ide o telefónny hovor sa uloží do poľa a pokračuje sa ďalším segmentom. V prípade, že sa prejde všetkými rámcami segmentu a nikde hodnoty výsledku FFT prah neprekročia, (a nižšie frekvencie majú istú spektrálnu hustotu výkonu) – je segment považovaný za telefónny hovor a táto skutočnosť sa aj uloží do poľa.



Obrázok 3. : Principiálny stavový diagram metódy detekcie I

Táto metóda je relatívne primitívna a sama o sebe bez akýchkoľvek ďalších úprav sa pravdepodobne nedá použiť na nekvalitné nahrávky. Detekciu na nahrávkach z DVB-T vysielaní však zvláda výborne. Detekcia však funguje spoľahlivo až po vhodnom určení parametrov detekcie (hlavne hodnoty prahu), ktoré sú pre každý zdroj nahrávky odlišné. Nasledujúci obrázok ukazuje situáciu, kedy vo všetkých rámcoch segmentu neprekročí hodnota vypočítaného PSD v skúmanej frekvenčnej oblasti prah metódy, ale v poslednom rámci ho prekročí. V takomto prípade je celý segment považovaný za wideband.

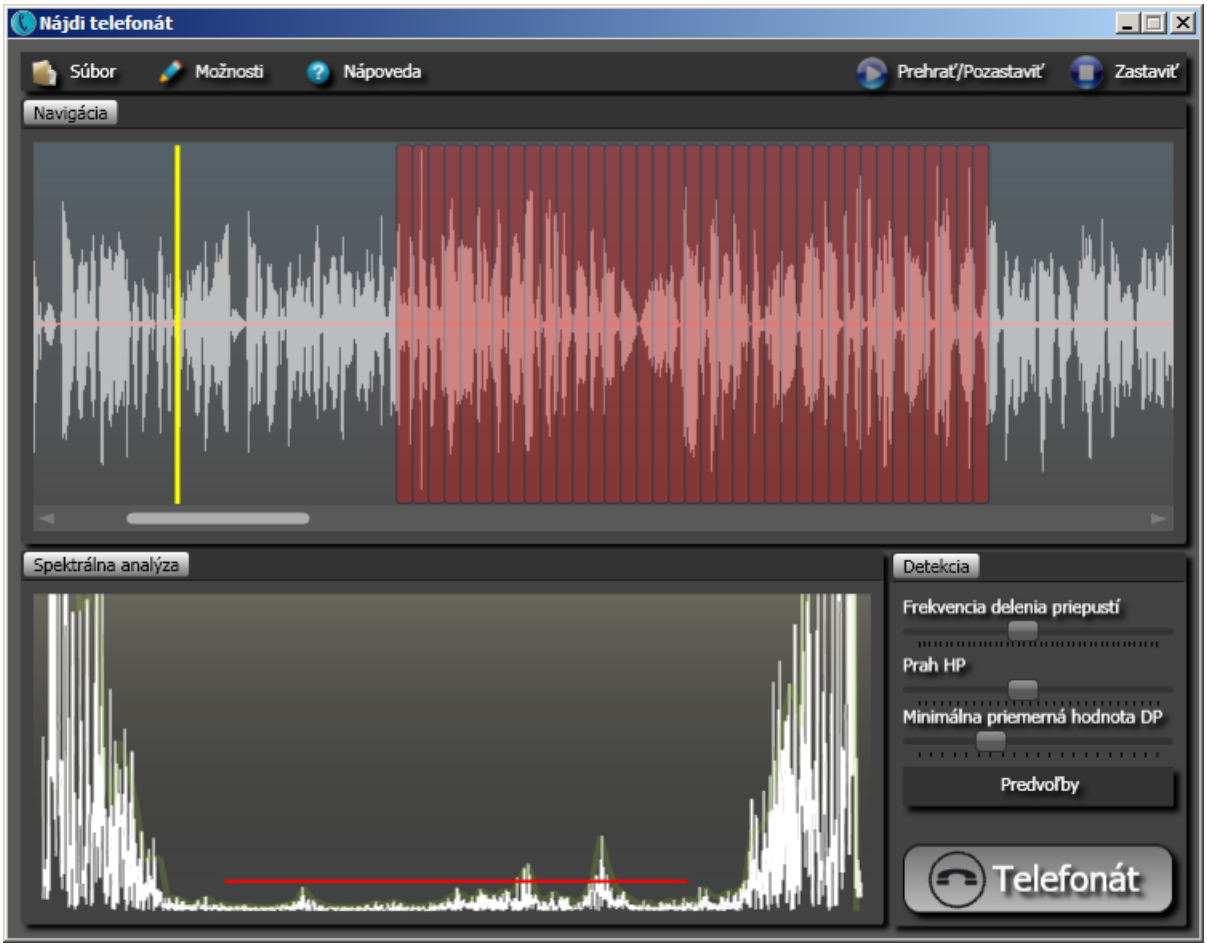


Obrázok 3b. : Červeným je vyznačený prah metódy. Modrá čiara je spektrálna hustota výkonu v danom rámci (5 rámcov na tento segment). Pri výskyte jediného rámca v segmente, takého v ktorom hodnota výsledku PSD prekročí prah metódy, je celý segment považovaný za wideband.

Parametrizovateľnosť metódy spočíva v určení frekvencie, od ktorej sa pri telefónnom hovore už nenachádzajú frekvencie nad túto hodnotu a taktiež „výška“, čiže hodnota prahu. Ďalším parametrom je určenie minimálnej sumy spektrálnej hustoty výkonu vo frekvenčnom pásme pod hraničnou frekvenciou, ktorú musí rámec obsahovať na to, aby bol považovaný za telefónny hovor. Princíp tejto metódy som vymyslel sám a na základe nej napísal program, ktorý je popísaný v nasledujúcej kapitole.

4 Program využívajúci metódu I

Tento program bol mojim prvým pokusom o napísanie detektoru telefónnych hovorov. Pôvodne nemal byť súčasťou tejto práce, jeho úlohou bolo len overiť, či je možné metódou I. detekovať nejaké telefónne hovory. Je napísaný pod platformu Windows a využíva technológiu .NET a WPF (Windows Presentation Foundation), čo je nová technológia na tvoru užívateľsky bohatého GUI. Je napísaný v jazyku C#. Screenshot tohto programu je na obrázku 4.



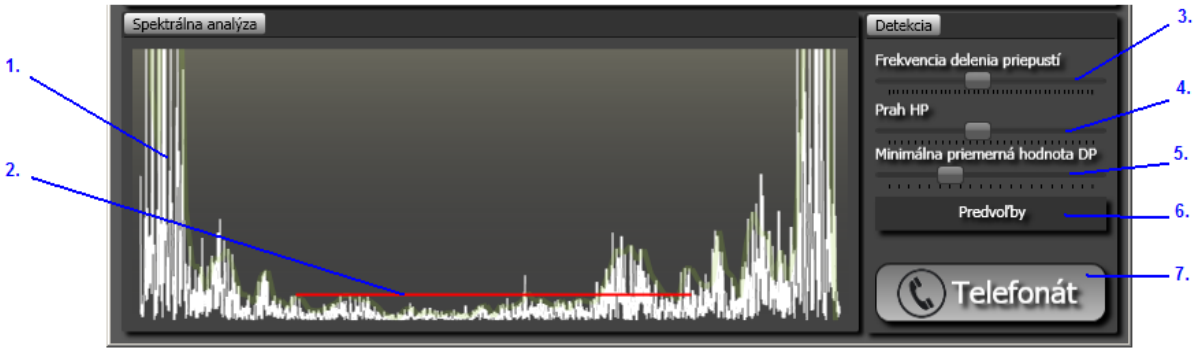
Obrázok 4. : Záber programu využívajúceho metódu I analyzujúceho nahrávku z DVB-T vysielania

Mojou snahou bolo skonvertovať tento program tak, aby ho bolo možné spustiť aj na UNIX-ových platformách, bohužiaľ po preskúmaní možností som sa dozvedel, že vývojári z projektu Mono sa rozhodli nepodporovať WPF (neplánujú ho podporovať ani do budúcnosti), čím som bol veľmi sklamaný.

Aplikácia podporuje veľkú škálu zvukových a video formátov. Sú nimi napríklad WAV, MP3, WMA, OGG, FLAC, AMR, RM, CDA, AC3, DTS, WMV, ASF, 3GP, MPG, AVI, AIF, AIFC,

AIFF, AU, SND. Podporu tak širokého spektra formátov som dosiahol pomocou špecializovanej knižnice *Alvas.Audio*, ktorá z formátov vyextrahuje zvuk a uloží ho vo formáte WAV, s ktorým potom aplikácia ďalej pracuje. Táto knižnica je v demo verzii k dispozícii zadarmo, ale pri konverzii vypíše dialóg s oznámením, že sa nejedná o plnú verziu. Mojím cieľom bolo dosiahnutie podpory aspoň formátu MP3. Na spektrálnu analýzu je použitá voľne šíriteľná knižnica FFTW. Konkrétne je použitý jej port do C#. Výhodou tejto knižnice je hlavne jej rýchlosť. Na detekciu hovorov, vykresľovanie zvuku, navigáciu v ňom, zvýrazňovanie detekovaných telefonátov a vykresľovanie spektra s maximálnymi hodnotami za určitý časový interval som si napísal vlastné funkcie. Po dokončení programu prišla požiadavka na to, aby bolo program možné spustiť aj pod UNIX-ovými platformami, pretože výsledný program bude pravdepodobne spúšťaný na školských počítačoch, na ktorých je Linux alebo BSD. A nakoľko aj metóda, ktorú využíva tento program nie je dokonalá, rozhodol som sa napísať úplne nový a dokonalejší program, ktorého vlastnosťou by bola možnosť spustiť ho na plaformách UNIX a zároveň Windows a samozrejme spoľahlivejšia detekcia telefónnych hovorov. Tento program som dlhú dobu vyvíjal a jeho finálna podoba je popísaná v kapitole 6.

Teraz popíšem parametrické časti programu. Na obrázku 5 je záber spodnej časti GUI s legendou.



Obrázok 5. : Parametrická časť UI prvej verzie programu

Legenda:

1 - **Miesto, kde sa vykresľuje spektrum** z aktuálnej pozície v nahrávke (spektrum je vykresľované z FFT dvoch rámcov, je zrkadlené - v strede sú vysoké frekvencie, na krajoch nízke) Nevýrazná žltá krivka obťahuje maximálne hodnoty výsledku FFT za poslednú sekundu – tzv. peaky. (viď obrázok 6)

2 - **Prah detekcie** – znázorňuje rozsah, ktorý sa vo výsledku FFT berie do úvahy (dĺžka čiary) a istú hodnotu prahu (výška čiary), ktorú ak niektorá z hodnôt výsledku prekročí, je celý segment považovaný za wideband.

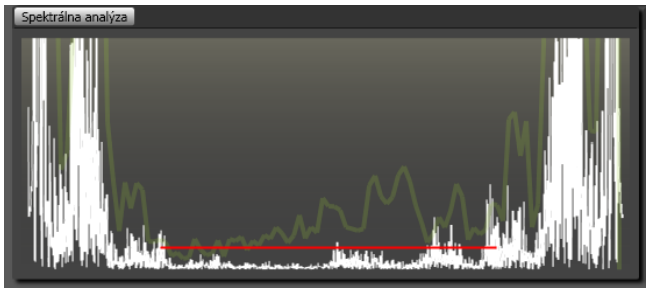
3 - **Posuvník frekvencie** - určuje, od akej frekvencie nahor sa bude detekovať, východziu hodnotu a hodnoty prednastavených hodnôt som volil empiricky. [1]

4 - **Posuvník hodnoty prahu** – určuje výšku prahu, teda hodnotu ktorú výsledok FFT v žiadnej zo skúmaných frekvencií nesmie prekročiť na to, aby bol segment považovaný za telefónny hovor.

5 - **Posuvník minimálnej priemernej hodnoty energie nížších frekvencií** – na to, aby bol segment považovaný za telefónny hovor, musí spektrálna hustota výkonu v pásme pod deliacou frekvenciou obsahovať istú energiu, vylúčia sa tým tiché segmenty.

6 - **Volič predvolieb** – hodnoty je možné nastaviť z niektorých z prednastavených možností, ktoré boli volené empiricky. (viď obrázok 7)

7 - **Indikátor telefónneho hovoru** – zmení farbu pri prehrávaní telefónneho hovoru.



