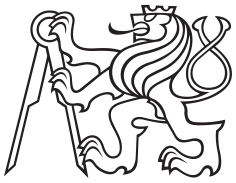
Bakalářská práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra teorie obvodů

Implementace kepstrálního detektoru řečové aktivity při výpočtu řečových příznaků

Michal Kosek

Vedoucí práce: Ing. Petr Mizera
Obor: Komunikace a elektronika
Studijní program: Komunikace, multimédia a elektronika
Leden 2018

ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Kosek** Jméno: **Michal** Osobní číslo: **305137**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra teorie obvodů**

Studijní program: **Komunikace, multimédia a elektronika**

Studijní obor: **Komunikace a elektronika**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Implementace kepstrálního detektoru řečové aktivity při výpočtu řečových příznaků

Název bakalářské práce anglicky:

Implementation of Cepstral Voice Activity Detector within Speech Feature Computation

Pokyny pro vypracování:

1) Seznamte se s algoritmy zpracování řečového signálu s užším zaměřením na detekci řečové aktivity.
2) Implementujte detektory na bázi energie/ MFCC/PLP kepstra s pevným a adaptivním prahem.
3) Integrujte detektory do nástroje CtuCopy pro výpočet řečových příznaků.
4) Analyzujte chování imlementovaných detektorů v různých akustických podminkách

Seznam doporučené literatury:

[1] J. Uhlir, P. Sovka, P. Pollak, V. Hanzl, R. Cmejla: Technologie hlasovych komunikaci. Nakladatelstvi CVUT, 2007.
[2] J. Psutka, L. Muller, J. Matousek, V. Radova: Mluvime s pocitacem cesky. Academia, 2006.
[3] M. Virius: Od C k C++. Kopp Ceske Budejovice 2000. ISBN 80-7232-110-2.
[4] B. Michal, P. Mizera, P. Pollák: Noise and Channel Normalized Cepstral Features for Far-Speech Recognition 2013, In Speech and Computer. LNCS Springer, p. 241-248
[5] P. Fousek, P. Pollak, 'Additive Noise and Channel Distortion-Robust Parametrization Tool - Performance Evaluation on Aurora 2 & 3', Eurospeech'03, Geneva, 2003.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

Ing. Petr Mizera, katedra teorie obvodů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **08.02.2017** Termín odevzdání bakalářské práce: 9.1.2018

Platnost zadání bakalářské práce: **30.09.2018**

[redacted] [redacted] [redacted]

Podpis vedoucí(ho) práce Podpis vedoucí(ho) ústavu/katedry Podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

21.12.2017 [redacted]

Datum převzetí zadání Podpis studenta

Poděkování

Děkuji vedoucímu mé bakalářské práce, Ing. Petru Mizerovi, za jeho aktivní a podporující vedení. Dále děkuji panu Doc. Petru Pollákovi za cenné konzultace a svojí rodině a přátelům za podporu při studiu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, dne 8. ledna 2018

Abstrakt

Cílem této bakalářské práce bylo vytvořit systém pro detekci přítomnosti řeči v diskrétním signálu. Vytvořený detektor používá krátkodobou energii signálu a změny v kepstrálních charakteristikách signálu jako kriteriální funkci. Pro rozhodnutí o přítomnosti řeči je použito několik různých heuristických metod stanovení prahové hodnoty kriteriální funkce. Implemetace detektoru byla provedena v programovacím jazyce C++ a při běžné výpočetní výkonnosti zvolené platformy je detektor schopen operovat v reálném čase. Detektor byl zaintegrován do softwarového nástroje CtuCopy jako interní funkční modul. V experimentální části bakalářské práce bylo provedeno testování funkčnosti detektoru a zkoumáno chování vytvořeného detektoru v různých akustických prostředích s různou úrovní šumu.

Klíčová slova: Detektor řečové aktivity, VAD, Výpočet řečových příznaků, Kepstrální analýza, Kepstrální vzdálenost

Vedoucí práce: Ing. Petr Mizera
Katedra teorie obvodů
ČVUT v Praze
Technická 2
166 27 Praha

Abstract

The aim of this bachelor thesis is to create a system for detection of human speech presence in a discrete signal. The created Voice Activity Detector (VAD) uses computation of short-time signal energy and cepstral distance as a criterion value. Several different methods of heuristic thresholding are used for decision making about speech and non-speech activity in current short-time signal segment. The implementation of the created VAD was written in the C++ programming language, allowing the detector to be capable of real-time operation at the average processing performance of a chosen platform. The detector has been integrated into the CtuCopy speech processing tool as an internal functional module. In experimental part of the thesis, VAD functionality has been tested and it's behavior in different acoustic conditions with different noise levels was studied.