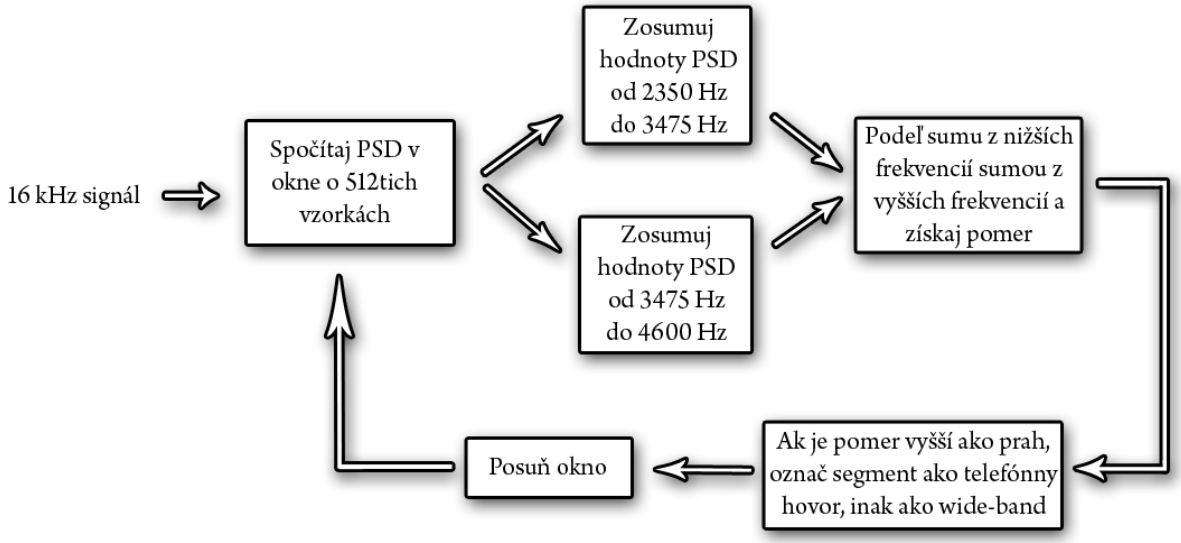
Obrázok 6. : Vykreslenie maximálnych hodnôt FFT za poslednú sekundu (žltým)

Obrázok 7. : Voľba prednastavených hodnôt

[1] Zdrojový kód programu nenasleduje najlepšie praktiky písania aplikácií a obsahuje množstvo magických konštánt. Niektoré hodnoty parametrov sú uvažované iba ako relatívne k nejakým iným konštantám. Napríklad východzia hodnota prahu je 13, a značí že výška prahu je 13% z konštanty určujúcej energiu spektra (v danom frekvenčnom pásme) pri normalizovanej reči. Vzhľadom na fakt, že vznikla nová 2. verzia programu (viď kapitola 6), v ktorom už sú správne praktiky programovania dodržiavané, zdrojové kódy 1. verzie neboli refaktorizované.

5 Metóda detekcie II

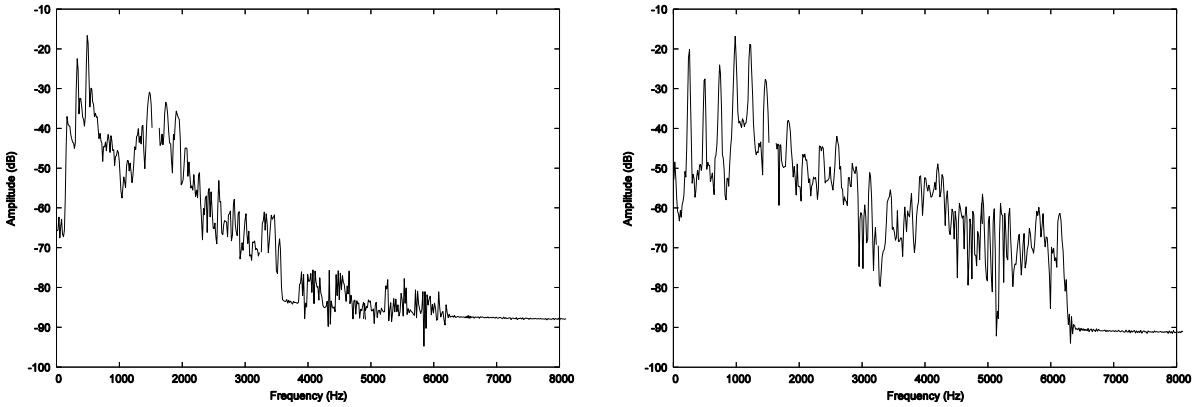
Autormi tohto algoritmu sú páni Ing. Oldřich Plchot a Ing. Petr Schwarz, Ph.D. Využíva dokonalejšiu myšlienku ako metóda I. Takisto ako metóda I je založená na analýze spektra – konkrétne spektrálnej hustoty energie. Využíva fakt, že telefónny kanál sa správa ako pásmová priepusť, ktorá prepustí len frekvencie od 400 Hz do 3,4 kHz. Na opačnej strane bežná wideband reč obsahuje značnú energiu až približne do 5 kHz.[3] Na obrázku 8 je zjednodušený stavový diagram tejto metódy. Pre detekciu je signál najprv prevzorkovaný na vzorkovaciu frekvenciu 16 kHz. Signál je rozdelený na rámce o 512tich vzorkách, bez prekrývania. Teoreticky je možné použiť aj inú dĺžku rámca. Pre každý rámec je spočítaná Fourierova transformácia. Na zistenie hraníc medzi wideband a telefónnym hovorom sa zameriame na frekvenčný rozsah medzi 2350 Hz a 4600 Hz. Využíva sa spektrálna hustota výkonu, ktorá je najprv normalizovaná na nulovú strednú hodnotu. Hodnoty z prvej časti (od 2350 Hz do 3475 Hz) a druhej časti (od 3475 Hz do 4600 Hz) sú zosumované. Vypočíta sa pomer týchto dvoch súm, ktorý je následne prahovaný a podľa výsledku sa spraví rozhodnutie. Ak je pomer väčší ako prah, v nižších frekvenciách je viac energie a považujeme tento segment za telefónny hovor. V opačnom prípade je segment považovaný za wideband signál.



Obrázok 8. : Stavový diagram metódy detekcie II

Bežné zdroje médií, akými sú satelitné alebo internetové rádiá, sú zvyčajne vzorkované na 22 kHz, čiže podporujú spomínané šírky pásma – znamená to, že keď do bežného rozhlasového vysielania vložíme telefónny hovor, uvidíme jasnú zmenu v spektre (viď obrázok 2).

Na tejto metóde je založený program, ktorý som spomínal v kapitole 1. Vyvinuli ho páni Ing. Oldřich Plchot a Ing. Petr Schwarz, Ph.D. Tento program bol použitý v praxi a ako vstupné dáta boli použité rozhlasové nahrávky z rádia Voice of America v rôznych jazykoch, vo formáte MP3 (bitrate 24 kbit/s, vzorkovacia frekvencia 22,050 Hz, bitová hĺbka 16 bitov, mono). Pre použitie týchto dát na trénovanie LRE systému sa muselo vyriešiť viacero problémov – originálne dáta obsahovali veľké množstvo hudby a reči s hudbou na pozadí. Na natrénovanie museli byť vyselektované iba úseky s čistou rečou. V mojom finálnom detektore (kapitola 6) je táto skutočnosť vyriešená segmentovaním nahrávky na rečové a nerečové segmenty a následným zhlukovaním segmentov na jednotlivých hovoriacich. (Proces zhlukovania segmentov podľa podobných vlastností hlasu sa volá diarizácia hovoriaceho.) Ďalším problémom bola nízka variabilita hovoriacich, napríklad v spravodajských programoch, ktoré sú moderované jediným moderátorom – preto museli byť vybrané také časti, kde tento problém nie je a variabilita je dostačujúca. Telefónne hovory vo vysielaní môžu byť prerušované moderátorom a preto nastala potreba rekonštruovať zo segmentov, v ktorých hovorí len osoba cez telefón. Na získanie týchto segmentov bol použitý postprocessing.



Obrázok 9. : Porovnanie spektier telefónneho hovoru (vľavo) a wideband reči (vpravo) z nahrávky zahraničného rádia [3]

Pre účely postprocessingu bol navrhnutý algoritmus, ktorý má na vstupe súbor so značkami (výstup programu na detekciu telefónnych hovorov) a jeho výstupom sú jednotlivé segmenty telefónnych hovorov označené ako `phonecall1`, `phonecall2`, ... Tento algoritmus označí jednotlivé segmenty telefónneho hovoru a snaží sa ich zoskupiť do dlhších celkov. Akceptuje segmenty, ktoré sú dlhšie ako 10 sekúnd, pretože výstup detektoru telefónnych hovorov produkuje krátke segmenty, ktoré niekedy obsahujú aj časti wideband signálu. Výstupným segmentom je priradené to isté poradové číslo len v prípade, že je medzi nimi maximálne 120 sekúnd wideband signálu. Ak je dĺžka wideband signálu medzi telefónnymi hovormi dlhšia ako 240 sekúnd, ďalšie segmenty budú mať nové poradové číslo (napríklad `phonecall3`).

Keď je spracovávaný súbor s pôvodnými segmentmi (výstup detektoru) spôsobom popísaným vyššie, segmenty s rovnakým označením sa rozdeľujú a spájajú. Na určenie pauzy v reči na hraniciach segmentov bol použitý fonémový rozpoznávač používaný na Speech@FIT. Výstupom boli časové značky, ktoré boli použité na vystrihnutie segmentov z pôvodných nahrávok. Následne boli segmenty s rovnakým označením zoskupené do jedného súboru, na získanie rekonštruovaného telefónneho hovoru. Pomocou tohto prístupu bolo získaných menej segmentov s telefónnymi hovormi ako produkoval pôvodný výstup detektoru, ale výhodou bolo, že segmenty obsahovali oveľa menej wideband častí, ktoré boli spôsobené chybnou detekciou. Navyše bola zvýšená aj variabilita hovoriacich, pretože sa získalo menej segmentov s rovnakým hovoriacim.

5.1 Možné vylepšenia detekcie

Nevylepšené spracovanie označkovaných výstupných rámcov detekcie metódou I alebo II je v podstate nepoužiteľné. Napríklad nespracovaný výstup metódy II produkuje segmenty s veľkým množstvom trhlín. Rámce, ktoré sa zoskupia tou najjednoduchšou metódou (pospájaním sekvencie za sebou nasledujúcich rámcov rovnakého kanálu do jedného segmentu) generujú segmenty prislúchajúce buď telefónnemu kanálu alebo wideband reči, avšak rozdelené na malé časti. Akési „zašumenie“ príslušnosti rámcov produkuje krátke segmenty s trhlinami. (Niektoré rámce v danom regióne (určitého kanálu) sú označené inak ako ostatné rámce príslušného regiónu a to rozdelí región na niekoľko malých segmentov.) Príslušnosti rámcov ku kanálu sa v niektorých časových intervaloch môžu aj často striedať a počet rámcov jedného a druhého kanálu môže byť v rovnováhe, a v takej situácii bez použitia dodatočnej metódy detekcie (napríklad metódy I, ktorá by v prípade „neistoty“ metódy II rozhodla a vice versa) je touto metódou takmer nemožné určiť, o aký kanál vlastne ide. Niektoré trhliny vo výsledných segmentoch do príslušného segmentu patria (ide o chybné označenie rámcov), iné nie (napríklad prerušenie telefónneho hovoru prasknutím spadajúcim do wideband kanálu). Tento fakt je ešte výraznejší, ak je prah rozdeľujúci pomer súm nastavený nevhodne – vzniká viac dier. Niektoré nahrávky a nastavenia prahu môžu produkovať také veľké množstvo krátkych segmentov, že po použití postprocessingu popísaného v predchádzajúcej kapitole stratíme veľké množstvo výsledných regiónov, pretože krátke regióny by boli zahodené vzhľadom na to, že táto metóda akceptuje iba segmenty dlhšie ako 10 sekúnd.

Naším cieľom je navrhnúť také vylepšenie algoritmu zaraďujúceho detekované rámce do segmentov (prípadne úplne novú metódu), ktorá by tieto krátke chybne zaevidované segmenty eliminovala a výsledkom by bolo čo najpresnejšie určenie regiónov príslušného kanálu. Takéto riešenie je zaimplementované v mojom finálnom projekte, využíva sa pokročilá segmentácia a diarizácia hovoriaceho.

5.1.1 Vylepšenie „A“ zoskupovania výstupných rámcov metód I a II do regiónov – mediánový filter

Prvým možným vylepšením, ktoré by jednoducho dokázalo odstrániť krátke trhliny vo výsledných segmentoch, je použiť na pole príslušnosti rámcov do určitého kanálu mediánový filter. Jeho efektom by bolo odstránenie krátkych a málo početných rámcov signálu zaevidovaných opačne, ako zvyšok rámcov v nejakej oblasti. Ide väčšinou o chybné označenie takýchto rámcov, ale existujú aj prípady, kde to neplatí, ako popíšem ďalej. Použitie mediánového filtru má svoje nevýhody. Napríklad v prípade, že v nejakej časti nahrávky by sa príslušnosť rámcov k telefónnemu kanálu a wideband kanálu striedali tak, že počet oboch kategórií rámcov by bol približne rovnaký, by medián nepomohol a jeho výsledkom by boli iné, ale tiež krátke segmenty. Toto by ale odstránil postprocessing, (napríklad taký, ako je popísaný v kapitole 5) ktorý krátke regióny eliminuje a na extrakciu príznakov sa použijú iba dlhšie regióny. Prišli by sme však o určitú časť nahrávky. Inou nevýhodou je, že ak by bol telefónny hovor pravidelne prerušovaný krátkym wideband signálom (dĺžky 1-2 rámce) – napríklad praskaním v nahrávke, tak by bol tento zvuk aj napriek tomu považovaný za telefónny hovor – čiže by vznikali false alarms. Pre porovnanie, metóda I by takýto segment vystrihla, ale v prevedení popísanom v kapitole 3 by zároveň odstránila z evidencie telefónnych hovorov ostatné rámce zo skúmaného segmentu (zhluku rámcov). Zvuk prasknutia je pre extrakciu príznakov hlasu nežiaduci, či už chceme detektorom vyfiltrovať telefónne hovory, alebo wideband reč. Tu sa naskytuje myšlienka pridať do programu detektor podobných nežiaducich zvukov, akým je praskanie – akýsi de-click filter, ktorý by rámce s podobnými zvukmi vyradil zo spracovania, ale vždy záleží na tom, aké kvalitné nahrávky chceme spracovať. My potrebujeme spracovať nekvalitné nahrávky, pretože reč, ktorá je v nich obsiahnutá je nedostupná z komerčných databáz. V mojej finálnej verzii detektoru sú segmenty s podobnými zvukmi zaevidované ako Non-Speech, a ak sú dlhšie ako 2 sekundy, užívateľ si môže vybrať, či chce tieto segmenty detekovať na telefónny hovor, alebo ich ignorovať.

5.1.2 Metóda „B“ vytvárania regiónov z výstupu metódy II

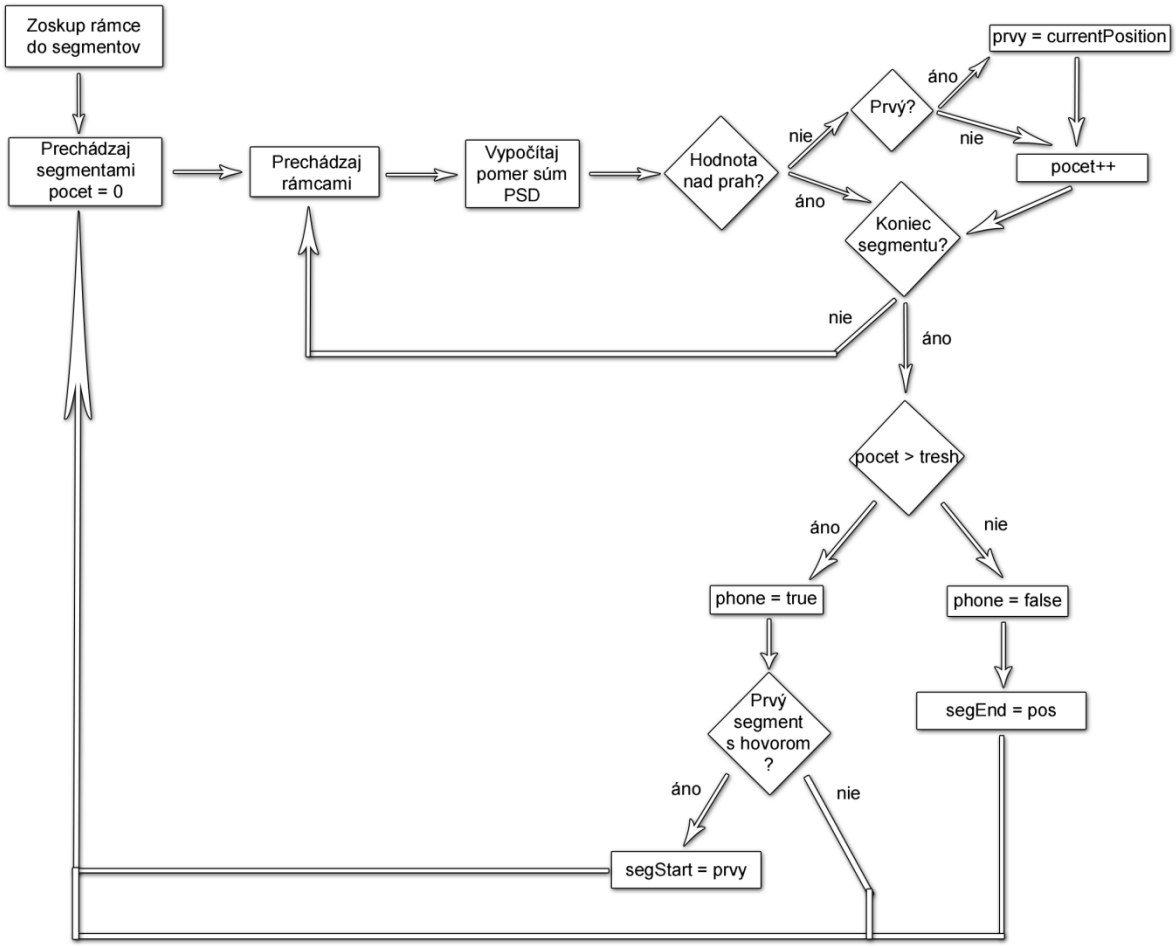
Iným vylepšením by bolo použiť metódu, ktorej princíp spočíva na fakte, že:

(**F1:**) Pri metóde detekcie II je pri telefónnom hovore rozdiel súm PSD v niektorých rámcoch značný (suma z nižších frekvencií je podstatne väčšia ako suma z vyšších frekvencií) a preto je výsledný pomer týchto súm veľmi vysoký (niekedy až 10-násobne oproti wideband). Stane sa tak vždy niekoľko krát v určitom počte skúmaných rámcov telefónneho hovoru.

Princíp tejto metódy by sa dal zhrnúť do piatich bodov:

1. Prah pomeru súm PSD sa zvolí vyššie ako obyčajne (učíme ho vhodným parametrom).
2. Rámce sa zoskupia do segmentov s istým počtom rámcov (počet určí iný parameter).
3. Spočítame počet rámcov v segmente, v ktorých hodnota pomeru súm PSD nepresiahne náš prah (narrowband rámce) a zaznamenáme pozíciu prvého takéhoto rámca.
4. Ak je počet narrowband rámcov v tomto segmente nad určitý (iný) prah, označíme celý segment za telefónny hovor a ak sa jedná o prvý takýto segment, zaznamenáme pozíciu prvého narrowband rámca v tomto segmente.
5. V opačnom prípade označíme segment za wideband reč a pokračujeme ďalším segmentom. (Ak boli predchádzajúce segmenty označené ako telefónny hovor a tento ako wideband, zistíme pozíciu posledného narrowband rámca v predchádzajúcom segmente (prvý už poznáme) a región zapíšeme do kolekcie.)

Na nasledujúcom obrázku je stavový diagram tejto metódy:



Obrázok 9b. : Stavový diagram metódy „B“

Princíp tejto metódy som navrhol po pozorovaní správania sa výsledného pomeru súm PSD druhej metódy detekcie v novej verzii programu na DVB-T nahrávkach. Potenciál tejto metódy som zbežne vyhodnocoval a zistil, že má niekoľko nevýhod: **a)** metóda nedetekuje správne začiatočné a koncové pozície segmentov z rôznych kanálov **b)** s použitím tejto metódy sú niektoré segmenty chybne zaevidované do určitého kanálu a **c)** nie pri všetkých skúmaných nahrávkach bol splnený predpoklad F1, a preto táto metóda nie je obecne použiteľná na všetky typy nahrávok, čo bolo moje záverečné zistenie pre obe metódy B a C a dôvod pre ich zavrhnutie.

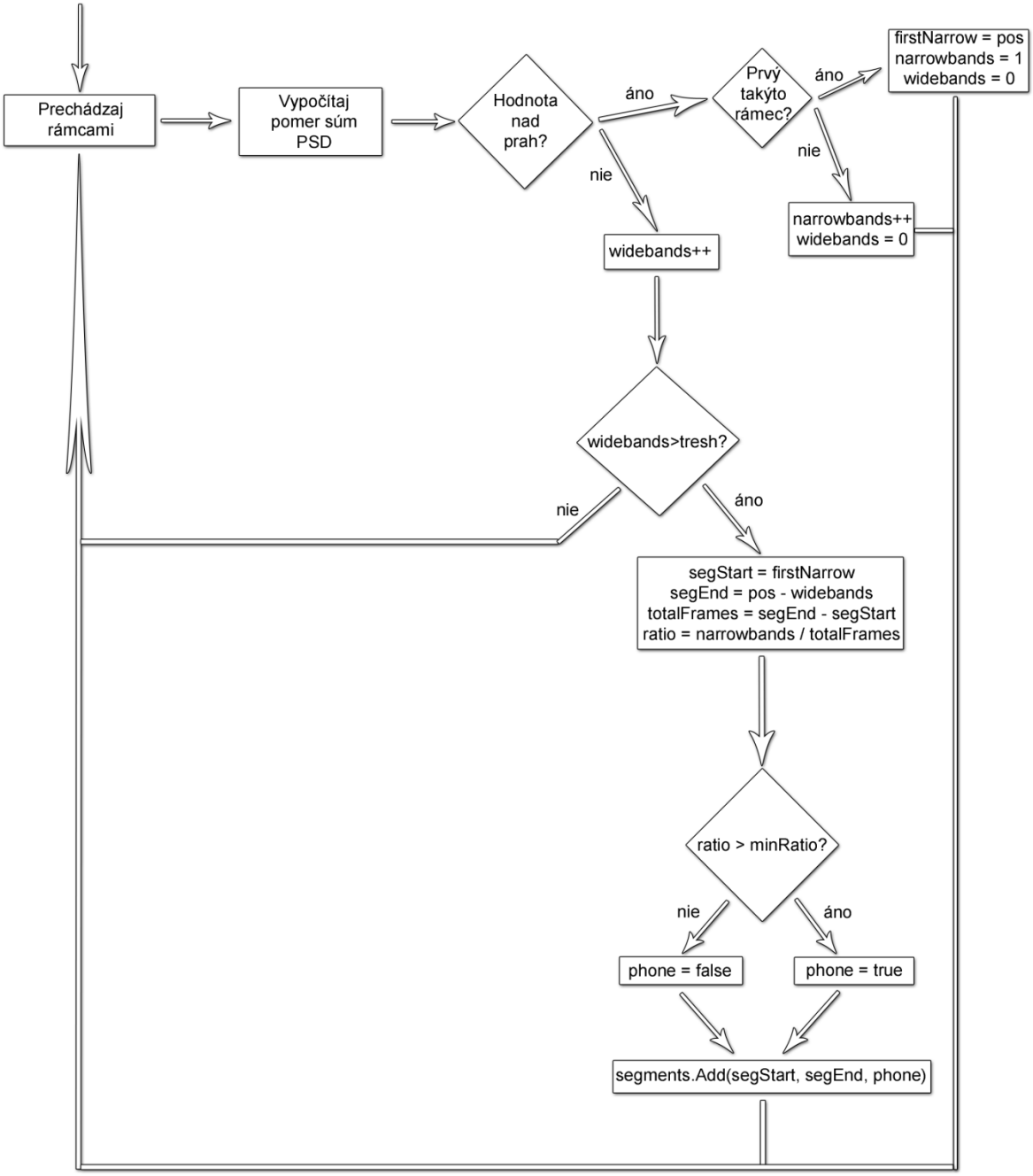
5.1.3 Metóda „C“ vytvárania regiónov z výstupu metódy II

Táto metóda sa snaží odstrániť nedostatky predchádzajúcej metódy. Je navrhnutá tak, aby čo s najväčšou istotou vyberala regióny, v ktorých je pravdepodobnosť telefónneho hovoru vysoká. V dobe navrhovania tejto metódy mi ešte nebola známa skutočnosť, že nie pri všetkých nahrávkach je splnený predpoklad F1 – teda že pri telefónnych hovoroch nedosahovala priemerná hodnota pomeru súm PSD takých vysokých hodnôt, ako na digitálne upravených nahrávkach (s použitím rôznych štúdiových kompresorov, prípadne stereo-imagerov a podobnej techniky v rádiu, na zvýšenie hlasitosti vysielania). Výhodou tejto metódy je fakt, že nerozdeľuje nahrávku na krátke segmenty zložené z viacerých rámcov, ktoré sa potom celé považujú za telefónny hovor alebo wideband a iba v prvom a poslednom rámci sa zistí presná pozícia začiatku a konca daného kanálu.