Keywords: Voice Activity Detector, VAD, Speech Feature Computation, Cepstral analysis, Cepstral Distance

Title translation: Implementation of Cepstral Voice Activity Detector within Speech Feature Computation

Obsah

Obrázky

Tabulky

Seznam použitých zkratek

A/D	analogově-digitální převodník
AR	autoregresní model
DCT	diskrétní kosínová transformace
DFT	diskrétní Fourierova transformace
FAR	relativná chyba vyhodnocení neřečových segmentů (False Alarm Rate)
FFT	algoritmus rychlé Fourierovy transformace
FIR	filtr s konečnou impulzní odezvou
GMM	model Gaussových hustotních směsí (Gaussian Mixture Model)
HTER	kombinovaná relativní chyba detekce (Half Total Error Rate)
HTK	standardní formát souboru pro ukládání řečových příznaků
iFFT	algoritmus rychlé inverzní Fourierovy transformace
LPC	lineární prediktivní kódování (Linear Predictive Coding)
MFCC	mel-kepstrum (Mel-Frequency Cepstral Coefficients)
MR	relativná chyba vyhodnocení řečových segmentů (Miss Rate)
OLA	metoda sčítání přesahů (Overlap-Add)
PCM	pulzně-kódová modulace
PLP	řečová parametrizace PLP (Perceptual Linear Predictive Coding)
SNR	odstup signálu od šumu (Signal-to-Noise Ratio)
SS	metoda spektrální odečítání (Spectral Substraction)
VAD	detektor řečové aktivity (Voice Activity Detector)
ZCR	průměrný relativní počet průchodů nulou (Zero-Crosing Rate)

Kapitola 1

Úvod

Hlasové technologie se postupně stávají nedílnou součástí našich každodenních životů a usnadňují komunikaci mezi strojem a člověkem. Tyto technologie nalézají uplatnění v různých oblastech. V call centrech pomáhají zkvalitnit péči o zákazníky (hlasové dialogové systémy), umožňují automatické titulkování pořadů, právníkům či lékařům usnadňují práci při diktování zpráv, dále umožňují hlasové ovládání různých zařízení (GPS navigace v automobilu) a hlasová syntéza zase pomáhá zejména zrakově postiženým lidem při automatickém čtení emailů, SMS a knih.

Základem výše uvedených systémů jsou moduly rozpoznáváni řeči, mluvčího, jazyka nebo diarizace mluvčího. Na vstup těchto modulů není přímo přiváděn řečový signál, ale krátkodobé řečové příznaky, které jsou z řečového signálu extrahovány. Dále jsou v reálných aplikacích na vstup modulů přiváděny pouze příznaky z částí signálu, kde byla detekována řečová aktivita. Tuto úlohu zajišťuje modul detekce řečové aktivity (VAD, Voice Activity Detector). VAD modul je klíčový pro úlohy rozpoznávání mluvčího/jazyka nebo úlohu diarizace a pomáhá snížit dobu zpracování (RTF, Real Time Factor) při dekódování řeči (nedékodují se části s tichem/šumem).

Cílem této bakalářské práce bylo vytvořit detektor řečové aktivity, který rozhoduje o přítomnosti řeči ve zpracovávaném signálu na základě kepstrální analýzy. Jako kriteriální funkce je použita kepstrální vzdálenost kepstra aktuálního segmentu a odhadu průměrného kepstra pozadí. Na jejím základě je možné provádět rozhodování o přítomnosti řeči pomocí vhodného prahu. Pro samotné rozhodování mezi řečovými a neřečovými segmenty byly implementovány dvě různé heuristické metody stanovení prahové hodnoty.

Vytvořený detektor byl napsán v programovacím jazyce C++ a byl zaintegrován do existujícího softwarového nástroje CtuCopy jako funkční modul. Chování detektoru bylo analyzováno v různých akustických podmínkách za použití signálové databáze `QUT-NOISE-TIMIT`, která obsahuje řečové signály smíchané s šumy různých charakterů (jedoucí automobil, domácí prostředí apod.) a s různými hodnotami odstupu signálu od šumu SNR.