Princíp je nasledovný:

1. Prah pomeru súm PSD sa zvolí vyššie ako obyčajne (učíme ho vhodným parametrom).
2. Sekvenčne sa prechádza rámcami a sleduje sa, či hodnota pomeru súm PSD prekročí prah.
3. V prípade, že je prah prekročený, a je to prvý takýto rámec, zaznamená sa pozícia tohto rámca, vynuluje sa premenná, v ktorej sumujeme počet takýchto rámcov od výskytu prvého a premenná do ktorej budeme sumovať počet za sebou nasledujúcich wideband rámcov (s pomerom nižším ako prah).
4. Od tejto chvíle sa pri každom prekročení prahu zvýši premenná s počtom takýchto rámcov a vynuluje sa suma za sebou nasledujúcich wideband rámcov (v ktorých hodnota pomeru prah neprekročila).
5. Pri narazení na rámec, ktorý neprekročil prah, zvýšime hodnotu príslušnej sumy.
6. Ak je suma za sebou nasledujúcich wideband rámcov vyššia ako nejaký zvolený prah (určený parametrom), prejde sa na bod 7, inak sa pokračuje bodom 2.
7. Telefónny hovor pravdepodobne skončil, vypočíta sa pozícia posledného rámca prislúchajúceho telefónnemu kanálu (podľa aktuálnej pozície a sumy za sebou nasledujúcich wideband rámcov).
8. Z pozícií prvého a posledného narrowband rámca sa vypočíta celkový počet rámcov detekovaného segmentu.
9. Vypočíta sa, aká je pravdepodobnosť, že detekovaný segment je telefónny hovor (podelí sa počet narrowband rámcov v detekovanom segmente celkovým počtom rámcov tohto segmentu a získa sa pomer).
10. Tento pravdepodobnostný pomer môže byť ďalej buď prahovaný (parametrom) a v prípade prekročenia prahu je segment označený za telefónny hovor, alebo je segment automaticky

považovaný za telefónny hovor s pridruženou percentuálnou informáciou o konfidencii detekcie.



Obrázok 9c. : Stavový diagram metódy „C“

Táto metóda sa po preskúmaní javila ako nádejný kandidát, pretože poskytovala lepšie výsledky ako metóda „B“. Naviac pridávala percentuálnu informáciu o „istote“ detekcie, avšak bola zavrhnutá z dôvodu nesplnenia predpokladu F1 na nahrávkach, ktoré neboli pred odvysielaním digitálne upravené a nahradená metódou ktorá bude popísaná v kapitole 5.1.5.

Metódy „A“, „B“ a „C“ vytvárania regiónov z označených rámcov prislúchajúcich nejakému kanálu by sa teoreticky dali využiť aj pri metóde detekcie I, ktorá by bola upravená tak, že by jednotlivé rámce neboli zoskupené do segmentov, ale u každého rámca by bolo poznačené, či prah prekročil, alebo nie.

5.1.4 Predspracovanie nahrávky s použitím segmentátora a následnou diarizáciou hovoriaceho

Na spresnenie pozicovania detekovaných regiónov s telefónnymi hovormi sa nám ponúka možnosť použiť systém na automatickú segmentáciu na rečové a nerečové (Speech/Non-speech) segmenty a diarizáciu hovoriaceho – čiže určenie, ktorý hovoriaci rozpráva v jednotlivých častiach nahrávky. V nasledujúcej podkapitole popíšem tieto systémy.

5.1.4.1 Segmentácia a zhlukovanie

Segmentačné a zhlukovacie moduly sú súčasťou väčšiny LVCSR[2] systémov. Segmentačný modul je zodpovedný za segmentáciu hlasového vstupu na menšie kusy, ktoré môžu byť rozpoznávacím systémom priamo spracované. Často krát segmentátor filtruje aj nerečové (Non-Speech) časti, akými sú napríklad ticho, mľaskanie, smiech, a dokonca aj pesničky alebo zvukové efekty. Zhlukovací modul má za úlohu zoskupiť do kopy segmenty s podobnými charakteristikami. Obvyklé charakteristiky, podľa ktorých sa zhlukuje signál sú napríklad zvukový kanál (wideband / telefón) alebo pohlavie. Niektoré zhlukovacie systémy, nazývané *Systémy pre diarizáciu hovoriaceho*, sú schopné zhlukovať fragmenty od individuálnych hovoriacich. S použitím informácií o zhlukoch môže rozpoznávač optimálne spracovať každý zhluk. Napríklad, špeciálne modely závislé na pohlaví hovoriaceho môžu byť aplikované keď je dostupná informácia o pohlaví, alebo môžu byť aplikované techniky adaptujúce model pre jednotlivých hovoriacich, ak bol použitý systém na diarizáciu hovoriaceho.

Segmentácia

Bežné zvukové nahrávky nemusia obsahovať iba reč, ale aj rôzne druhy iných zvukov. Tieto nerečové fragmenty, akými sú napríklad šum v pozadí, hudba alebo zvukové efekty musia byť oddelené od rečových segmentov. Bežnou metódou je modelovať reč, a každú akustickú nerečovú udalosť tak, aby sa dali identifikovať pri segmentácii. Tento prístup vyžaduje, aby bol známy druh

[2] Large Vocabulary Continuous Speech Recognition – rozpoznávanie reči s použitím veľkého slovníka

nerečového zvuku, ktorý bol zaznamenaný pri segmentácii, čo často nie je. Táto metóda taktiež vyžaduje, aby trénovacie dáta boli reprezentatívne pre dáta, ktoré majú byť segmentované. Aj keď v nahrávke nie sú prítomné žiadne počuteľné nerečové zvuky, úspešnosť systému významne klesá pri použití štatistických modeloch, ktoré sú natrénované na nevhodných dátach.

Zhlukovanie hovoriacich

Systémy pre zhlukovanie hovoriacich nemajú také problémy, ako segmentačné systémy. Ale pretože hovoriaci nie sú dopredu známi, je ťažké natrénovať pre nich dokonalé štatistické modely. Dáta určené na natrénovanie modelu nebudú vždy dokonale odpovedať dátam, ktoré obsahujú reč skutočných hovoriacich. Dobrým nápadom by bolo mať oddelene dva rôzne akustické modely pre mužský a ženský hlas, ale to by sme dopredu museli vedieť, kto rozpráva na nahrávke, ktorú skúmame.

5.1.4.2 Toolkit Shout

Toolkit *Shout*, s ktorým ma zoznámil môj pán vedúci, je LVCSR systém. Jeho autorom je Marijn Huijbregts, človek ktorý pôsobil svojho času aj na Fakulte Informačních Technologií VUT. Tento toolkit slúži na rozpoznávanie reči s použitím veľkého slovníka. Okrem iného sú súčasťou tohto systému aj moduly na segmentáciu a diarizáciu hovoriaceho. Segmentačný modul filtruje nahrávku na rečové a nerečové segmenty a diarizačný modul zhlukuje časti s podobnými charakteristikami do väčších segmentov, výsledkom čoho je rozdelenie nahrávky na časti podľa hovoriaceho. Autor systému ho použil aj na úlohu automatickej diarizácie v NIST RT 06. Výstupom tohto programu je zoznam segmentov, ktorý by sa dal použiť na presnejšie pozicovanie začiatkov a koncov regiónov s telefónnymi hovormi. Autora sme kontaktovali s prosbou, či by sme mohli jeho toolkit použiť pre naše ciele. S nadšením súhlasil. Mojim plánom bolo zapracovať toto vylepšenie do finálnej verzie detektoru, čo sa podarilo.

Použitie modulov toolkitu Shout pre naše účely

Segmenty sa generujú v dvoch priechodoch. Prvý priechod (program *shout_segment*) generuje zoznam regiónov Speech/Non-Speech, čím sa zo spracovania nahrávky vysielania nejakého rádia vylúči napríklad hudba a ďalej sa pracuje už len so segmentmi, ktoré obsahujú čistú reč.

Vstupom sú:

1. RAW audio súbor (musí byť 16kHz, 16bit, little endian)
2. Súbor SAD s akustickým modelom

3. Názov výstupného RTTM súboru

Segmentácia má svoje nevýhody – niektoré (hlavne digitálne upravované) nahrávky telefónnych hovorov považuje *Shout* za Non-Speech. Tento fakt som vyriešil, ako popíšem v kapitole 6.

V druhom priechode požadované segmenty s informáciou o konkrétnom hovoriacom produkuje nástroj *shout_cluster*.

Jeho vstupom sú:

1. Raw zvukový súbor
2. Vstupný RTTM súbor označujúci segmenty s hlasom / bez hlasu (výstup programu *shout_segment*)
3. Názov výstupného RTTM súboru (súbor s regiónmi).

Použiteľnosť

Dlhú dobu nebolo jasné, do akej miery bude výstup segmentov tohto programu presnejší ako výstup samotnej detekcie. Ukázalo sa však, že nástroj funguje dobre a okrem spoľahlivej a presnej segmentácie priraďuje aj segmentom, v ktorých je čistá reč ID hovoriaceho. Táto informácia však nie je dostatočne presná na to, aby na jej základe boli všetky segmenty s nejakým ID považované za telefónny hovor len pre to, že niektoré segmenty s daným ID telefónnym hovorom sú – ID nebolo vždy priradené správne. Po sérii testov som zistil, že generovanie regiónov funguje najlepšie na dlhých nahrávkach. Vo všeobecnosti platí, že čím dlhšia nahrávka, tým lepšie sú zaevidované jednotlivé výstupné regióny. Nevýhodou je, že sa programu *shout_cluster* niekedy stáva, že nedokáže vystrihnúť regióny, ak je jeden hovoriaci prerušený iným a toto prerušenie je príliš krátke – napríklad jediné slovo. Pre naše účely je toto obzvlášť nevhodné v prípade, že prerušenie je wideband signál v telefónnom hovore, pretože nám vznikajú false alarms a tým pádom by sa nám dostávali do nahrávok obsahujúcich čisto telefónne hovory rámce s wideband signálom, čo je nežiadúce pri trénovaní systémov na rozpoznávanie reči, alebo jazyka hovoriaceho.

5.1.5 Metóda „D“ vytvárania regiónov z výstupu metód I a II s použitím segmentátora

Spoločnou nevýhodou metód „A“, „B“ a „C“ bolo, že nerozlišovali reč od hudby. Hudba síce bola zaevidovaná ako wideband, ale toto by bolo zle v prípade, že by sme na natrénovanie chceli použiť wideband signál – dostala by sa nám tam hudba, čiže ak by sme skúmali wideband reč, vznikali by nám false alarms. Táto metóda kombinuje metódy detekcie I a II, pričom využíva výstup z toolkitu *Shout*, kde vyberá po segmentácii iba rečové segmenty. Ako dobrý nápad sa ukázala možnosť vypočítať z oboch metód percentuálnu pravdepodobnosť, s ktorou daný segment prislúcha telefónnemu pásmu. Tá môže byť následne prahovaná, pričom pre každú metódu po nastavení jej vnútorných parametrov môžeme zvoliť iný prah, a za telefónny hovor považujeme situáciu, kedy obe percentuálne hodnoty metód prekročili naše nastavené prahy. Vo finálnej verzii detektoru sa pomocou GUI dajú zistiť a nastaviť správne prahy a vnútorné hodnoty oboch metód, s ďalšími nastaveniami z príkazového riadku. Pre metódu I sa jej percentuálna pravdepodobnosť telefónneho hovoru vypočíta ako pomer počtu rámcov, ktoré spadajú do telefónneho kanálu k celkovému počtu rámcov v danom segmente.

$$p_{m1} = s_p / s_{total}$$

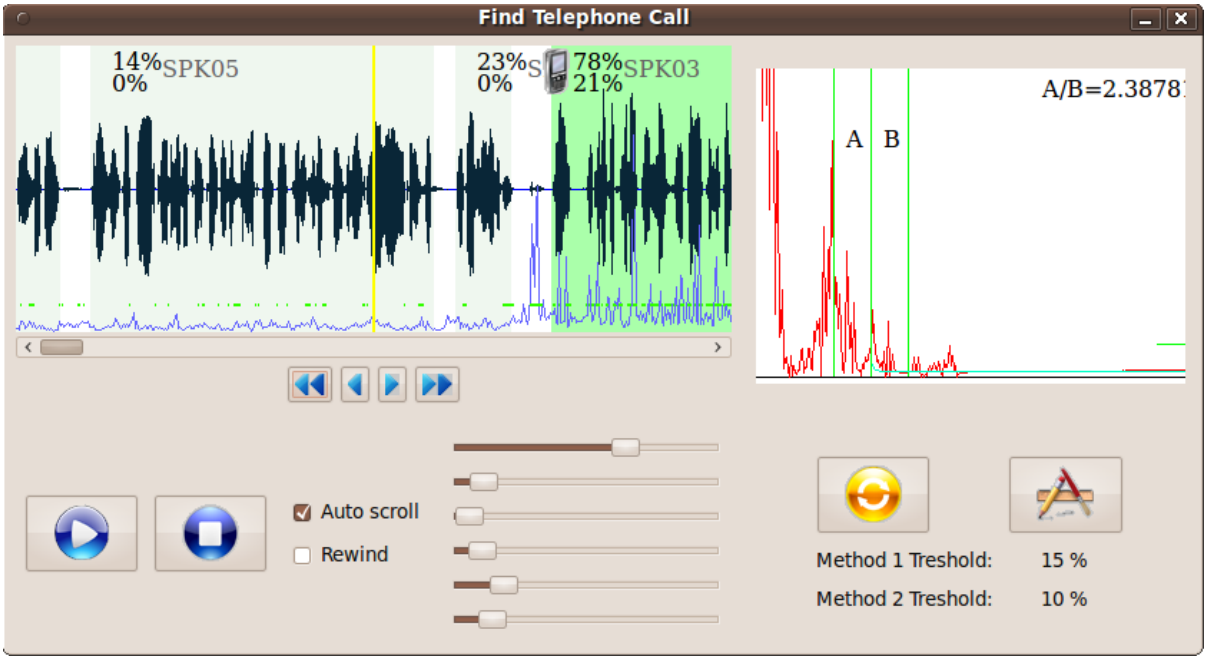
Pre metódu II je princíp obdobný, pravdepodobnosť sa vypočíta ako pomer počtu rámcov, ktoré prekročili vnútorný prah metódy k počtu celkových rámcov v segmente.