V druhé kapitole jsou shrnuty použité metody zpracování číslicového signálu. V kapitole třetí následuje přehled současných systémů pro detekci řeči a teoretický popis vytvořeného detektoru. Ve čtvrté kapitole lze nalézt popis implementace detektoru a podrobný popis použití. Pátá kapitola dokumentuje experimentální část této bakalářské práce a diskutuje dosažené výsledky.

Kapitola 2

Metody zpracování řečového signálu

Pro zpracování řečového signálu v počítači je nutné signál nejprve převést do číslicové podoby, což znamená převést analogový signál na signál diskrétní v čase i v hodnotách. To zajistí procesy vzorkování a kvantování. Základním parametrem vzorkování je vzorkovací frekvence f_s, která udává, s jakou frekvencí budou v čase snímány vzorky signálu. Podle Shannonova (též Nyquistova nebo Kotělnikova) teorému musí být vzorkovací frekvence f_s nejméně dvojnásobná vzhledem k nejvyšší složce f_{max} ve spektru signálu:

$$f_s \geq 2f_{max}. \tag{2.1}$$

V opačném případě by došlo k prolnutí spekter vzorkovaného signálu (k tzv. aliasingu). Tento požadavek je potřeba zajistit dolnofrekvenční propustí na vstupu A/D převodníku. Vzhledem k tomu, že většina podstatné informace je v řečovém signálu obsažena v pásmu do 8 kHz [11], pro zpracování řeči se běžně používá vzorkovací frekvence $f_s = 16$ kHz.

Důsledkem kvantování signálu na konečný počet hladin je pak vznik kvantizačního šumu. Počet kvantizačních hladin, na které je signál kvantován, přímo ovlivňuje úroveň kvantizačního šumu v signálu. Běžně používané 16-bitové kódování 16-PCM zajišťuje dostatečný počet kvantizačních hladin.

Protože je řečový signál stacionární pouze v krátkodobých úsecích (je kvazistacionární), je potřeba použít tzv. *krátkodobou analýzu signálu*. To

znamená vybírat pro analýzu ze signálu časové úseky (segmenty) relativně krátké délky. Krátkodobé segmenty je (pro omezení prosakování ve spektru) v časové oblasti nutné vybrat ze signálu vhodným váhovacím oknem. Při zpracování řečového signálu se nejčastěji používá tzv. *Hammingovo okno* (vztah 2.2).

$$w[n] = 0,54 - 0,46 \cos \frac{2\pi n}{N}, \quad \text{pro} \quad n = 0, \ldots, N-1, \tag{2.2}$$

kde N je délka okna (počet vzorků). Krátkodobý i-tý segment $x_i[n]$ tedy bude z vstupního signálu vybrán takto:

$$x_i[n] = x[N \cdot i + n].w[n], \quad n = 0, \ldots, N-1. \tag{2.3}$$

Velikost krátkodobého segmentu N je pro krátkodobou analýzu řeči vhodné volit tak, aby časově odpovídala době trvání 10 až 30 ms, což odpovídá délce kvazistacionárních úseků lidské řeči.

2.1 Analýza v časové oblasti

Mezi řečové charakteristiky používané pro analýzu řečového signálu v časové oblasti patří energetické charakteristiky (krátkodobá energie a krátkodobý výkon signálu) a počet průchodů nulou (*Zero-Crossing Rate*). Krátkodobá energie E_i z i-tého segmentu signálu je definována vztahem 2.4 a krátkodobý výkon vztahem 2.5.

$$E_i = \sum_{n=0}^{N-1} x_i^2[n] \tag{2.4}$$

$$P_i = \frac{1}{N} \sum_{n=0}^{N-1} x_i^2[n] \tag{2.5}$$

Počet průchodů nulou (ZCR) udává, kolikrát za sekundu signál průměrně změní znaménko (vztah 2.6).

$$\mathrm{ZCR} = \frac{1}{N}\sum_{n=0}^{N-1}\frac{|sgn(x[n]) - sgn(x[n-1])|}{2}\cdot f_s \ \ [\mathrm{Hz}], \tag{2.6}$$

kde

$$sgn(x) = \begin{cases} 1, & \text{pokud} \quad x \geq 0 \\ -1, & \text{pokud} \quad x < 0. \end{cases} \tag{2.7}$$

2.2 Spektrální analýza

Spektrum (krátkodobého) diskrétního signálu o konečném počtu vzorků N definuje Diskrétní Fourierova Transformace (DFT):

$$X_i[k] = \sum_{n=0}^{N-1} x_i[n] e^{-jkn\frac{2\pi}{N}}, \tag{2.8}$$

kde $x_i[n]$ je hodnota n-tého vzorku i-tého krátkodobého segmentu a $X_i[k]$ je k-tá složka jeho spektra. DFT spektrum krátkodobého signálu o délce N vzorků tedy obsahuje N spektrálních složek. Implementace výpočtu krátkodobého spektra používá FFT algoritmus.