$$p_{m2} = s_o / s_{total}$$

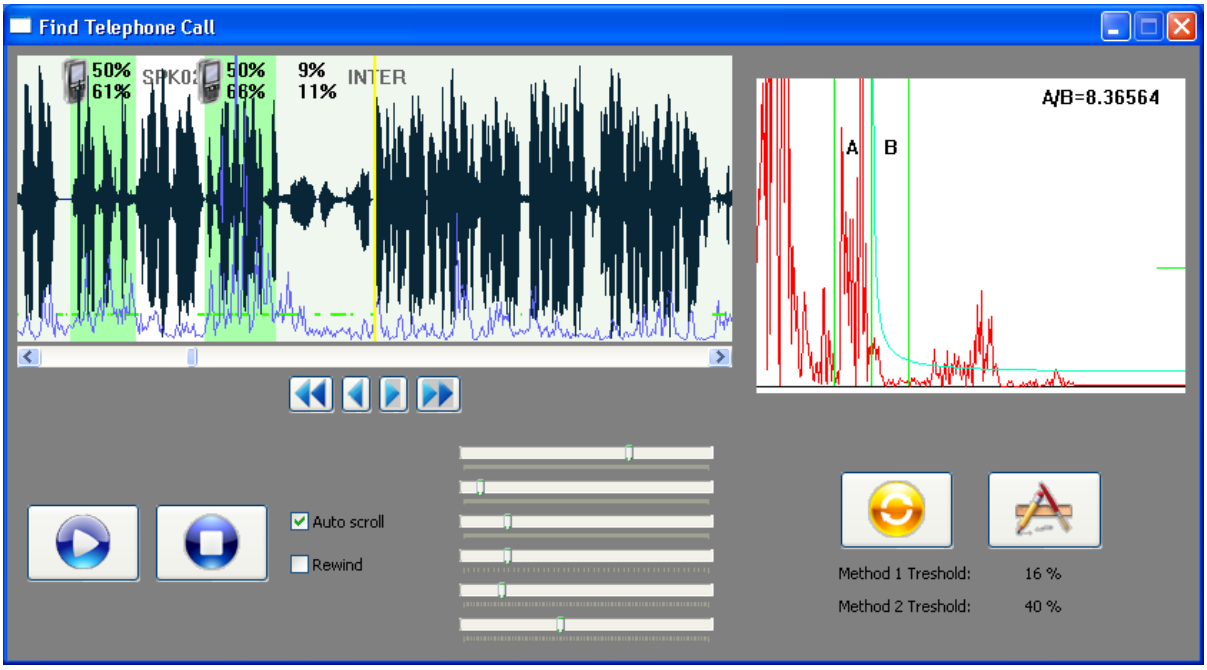
Týmto spôsobom získame pre každý rečový segment dva percentuálne údaje, za každú metódu jeden. Segment považujeme za telefonát, v prípade že oba percentuálne údaje prekročia príslušné prahy.

6 Finálny detektor – projekt FindCall

V súčasnosti som dokončil prácu na tejto novej verzii, ktorá sa dá skompilovať pre platformy Windows a UNIX pomocou kompilátoru **g++**. Kombinuje metódy detekcie I a II a z ich výsledkov získava pravdepodobnosti, s akými sa jedná o telefónny hovor. Program sa dá spustiť s GUI, alebo kompletne ovládať priamo z konzoly a GUI vôbec nezobraziť. Použiť GUI je vhodné na nastavenie vnútorných parametrov obidvoch metód (napríklad na kratšej nahrávke) a potom spustiť analýzu celej nahrávky bez GUI na výkonnejšom počítači, ktorý ani nemusí mať X-Server. V GUI programu je okrem iného zobrazovaný výsledok pomeru súm PSD metódy II, ktorý je znázorňovaný graficky. Ďalej obsahuje spektrálny analyzátor s vykreslením prahu pre metódu I a frekvenčné rozmedzia pre metódu II, vykreslenie zvuku (na aké sme zvyknutí z rôznych programov na editáciu zvuku – napríklad Audacity alebo SoundForge) spolu s vykreslením regiónov s ID hovoriaceho a po analýze aj telefónnych hovorov označených ikonou telefónu. Z názorného zobrazenia výsledkov oboch metód po analýze nahrávky sa dajú ľahko odhadnúť správne hodnoty pre obe metódy detekcie, ktoré môžu byť upravované z GUI a nahrávka reanalyzovaná. Analýza je vďaka použitiu FFTW veľmi rýchla, 10 minútovú nahrávku zanalyzuje program za 2,5 sekundy. Väčšinu času pri spracovávaní nejakej nahrávky zaberá segmentácia a následná diarizácia hovoriaceho pomocou toolkitu *Shout*. Aplikácia je navrhnutá tak, aby sa segmentácia vykonala iba prvý krát a pri opakovanom spustení sa používali segmenty, ktoré vznikli pri prvej segmentácii. Samozrejme, je tu možnosť spustiť segmentáciu opakovane, toto sa dá dosiahnuť použitím príslušného prepínača (**-nr**).

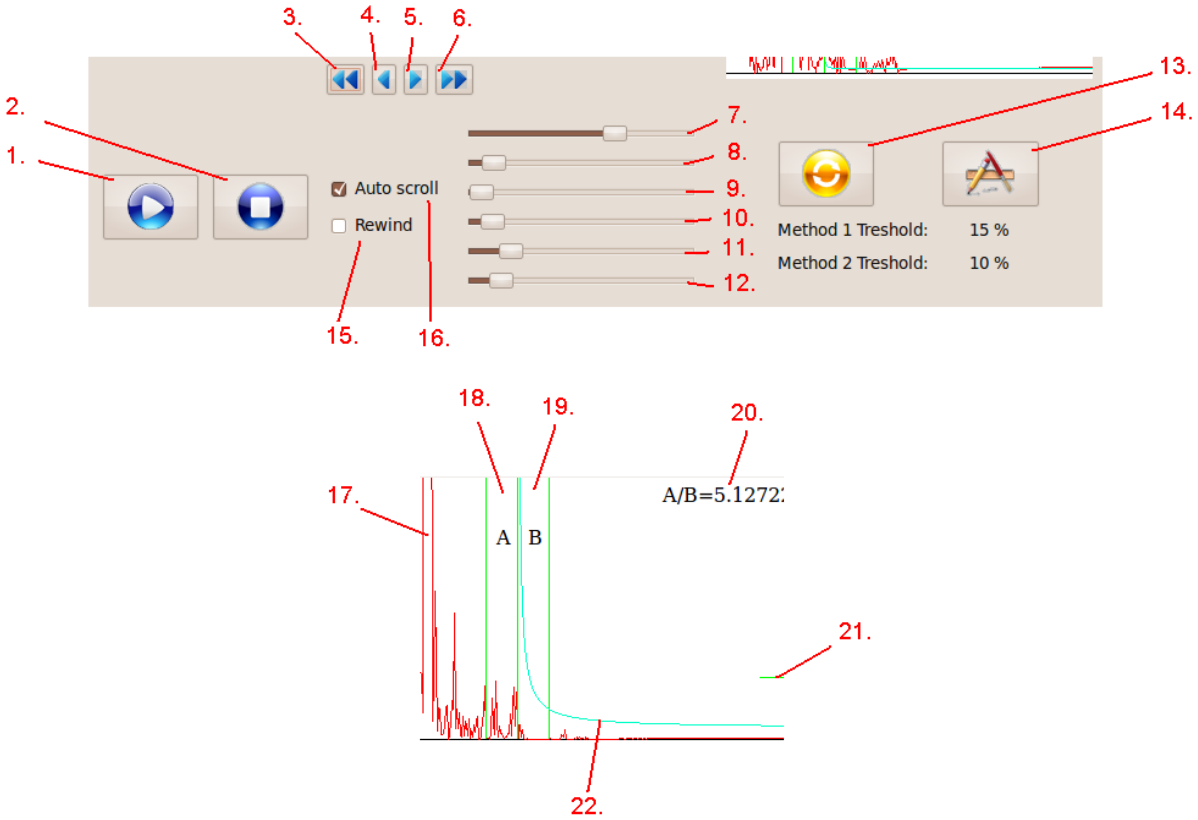


Obrázok 10. : Druhá verzia detektoru bežiaca pod platformou Linux

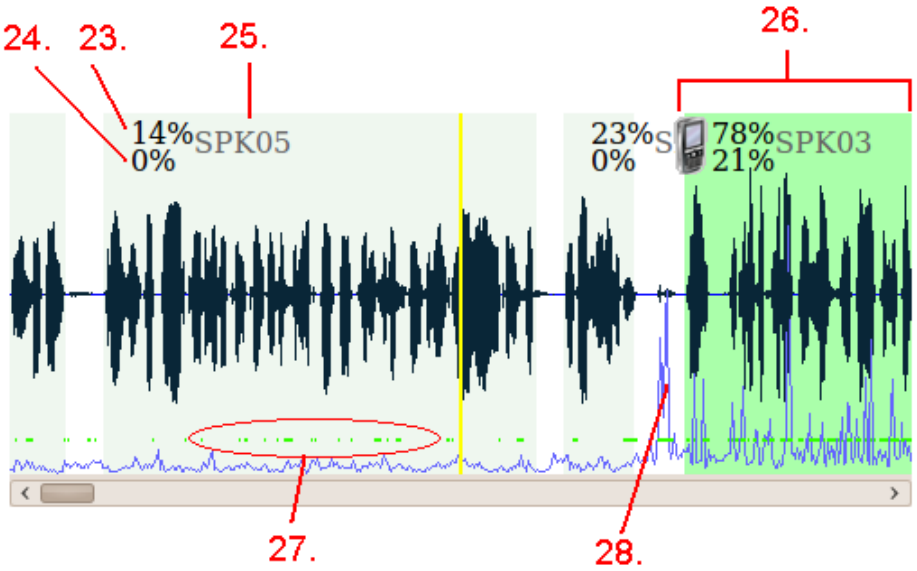


Obrázok 11. : Druhá verzia detektoru bežiaca pod platformou Windows

Ďalej popíšem niektoré časti programu, ktoré sú ukázané na obrázkoch s legendou.



Obrázok 12. : Časti 2. verzie programu



Obrázok 12. : Časti 2. verzie programu

Legenda:

1. Tlačidlo na spustenie prehrávania vstupnej nahrávky
2. Tlačidlo na pozastavenie prehrávania
3. Tlačidlo na posunutie kurzora aktuálnej pozície v nahrávke o 10 sekúnd dozadu
4. Tlačidlo na posunutie kurzora aktuálnej pozície o 1 sekundu dozadu
5. Tlačidlo na posunutie kurzora o 1 sekundu dopredu
6. Tlačidlo na posunutie kurzora o 10 sekúnd dopredu
7. Voľba začiatočnej frekvencie pásma segmentu „A“ pri druhej metóde detekcie (legenda sa zobrazí pri nájdení na posuvník kurzorom myši)
8. Voľba koncovej frekvencie pásma segmentu „B“ pri druhej metóde detekcie
9. Tvar prahu metódy I (softness)
10. Výška prahu metódy I (value)
11. Percentuálna hodnota pre akceptáciu výsledku metódy I za telefónny hovor
12. Percentuálna hodnota pre akceptáciu výsledku metódy II za telefónny hovor
13. Tlačítko pre spustenie analýzy nahrávky
14. Tlačítko pre uloženie výstupného LAB súboru
15. Prepínač automatického posúvania nahrávky pri prehrávaní
16. Prepínač automatického pretáčania na začiatok nahrávky pri zastavení prehrávania
17. Spektrum rámca na aktuálnej pozícii v nahrávke
18. Indikátor rozsahu nižšieho frekvenčného pásma pri metóde II
19. Indikátor rozsahu vyššieho frekvenčného pásma pri metóde II

20. Percentuálny výsledok pomeru sumy PSD nižších frekvencií a sumy PSD vyšších frekvencií (metóda II)
21. Ukazateľ pomeru súm PSD
22. Prah metódy I
23. Percentuálny výsledok metódy I v danom segmente
24. Percentuálny výsledok metódy II v danom segmente
25. ID hovoriaceho
26. Telefónny hovor
27. Indikátor, či prekročil pri metóde I výsledok spektrálnej analýzy prah (zelené znamená že neprekročil)
28. Výsledok pomeru súm PSD metódy II

Celý program je napísaný v jazyku C++. Na vytvorenie GUI bol použitý integrovaný dizajnér vývojového prostredia *wxDevCpp* – na komponenty GUI využíva *wxWidgets*, ktoré je dostupné pod obidvomi platformami. Na výpočet Fourierovej transformácie je rovnako ako v prvej verzii programu použitá knižnica FFTW. Na prehrávanie zvuku som použil knižnicu FMOD (www.fmod.org), ktorá je voľne použiteľná pre nekomerčné účely a jej hlavným prínosom je, že existuje aj verzia pre Linux, a preto som nemusel písať oddelene 2 verzie programu a stačilo vytvoriť len vhodný makefile.

Tvar prahu u metódy I je v tejto verzii detektoru upravený tak, aby sa lepšie prispôsoboval hodnotám výsledku FFT pri telefónnom hovore aj pri nekvalitných nahrávkach (viď obrázok 9 vľavo). Je pridaná hodnota softness, ktorá zaobľuje tvar prahu (viď legenda č. 22). Rámce, v ktorých hodnota výsledku FFT neprekročí tento prah, sú v ľavej hornej časti vyznačené zelenou prerušovanou čiarou (legenda č. 27).

Pri segmentácii a následnej diarizácii hovoriaceho nám vznikajú Non-Speech regióny, ktoré sú väčšinou hudba, alebo ticho. Stáva sa ale, že Shout sa pri niektorých telefónnych hovoroch „pomýli“ a zaeviduje ho ako hudbu – teda Non-Speech. Toto vzniká najčastejšie pri telefónnych hovoroch, ktoré boli nejakým spôsobom v štúdiu rádia upravené, alebo pri veľmi nekvalitných hovoroch. Ja som tento problém vyriešil tým spôsobom, že z Non-Speech regiónov, ktoré sú dlhšie ako 2 sekundy program automaticky vytvára nové segmenty a vznikajú tak tzv. INTER regióny. Pomocou prepínača „**-ni**“ z príkazového riadku sa dá vytváranie týchto regiónov vypnúť. Na týchto regiónoch sa potom vykonáva analýza a detekcia telefónnych hovorov. Z testovania na niekoľkých nahrávkach som zistil, že niekedy je dobré, že sa tieto regióny vytvárajú, pre iné nahrávky je to nepotrebné, až nežiaduce.

Ako som už spomínal, celý program sa dá ovládať z príkazového riadku. Tu je zoznam parametrov:

```
Usage: FindCall [-i <str>] [-o <str>] [-ng] [-nr] [-ni] [-v <num>] [-h <num>] [-s
<num>] [-e <num>] [-m1 <num>] [-m2 <num>]
  -i, --input=<str>              Input .WAV file
  -o, --output=<str>             Output .LAB file
  -ng, --nogui                   No GUI
  -nr, --newregions              Create new regions (cluster)
  -ni, --nointer                 No INTER regions
  -v, --m1v=<num>                Method 1 treshold value (default: 5)
  -h, --m1shp=<num>              Method 1 treshold shape (softness = 1-250.
default: 1)
  -s, --m1s=<num>                Method 2 starting frequency (default: 2350 Hz)
  -e, --m1e=<num>                Method 2 ending frequency (default: 4600 Hz)
  -m1, --m1percentual=<num>      Method 1 percentual acceptation (default: 15)
  -m2, --m2percentual=<num>      Method 2 percentual acceptation (default: 10)
```

Vysvetlenie:

-i <názov> Vstupný súbor (WAV, 44.1 kHz, 16 bit, mono)

-o <názov> Výstupný súbor s regiónmi (LAB, textový, používa sa s **-ng**)

-ng Vypnúť GUI (ovládanie len z príkazového riadku)

-nr Vygeneruje nové regióny (súbor RTTM) toolkitom *Shout*

-ni Negeneruje INTER regióny

-v <číslo> Výška prahu metódy I (východzia hodnota je 5)

-h <číslo> Tvar prahu metódy I – softness (východzia hodnota je 1 - ostrý prah)

-s <číslo> Začiatočná frekvencia rozsahu "A" metódy II (východzia hodnota je 2350 Hz)

-e <číslo> Koncová frekvencia rozsahu "B" metódy II (východzia hodnota je 4600 Hz)

-m1 <číslo> Akceptuj za telefónny hovor, len ak je hotnota výsledku metódy I nad túto hodnotu (východzia hodnota je 15 %)

-m2 <číslo> Akceptuj za telefónny hovor, len ak je hotnota výsledku metódy II nad túto hodnotu (východzia hodnota je 10 %)

7 Výsledky merania presnosti prvej verzie detektoru

V tejto kapitole zhrniem svoje výsledky merania presnosti detekcie prvej verzie programu. Na tento účel som do programu pridal pole ukazujúce pozíciu kurzora v milisekundách a následne som odpísal pozície začiatkov a koncov detekovaných segmentov. Na analýzu som použil celkom 3 nahrávky: prvé dve sú nahraté z DVB-T vysielania jednej českej televízie (*poplatky.wav* a *viza_do_kanady.wav*) a druhá je pomerne nekvalitná nahrávka z afrického rádia (*amha.wav*). Telefónne hovory v nahrávkach som ešte pred porovnávaním s detekovanými segmentmi manuálne anotoval. Tu sú dosiahnuté výsledky:

Nahrávka: *poplatky.wav*

Dĺžka: 1 minúta 41,529 sekundy

Počet telefónnych hovorov: 2

Kódovanie: PCM, 44 kHz, 16 bit, 705 kbps, mono

Segment	Skutočný začiatok	Skutočný koniec	Detekovaný začiatok	Detekovaný koniec	Rozdiel začiatkov	Rozdiel koncov	Počet dier
1.	11,379 s	20,407 s	11,598 s	20,190 s	0,219 s	0,217 s	0
2.	52,767 s	62,600 s	52,691 s	62,468 s	0,076 s	0,132 s	0

Priemerná odchylka všetkých začiatkov a koncov od referenčných pozícií v tejto nahrávke je: **0,161 s**

Nahrávka: *viza_do_kanady.wav*

Dĺžka: 51,710 sekundy

Počet telefónnych hovorov: 1

Kódovanie: PCM, 44 kHz, 16 bit, 705 kbps, mono

Segment	Skutočný začiatok	Skutočný koniec	Detekovaný začiatok	Detekovaný koniec	Rozdiel začiatkov	Rozdiel koncov	Počet dier
1.	34,356 s	39,559 s	34,094 s	38,343 s	0,262 s	1,216 s	0

Priemerná odchylka všetkých začiatkov a koncov od referenčných pozícií v tejto nahrávke je: **0,739 s**

Nahrávka: amha.*wav*

Dĺžka: 1 minúta 40,755 sekundy

Počet telefónnych hovorov: 3

Kódovanie: MP3, 11 kHz, 16 bit, 24 kbps, stereo

Segment	Skutočný začiatok	Skutočný koniec	Detekovaný začiatok	Detekovaný koniec	Rozdiel začiatkov	Rozdiel koncov	Počet dier
1.	11,145 s	22,465 s	11,796 s	22,501 s	0,651 s	0,036 s	4
2.	25,620 s	30,402 s	25,591 s	30,210 s	0,029 s	0,192 s	1
3.	50,224 s	97,814 s	50,134 s	97,504 s	0,09 s	0,31 s	>20

Priemerná odchylka všetkých začiatkov a koncov od referenčných pozícií v tejto nahrávke je **0,218 s**.

Celková priemerná odchylka pozície začiatkov a koncov segmentov oproti referenčným pozíciám vo všetkých nahrávkach je 0,373 sekundy. Pri nekvalitných nahrávkach, akou je posledná analyzovaná, vznikali v segmentoch diery, ktoré v novej verzii programu už nevznikajú, pretože program pracuje s regiónmi, ktoré vznikli pri segmentácii ako s celkami. Toto vyhodnotenie úspešnosti bolo predmetom semestrálneho projektu a nebolo dostačujúce, a preto som si na analýzu výsledkov konečnej verzie programu napísal automatický vyhodnocovací nástroj AutoScore, ktorý vyhodnocuje na úrovni rámcov, pričom každý rámec môže patriť do jednej zo skupín Hit / Miss / False Alarm / Non-Speech.

8 Automatický vyhodnocovací nástroj AutoScore

Tento nástroj poslúži na presné vyhodnotenie úspešnosti detekcie vo finálnej verzii detektoru a umožní nám spočítať frame accuracy. Jeho vstupom sú 2 názvy súborov – prvý je porovnávaný LAB súbor a druhý je ručne anotovaný (referenčný) LAB súbor. Program predpokladá, že vstupný WAV súbor bol vzorkovaný na 44,1 kHz a že dĺžka rámca je 1024 vzorkov. Program je síce napísaný v jazyku C#, ale dá sa skompilovať aj pod platformu UNIX pomocou kompilátoru Mono. Takto vyzerá výstup programu:

```
Čítam porovnávaný .lab súbor "output.lab"
Čítam referenčný .lab súbor "reference.lab"
Inicializujem polia príslušnosti k segmentu.
Generujem polia príslušnosti k segmentu.
Generujem polia príslušnosti k telefónnemu hovoru.
Vyhodnocujem úspešnosť.
----------------------------------------------------------------------
Hits:           21469
Misses:         13
False alarms:   15
Non-Speech:     3262
```

Program porovnáva vstupný LAB súbor (výstup detektoru) s referenčným LAB súborom. Časti nahrávky, ktoré nepatria žiadnemu segmentu sú považované za Non-Speech. Tu je zdrojový kód časti projektu AutoScore, ktorý zaraďuje jednotlivé rámce do kategórií.

```
for (int i = 0; i < compareArray.Count; i++)
{
     if (compareArray[i].output.isValid && compareArray[i].reference.isValid)
     {
          if (compareArray[i].output.isPhone && compareArray[i].reference.isPhone) hits++;

          if (!compareArray[i].output.isPhone && !compareArray[i].reference.isPhone) hits++;

          if (compareArray[i].output.isPhone && !compareArray[i].reference.isPhone)
          {
                falseAlarms++;
                Console.WriteLine("FA: " + i.ToString()); // vypiseme poziciu false alarmu
          }

          if (!compareArray[i].output.isPhone && compareArray[i].reference.isPhone) misses++;
     }
     else
     {
          silences++;
     }
}
```

Ako vidieť zo zdrojového kódu, v prípade že:

- Aktuálny rámec porovnávaného súboru je telefónny hovor a aktuálny rámec referenčného súboru je tiež telefónny hovor, zvýši sa počet zásahov.
- Aktuálny rámec porovnávaného súboru je wideband reč a rámec referenčného súboru je tiež wideband, zvýši sa počet zásahov.
- Aktuálny rámec porovnávaného súboru je telefónny hovor, ale rámec referenčného súboru je wideband, zvýši sa počet false alarms a pre informáciu sa číslo rámca vypíše.
- Aktuálny rámec porovnávaného súboru je wideband, ale rámec referenčného súboru je telefónny hovor, zvýši sa počet misses.

9 Výsledky merania úspešnosti finálnej verzie detektoru

V tejto kapitole zmeriam úspešnosť vyhodnocovania mojej finálnej verzie detektoru. Na tento účel som si napísal dva programy. Prvý je projekt AutoScore, popísaný v predchádzajúcej kapitole. Druhý program je projekt ReferenceCreator, ktorý slúži na vytváranie referenčných súborov. Jeho princíp je, že z ľubovoľného programu na úpravu zvuku do programu odpisujem začiatočné a koncové časy regiónov s telefónnymi hovormi a wideband signálom a program ich automaticky konvertuje na stovky nanosekúnd a spolu s informáciou o tom, či sa jedná o telefónny hovor alebo wideband reč, zapisuje regióny do výstupného LAB súboru. Ten je potom použiteľný ako referenčný súbor s regiónmi pre vstup programu AutoScore.

9.1 Nahrávka amha_c.wav

Táto nahrávka je starou nahrávkou z afrického rádia Voice of America, ktorú mi poskytol môj pán vedúci Ing. Oldřich Plchot. Nahrávka je nekvalitná, ale jedine takéto nahrávky sú dostupné pre niektoré menej známe jazyky. Nahrávka obsahuje 4 telefónne hovory a jej celková dĺžka je 11 minút a 44 sekúnd. Pôvodná verzia nahrávky nebola toolkitom Shout rozsegmentovaná korektne, preto som musel nahrávku pred spracovaním a následnou detekciou normalizovať. Použil som normalizáciu RMS na -10 dB, čo je štandardná hodnota pre reč.