2.3 Kepstrální analýza

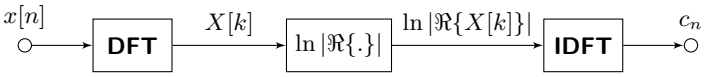
Vhodnou metodou pro zpracování řeči je Kepstrální analýza. Důležitou vlastností vektoru kepstrálních koeficientů je fakt, že rozdílnost dvou zvuků je úměrná euklidovské vzdálenosti vektorů jejich kepster v L-rozměrném prostoru (kde L je počet kepstrálních koeficientů).

2.3.1 Reálné kepstrum

Kepstrální analýza je algoritmus s dvojí transformací. Reálné kepstrum lze vypočítat jako inverzní Fourierovu transformaci logaritmu amplitudového

spektra. Na spektrum signálu se tedy díváme jako na periodický signál a harmonické složky signálu, které se ve spektru zobrazí jako periodicky rozmístěné složky, se po druhé transformaci zobrazí jako významná špička, odpovídající základní frekvenci. Kepstrum lze dobře použít na měření periodicity signálu. Výpočet reálného kepstra ukazuje vztah 2.9 a blokové schéma jeho výpočtu ukazuje obrázek 2.1. Vektor koeficientů c_n značí reálné kepstrum signálu.

$$c_n = IDFT\{\ln|\Re\{X[k]\}|\} = \sum_{k=0}^{N-1} \ln|\Re\{X[k]\}|e^{jkn\frac{2\pi}{N}} \tag{2.9}$$



Obrázek 2.1: Blokové schéma výpočtu reálného kepstra

2.3.2 LPC kepstrum

Jiným způsobem výpočtu kepstra lze získat *LPC kepstrum* s odlišnými vlastnostmi. Výpočet LPC (*Linear Predictive Coding*) kepstra vychází z *AR (autoregresního) modelu signálu*. Ten předpokládá, že vzorek $x[n]$ zkoumaného signálu lze predikovat jako lineární kombinaci p minulých vzorků. Cílem AR analýzy je určit *autoregresní koeficienty signálu* a_k tak, aby chyba predikce $\tilde{x}[n]$ byla co nejmenší, tzn. aby energie rozdílového signálu $e[n]$ byla co nejmenší.

$$e[n] = x[n] - \tilde{x}[n] = x[n] + \sum_{k=1}^{p} a_k x[n-k] \tag{2.10}$$

Na rovnice 2.10 se lze dívat jako na diferenční rovnicí FIR filtru s přenosovou funkcí $A(z)$:

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}. \tag{2.11}$$

V případě, že chybový signál $e[n]$ na výstupu tohoto filtru má charakter bílého šumu, filtr s koeficienty a_k se nazývá *syntetizujícím filtrem AR modelu signálu* s přenosovou funkcí $H(z)$.

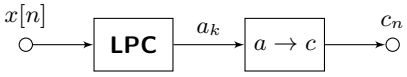
$$H(z) = \frac{G}{A(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{2.12}$$

AR koeficienty lze pomocí vztahů 2.13 a 2.14 použít k výpočtu LPC kepstrálních koeficientů c_n:

$$c_0 = \ln G, \tag{2.13}$$

$$c_n = -a_n - \frac{1}{n} \sum_{k=1}^{n-1} (n-k) a_k c_{n-k} \qquad \text{pro } n = 1, 2, \ldots, p. \tag{2.14}$$

Na rozdíl od DFT kepstra je LPC kepstrum vyhlazené. Důvodem je fakt, že LPC spektrum neobsahuje informaci o periodicitě, která je obsažena v chybovém signálu $e[n]$. Pro nalezení odhadu koeficientů a_k a zizku G, které odpovídají AR modelu 2.10, lze použít tzv. *Levinsonův-Durbinův algoritmus* (popsán v [12]) anebo *Burgův algoritmus*.