Konfigurácia detektoru:

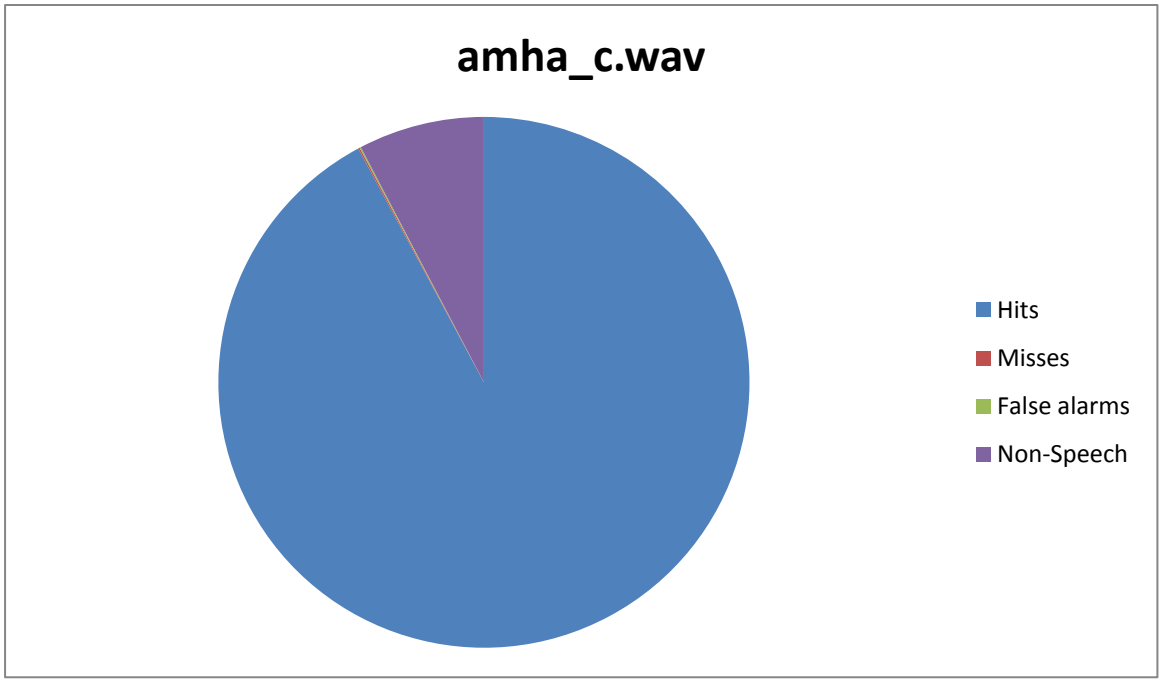
Prah akceptácie metódy I: 50%

Prah akceptácie metódy II: 20%

Ostatné parametre ostali na štandardných hodnotách.

Výsledky úspešnosti detekcie:

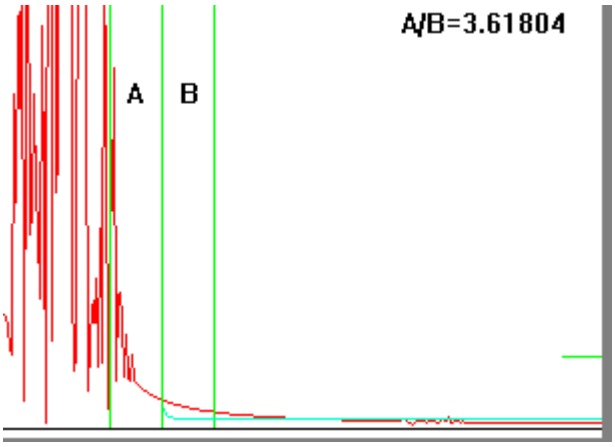
```
Hits:           27956
Misses:         29
False alarms:   24
Non-Speech:     2329
Frame accuracy: 99,811%
```



Obrázok 13. : Graf úspešnosti nahrávky amha_c.wav

Záver:

Detekcia je presná, vyskytlo sa 24 rámcov false alarms. Toto číslo je zanedbateľné, ale false alarms sú oveľa viac nežiadúce, ako napríklad misses, pretože keby sme výstup chceli použiť na natrénovanie rozpoznávača jazyka, dostávali by sa nám tam wideband rámce. Keď to prepočítame, 24 rámcov false alarms je 0,56 sekundy. To je 0,0008 % z nahrávky. Všetky 4 telefónne hovory patria do INTER regiónov, čo znamená, že toolkit *Shout* považoval tieto telefónne hovory za Non-Speech, ale detektor na týchto miestach vytvoril umelé regióny, na ktorých bola vykonaná detekcia. Dĺžka prvého telefónneho hovoru v nahrávke je 215 sekúnd a výsledky metód na tomto telefónnom hovore sú nasledovné: Metóda I = 65%, Metóda II = 42%. Tieto výsledné hodnoty pre tento segment s telefónnym hovorom sú nad naše nastavené prahy pre akceptáciu segmentu za telefónny hovor, takže detektor o ňom prehlásil, že to telefónny hovor je. Hodnoty sú oproti ostatným telefónnym hovorom v nahrávke nízke, hlavne kvôli tomu, že spektrum za deliacou frekvenciou pre telefónne hovory sa dvíha (viď obrázok 14). Pre tento telefónny hovor by pomohlo zaoblenie prahu metódy I, to by však mohlo mať za následok nesprávne zaevidovanie ostatných telefónnych hovorov v nahrávke. **Oproti prvej verzii detektoru pre platformu Windows a verzii od pána Ing. Oldřicha Plchota sú výsledky neporovnateľne lepšie**, s použitím kombinácie obidvoch metód a segmentáciou sa dosiahlo výrazné zlepšenie presnosti detekcie.



Obrázok 14. : Spektrum v nahrávke amha_c.wav sa pri telefónnom hovore dvíha nad prah metódy I

9.2 Nahrávka hydepark.wav

Hydepark.wav je nahrávka z internetového vysielania televízie ČT24, nahraná zo streamu v strednej kvalite. Je to relácia, kde ľudia môžu telefonovať a pýtať sa hostí relácie otázky. Táto nahrávka obsahuje 2 telefónne hovory. Celková dĺžka nahrávky je 18 minút a 33 sekúnd.

Konfigurácia detektoru:

Prah akceptácie metódy I: 70%

Prah akceptácie metódy II: 14%

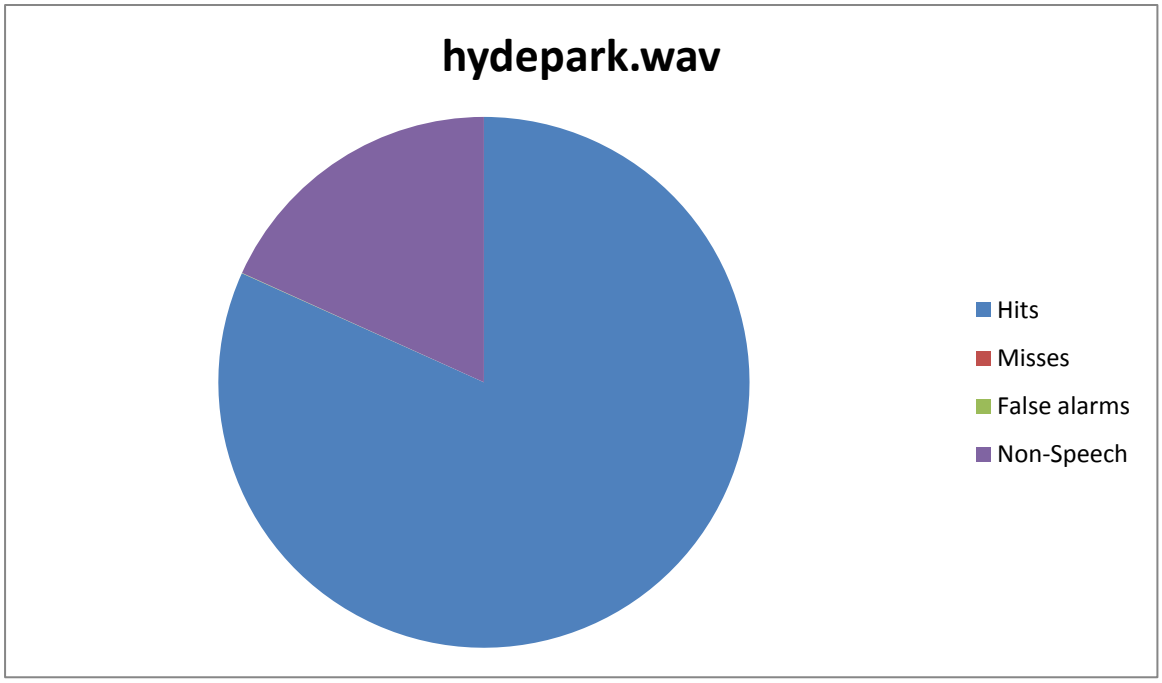
Výška prahu metódy I: 20

Tvar prahu metódy I (softness): 50

Ostatné parametre ostali na štandardných hodnotách.

Výsledky úspešnosti detekcie:

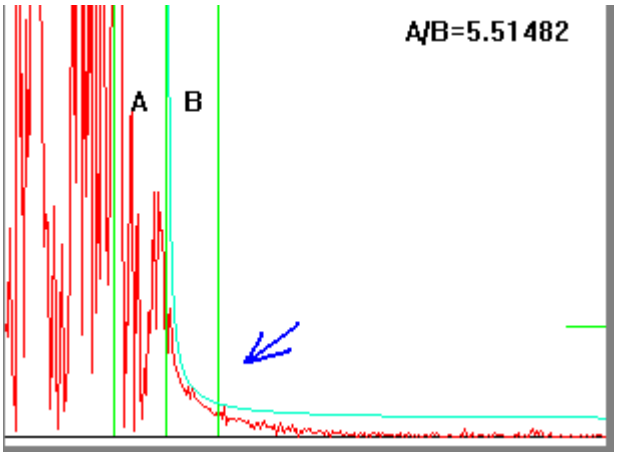
```
Hits:           39202
Misses:         0
False alarms:   11
Non-Speech:     8750
Frame accuracy: 99,971%
```



Obrázok 15. : Graf úspešnosti nahrávky hydepark.wav

Záver:

Priaznivými výsledkami vyhodnotenia úspešnosti na tejto nahrávke som bol milo prekvapený, musím ale priznať, že som parametre detektora pre túto nahrávku ladil dovtedy, až boli všetky segmenty zaradené správne. Ukázalo sa, že zaoblenie prahu metódy I je dobrý nápad, ako vidno na obrázku 16. Bez tohto vylepšenia by nebola detekcia taká presná. Nakoľko potrebujeme, aby pri telefónnom hovore výsledok spektra neprekročil prah metódy I a spektrum sa nedrží pri nulových hodnotách, ale takisto ako v prvom telefónnom hovore v nahrávke amha_c.wav sa smerom k deliacej frekvencii pre telefónny hovor dvíha, potrebujeme upravený prah (viď obr. 16).



Obrázok 16. : Spektrum nahrávky hydepark.wav pri telefónnom hovore a nastavenie prahu pre metódu I (tyrkysovým)

9.3 Nahrávka soma_b.wav

Ďalšia nahrávka z rádia Voice of America, tentokrát v somálštine. Dĺžka nahrávky je 22 minút a 29 sekúnd a celkovo obsahuje 17 telefónnych hovorov. Pre lepšiu segmentáciu pomocou *Shout* bola aj táto nahrávka normalizovaná s RMS na -10 dB.

Konfigurácia detektoru:

Prah akceptácie metódy I: 55%

Prah akceptácie metódy II: 38%

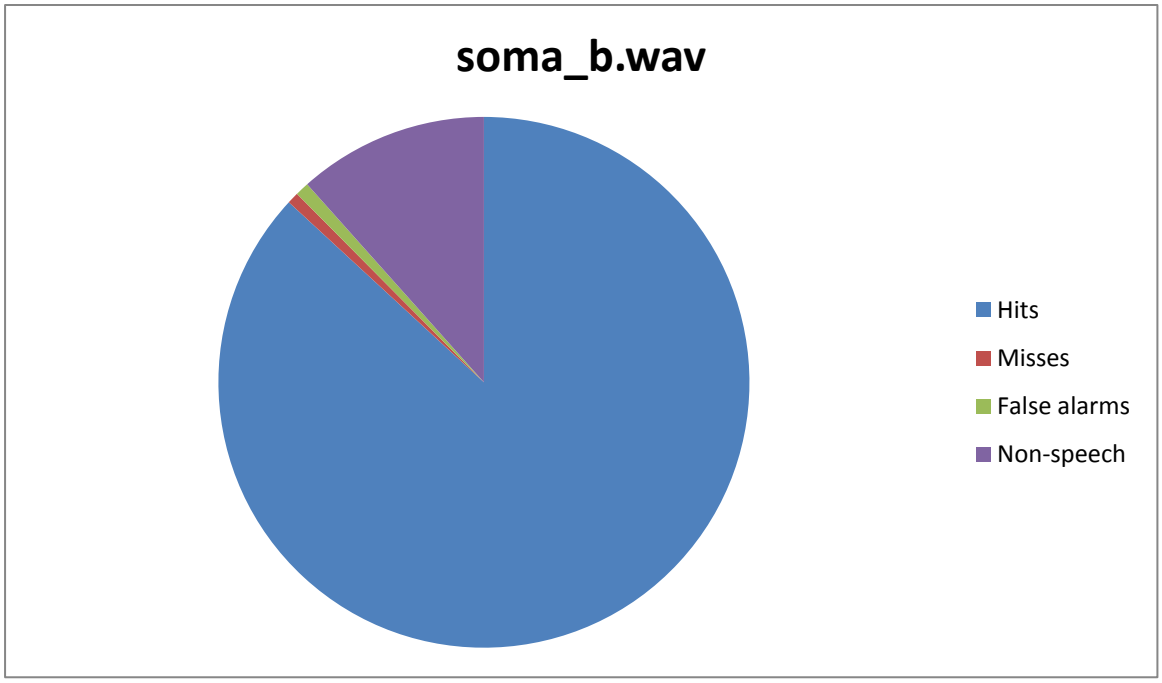
Výška prahu metódy I: 5

Tvar prahu metódy I (softness): 33

Ostatné parametre ostali na štandardných hodnotách.

Výsledky úspešnosti detekcie:

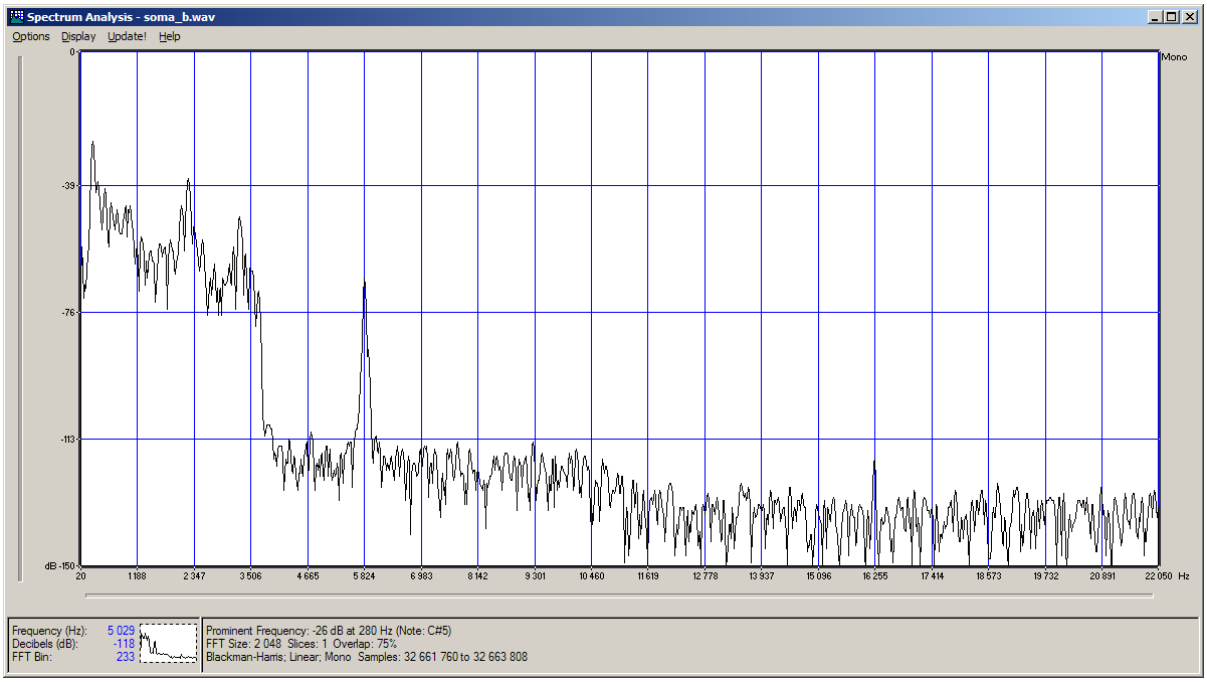
```
Hits:           50480
Misses:         408
False alarms:   488
Non-Speech:     6741
Frame accuracy: 98,225%
```



Obrázok 17. : Graf úspešnosti nahrávky soma_b.wav

Záver:

Na tejto nahrávke sa ukázalo, že ani finálna verzia detektoru nemusí na všetkých nahrávkach dávať stopercentné výsledky. Parametre detekcie som určil empiricky. Je možné, že po jemnom doladení parametrov by boli výsledky lepšie. V nahrávke sa vyskytol krátky úsek telefónneho hovoru, o ktorom detektor povedal, že je wideband signál, ale toto bolo spôsobené tým, že v tomto telefónnom hovore píska tón na frekvencii 5824 Hz. (Obr. 18)



Obrázok 18. : Nežiadúci tón na frekvencii 5824 Hz pri jednom z telefónnych hovorov v nahrávke soma_b.wav

Prevažná väčšina false alarms v tejto nahrávke bola spôsobená tým, že referenčné regióny telefónnych hovorov začínajú až vo chvíli, keď osoba na telefóne začne rozprávať, ale program niektoré INTER regióny zaevidoval aj s úvodným tichom. Keď sa uvažuje úvodné ticho na začiatku telefónneho hovoru a na jeho konci, spolu asi 2 sekundy ticha (ale uvažujeme len polovicu kôli spôsobu značkovania v referenčnom súbore), vynásobí sa to počtom telefónnych hovorov – dostaneme číslo, ktoré približne odpovedá nášmu množstvu false alarms. V nahrávke som ináč žiadny zle zaevidovaný telefónny hovor okrem už spomínaného nezaevidoval.

9.4 Nahrávka poplatky.wav

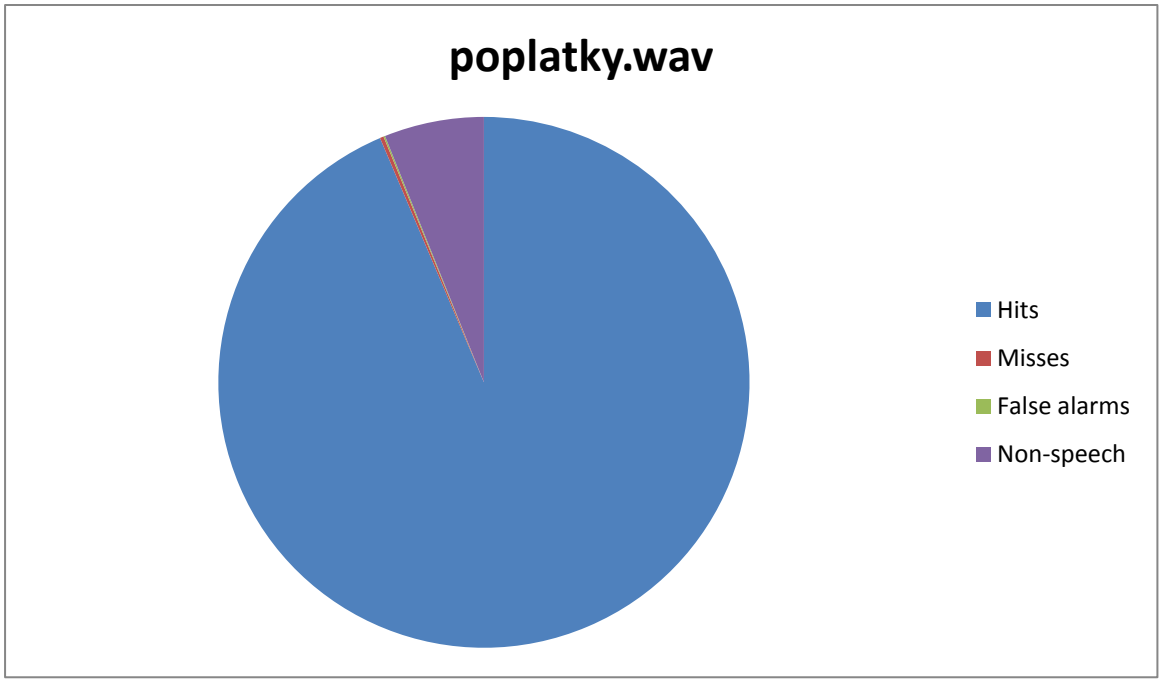
Túto nahrávku som zaradil do vyhodnotenia len pre dôkaz, že na kvalitných vstupných dátach je detekcia presná. Nahrávku nebolo treba normalizovať, pretože zvukový signál bol ešte pred odvysielaním počuteľne spracovaný pomocou štúdiového kompresoru. Nahrávka pochádza z DVB-T vysielania, je dlhá 41 sekúnd a obsahuje 2 telefónne hovory.

Konfigurácia detektoru:

Všetky parametre ostali na štandardných hodnotách.

Výsledky úspešnosti detekcie:

```
Hits:           4093
Misses:         10
False alarms:   4
Non-Speech:     266
Frame accuracy: 99,658%
```



Obrázok 19. : Graf úspešnosti vyhodnotenia nahrávky poplatky.wav

Záver:

Výsledky sú dobré, na kvalitných nahrávkach je detekcia oveľa menej chybová, ako na nahrávkach nekvalitných. Detektor zvládol túto nahrávku zaevidovať úspešne aj s použitím prednastavených hodnôt. Segmentácia je tiež presná, je vidieť, že akustický model ktorý používam na segmentáciu je pre český jazyk vhodný.

10 Záver

V tejto práci som zhrnul niekoľko možných prístupov na detekciu telefónnych hovorov v zvukových dátach. Popísal som 2 rôzne metódy, ktorými sa dajú telefónne hovory detekovať. Prvú metódu som vymyslel sám, a navrhol som niekoľko možných vylepšení detekcie oboch metód. Jednoduchšia metóda I hľadá v spektre časti nahrávky bez vysokých frekvencií, metóda II počíta pomer sumy dvoch častí spektrálnej hustoty energie, ktorý prahuje. Výsledkom práce je hotová prvá verzia programu, ktorá používa jednoduchšiu metódu I detekcie a druhá, finálna verzia detektoru, ktorá kombinuje metódy I a II a využíva toolkit *Shout* na presnejšie určenie segmentov s rečou a zároveň filtruje nerečové segmenty a pre detekciu vyberá čistú reč. Táto verzia je upravená tak, aby ju bolo možné spustiť na platformách Windows aj Linux. Spolu s GUI a ovládaním z príkazového riadku sa z tohto programu stáva nástroj využiteľný pre praktické účely, napríklad na vyselektovanie telefónnych hovorov alebo priamej reči pre natrénovanie systémov na rozpoznávanie jazyka alebo hovoriaceho. Vďaka multiplatformovovej povahe aplikácie by nebolo zložité začleniť tento detektor do rôznych iných systémov, napríklad na predspracovanie dát. Úspešnosť detektoru som preskúmal a zhrnul výsledky, ktoré sú priaznivé. Oproti verzii pôvodného detektoru od pána Ing. Oldřicha Plchota a mojej prvej verzie pre platformu Windows som dosiahol v druhej, finálnej verzii detektoru výrazné zlepšenie úspešnosti detekcie s použitím kombinácie metódy II s metódou I, ktorá zahŕňa niekoľko vylepšení, akými je napríklad zaoblenie prahu metódy I, vylúčenie tichých segmentov a segmentácia. Na každej nahrávke dávajú tieto dve použité metódy iné výsledky, závisí to hlavne od nastavenia vnútorných hodnôt metód. Na nekvalitných nahrávkach je percentuálny pomer počtu rámcov prislúchajúcich telefónnemu hovoru k celkovému počtu rámcov v segmente ťažko určiteľného telefónneho hovoru asi 65% pre metódu I a 42% pre metódu II. Pri kvalitných nahrávkach sú výsledky metód v telefónnom hovore asi 97% pre metódu I a 58% pre metódu II. Preto treba akceptačné prahy pre obe metódy nastaviť pre každý zdroj nahrávky zvlášť. Detekcia je výpočetne nenáročná, dosahuje rýchlosti asi 310-krát realtime. Frame accuracy na kvalitných nahrávkach je asi 99,5 % a na nekvalitných približne 98,2 %. Diarizácia nahrávky na segmenty od jednotlivých hovoriacich funguje dobre, ale môže sa stať, že vo vnútri telefónneho hovoru sa vyskytne zle zaevidovaný krátky úsek wideband signálu. V tomto vidím budúcu prácu na projekte, minimalizáciu takýchto chýb s pokročilými algoritmami, delenie segmentov ktoré produkuje Shout, ak je zrejmé, že telefónny hovor je prerušený priamou rečou, použitie špecializovaných akustických modelov pre vstup segmentácie, alebo napríklad dynamickú úpravu hodnôt prahov metód podľa štatistických informácií z jednotlivých segmentov.

Literatura

[1] Telefón, Wikipedia (http://sk.wikipedia.org/wiki/Telef%C3%B3n)

[2] O. Plchot, V. Hubeika, L. Burget, P. Schwarz, P. Matějka: Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition, Speech@FIT, Brno University of Technology

[3] O. Plchot, V. Hubeika, L. Burget, P. Schwarz, P. Matějka, J. Černocký: Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition, Technnical Report, Speech@FIT, Brno University of Technology

[4] Spraak-Herkennings-Onderzoek Universiteit Twente: Shout, Marijn Huijbregts (http://wwwhome.cs.utwente.nl/~huijbreg/shout/index.html)

[5] Štandardné metódy rozpoznávania hlasu, Michal Gaspierik, 2009 (http://keeper.blog.matfyz.sk/p15244-standardne-metody-rozpoznavania-hlasu)

[6] Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled, Marijn Huijgbregts, ISBN 978-90-365-2712-5

[7] Notes on implementation strategies for a Mono-based WPF, The Mono Project, 2009 (http://www.mono-project.com/WPF)

[8] Biometrické systémy založené na srovnání hlasu, Tomáš Farkas, Masarykova univerzita Fakulta informatiky, 2007 (http://is.muni.cz/th/60398/fi_m/diplomka.txt)

[9] Telephone Line Characteristics, Biblioteca Online (http://www.rigacci.org/docs/biblio/online/intro_to_networking/c3869.htm)

Seznam příloh

Příloha 1. DVD so zdrojovými kódmi obidvoch verzií detektorov a vstupnými nahrávkami