Obrázek 2.2: Blokové schéma výpočtu LPC kepstra

2.4 Parametrizace řečového signálu

Pro klasifikaci řečového signálu je nutné ze signálu extrahovat vektor příznaků, který obsahuje potřebnou informaci při celkové redukci objemu dat. Tento postup se nazývá *parametrizace řečového signálu* a často používané metody jsou řečová parametrizace MFCC a PLP, které budou popsány níže.

2.4.1 Řečová parametrizace MFCC

Řečová parametrizace MFCC (též mel-kepstrum, anebo *Mel-Frequency Cepstral Coefficients*) zohledňuje nelineární vnímání frekvence lidským sluchem. Při

jejím výpočtu je na spektrum signálu $S[k]$ aplikována banka filtrů, která spektrum podle frekvenční osy rozděluje do M pásem. Rozdělení frekvenční osy je nelineární a odpovídá přechodu do melodické frekvenční osy v *melech*:

$$f_{mel} = \mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{2.15}$$

$$f = \mathrm{InvMel}(f_{mel}) = 700 \cdot (10^{\frac{f_{mel}}{2595}} - 1). \tag{2.16}$$

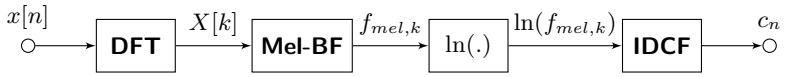
V melodické frekvenční stupnici mají všechny filtry stejnou šířku pásma (melodická frekvenční osa je rovnoměrně rozdělena do M pásem) a vzájemný překryv se rovná polovině šířky pásma. Frekvenční charakteristika filtrů je trojúhelníková. Dalším krokem je výpočet logaritmu energie v jednotlivých pásmech:

$$g_j = \ln \sum_{k=0}^{N/2} |S[k]|^2 H_{mel,j}[k], \tag{2.17}$$

kde $H_{mel,j}[k]$ je diskrétní frekvenční charakteristika filtru pro j-té pásmo. Aplikací kosinové transformace (vztah 2.18) na tyto hodnoty pásmové energie g_j získáváme koeficienty mel-kepstra c_n.

$$c_n = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} g_j \cos\left(\frac{\pi i}{P}(j - 0,5)\right) \tag{2.18}$$

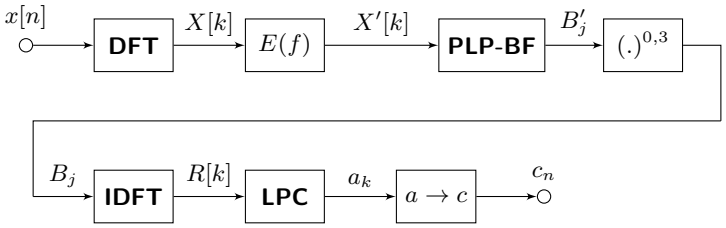
Blokové schéma výpočtu mel-kepstra zobrazuje obrázek 2.3.



Obrázek 2.3: Blokové schéma výpočtu mel-kepstra

2.4.2 Řečová parametrizace PLP

PLP parametrizace [5] (Perceptual Linear Predictive Coding) řečového signálu představuje další typ kepstrálních příznaků pro rozpoznávání řeči. Rozdíl ve výpočtu PLP oproti příznakům MFCC je možné vidět na obr. 2.4. PLP parametrizace používá při výpočtu odlišnou banku filtrů, která vychází z Barkovy frekvenční stupnice. Z blokového schématu je dále vidět, že se energie v jednotlivých pásmech umocňuje na 0,3, čímž se bere při výpočtu v úvahu tvz. zákon slyšení.



Obrázek 2.4: Blokové schéma výpočtu PLP kepstra

Kapitola 3

Detektor řečové aktivity

Vzhledem k širokému rozšíření hlasových technologií v různých aplikacích existují také různé požadavky na detektory řečové aktivity, jako jsou rychlost detekce, přesnost detekce řečových segmentů, anebo nízká četnost falešných detekcí. Z těchto důvodů existuje množství algoritmů pro tuto úlohu. Současné algoritmy pro detekci řeči je možné rozdělit do dvou hlavních skupin: *deterministické algoritmy* a *stochastické algoritmy*.

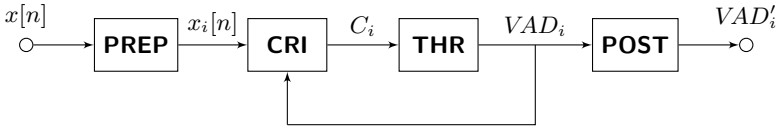
- Deterministické algoritmy pracují s vybranou charakteristikou signálu, vypočítanou danou kriteriální funkcí. O přítomnosti řeči v daném úseku signálu je rozhodováno na základě porovnávání hodnoty kriteriální funkce s určitou prahovou hodnotou. Sledovanou charakteristikou signálu může být například krátkodobá (spektrální) energie signálu, počet průchodů nulou (Zero-Crossing Rate), anebo rozdíly v kepstrálních charakteristikách, jako v případě detektoru popsaného v [13] . Pro stanovení rozhodovacího prahu je možné použít různé heuristické metody.
- Stochastické algoritmy jsou založeny na různých statistických metodách. Příkladem je algoritmus GMM (Gaussian Mixture Model), který obsahuje model řeči a model signálu bez řeči a na jejich základě stanovuje příslušnost analyzovaného signálu k jedné z těchto tříd. Příkladem jsou detektory popsané v [19], [17] a [25]. Jiné algoritmy z této skupiny zase pracují s principem umělé inteligence (neuronové sítě). Společným znakem algoritmů z této skupiny je nutnost podrobit detektor trénovací fázi na trénovacím vzorku dat.

V této bakalášské práci byl implementován a zkoumán detektor ze skupiny

deterministických algoritmů. Jako kriteriální funkce slouží změny v kepstrálních charakteristikách signálu a pro stanovení prahu jsou použity různé algoritmy heuristického prahování.

3.1 Popis vytvořeného detektoru

Detektor je možné popsat pomocí čtyř nezávislých funkčních bloků: předzpracování `(PREP)`, výpočet kriteriální funkce `(CRI)`, stanovení prahové hodnoty kriteriální funkce `(THR)` a post-processing `POST` (viz. schéma na obrázku 3.1). Následuje popis jednotlivých bloků.



Obrázek 3.1: Blokové schéma celého detektoru

3.1.1 Předzpracování (PREP)

V rámci předzpracování je provedeno načtení signálu, segmentace na krátkodobé segmenty (váhování) a výpočet spektra.

3.1.2 Výpočet kriteriální funkce (CRI)

Úlohou bloku pro výpočet hodnoty kriteriální funkce je převzít na vstupu vektor s hodnotami vzorků i-tého krátkodobého segmentu vstupního signálu a z něj vypočítat hodnotu kriteriální funkce C_i. K výpočtu hodnoty kriteriální funkce je možné použít jednu z následujících metod:

- krátkodobá energie i-tého segmentu signálu:

$$C_i = E_i = \sum_{n=0}^{N-1} x_i^2[n] \tag{3.1}$$

- krátkodobá energie v decibelech:

$$C_i = E_{dB,i} = 10 \log \sum_{n=0}^{N-1} x_i^2[n] \tag{3.2}$$

- kepstrální vzdálenost CD_i mezi kepstrem i-tého segmentu a odhadem kepstra pozadí $c_{bkg,i}$

$$C_i = CD_i = \sqrt{\sum_{k=1}^{L} (c_i[k] - c_{bkg,i}[k])^2}, \tag{3.3}$$

kde c_i je vektor s kepstrálními koeficienty aktuálního segmentu a $c_{bkg,i}$ reprezentuje aktuální odhad kepstra pozadí. L vyjadřuje počet použitých kepstrálních koeficientů.

Odhad kepstra pozadí $c_{bkg,i}$ je v neřečových segmentech ($VAD_i = 0$) a v inicializační části na začátku signálu ($i \leq N$) rekurentně aktualizován použitím následujícího vztahu:

$$c_{bkg,i} = p \cdot c_{bkg,i-1} + (1 - p) \cdot c_i, \tag{3.4}$$

kde p je parametr, který ovlivňuje rychlost zapomínání odhadu kepstra pozadí.

3.1.3 Stanovení prahové hodnoty kriteriální funkce (THR)

Prahová hodnota kriteriální funkce $C_{thr,i}$ je vyhodnocena pro každý i-tý krátkodobý segment a porovnávána s hodnotou kriteriální funkce C_i. Pokud je kriteriální hodnota větší než stanovený práh, segment je vyhodnocen jako segment obsahující řeč a hodnota detekce VAD_i pro daný segment bude 1 (vztah 3.5).

$$VAD_i = \begin{cases} 1, & \text{pokud} \quad C_i \geq C_{thr,i} \\ 0, & \text{pokud} \quad C_i < C_{thr,i} \end{cases} \tag{3.5}$$

Ve vytvořeném detektoru je možné zvolit jednu z několika různých metod pro stanovení prahové hodnoty:

- fixní práh daný absolutním číslem C_{thr}:

$$C_{thr,i} = C_{thr}. \tag{3.6}$$

- práh procentuelně umístěný v rozsahu daném minimální a maximální hodnotou kriteriální funkce. Použitá implementace vyhodnocuje hodnotu prahu $C_{thr,i}$ pro každý segment na základě dosud nalezené minimální ($C_{min,i}$) a maximální ($C_{max,i}$) hodnoty kriteriální funkce:

$$C_{thr,i} = C_{min,i} + \frac{T}{100} \cdot (C_{max,i} - C_{min,i}), \tag{3.7}$$

  kde parametr T představuje procentuelní umístění prahu v daném rozsahu.

- adaptivní práh sleduje změny hodnoty kriteriální funkce související se změnami v charakteristikách pozadí. Práh je stanoven na základě stochastických vlastností variability kriteriální funkce v neřečových segmentech ($VAD_i = 0$) a v inicializační části na začátku signálu ($i < N$) podle vztahu 3.8.

$$C_{thr,i} = \mu_{C,i} + z_\alpha . \sigma_{C,i} \tag{3.8}$$

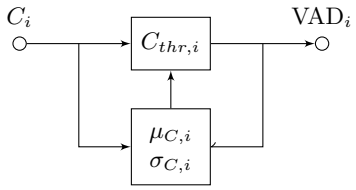
$$\mu_{C,i} = q \cdot \mu_{C,i-1} + (1 - q) \cdot C_i \tag{3.9}$$

$$\mu_{C2,i} = q \cdot \mu_{C2,i-1} + (1 - q) \cdot C_i^2 \tag{3.10}$$

$$\sigma_{C,i} = \sqrt{\mu_{C2,i} - (\mu_{c,i})^2} \tag{3.11}$$

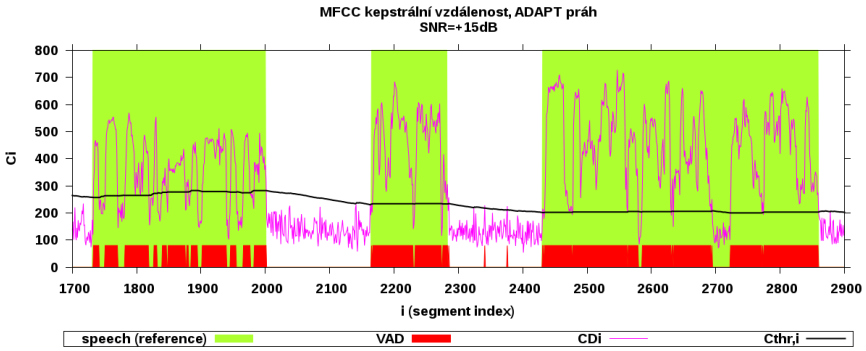
  $C_{thr,i}$ je hodnota prahu pro aktuální, i-tý segment, $\mu_{C,i}$ je střední hodnota kriteriální funkce v řečových pauzách a $\sigma_{C,i}$ je standardní odchylka kriteriální funkce v řečových pauzách. z_α je koeficient specifikující pravděpodobnostní interval variability hodnot kriteriální funkce. Střední hodnota $\mu_{C,i}$ a standartní odchylka $\sigma_{C,i}$ jsou rekurentně aktualizovány v neřečových segmentech a v inicializační fázi podle vztahů 3.9, 3.10 a 3.11. Parametr q ovlivňuje rychlost exponenciálního zapomínání v těchto vztazích. Adaptivní práh je systém se zpětnou vazbou a vyžaduje inicializaci. Proto musí být zaručeno, že určitý časový úsek (určitý počet segmentů N) na začátku signálu neobsahuje řeč. V opačném případě může práh v důsledku existující zpětné vazby zůstat saturován na jedné hladině.

  Obrázek 3.3 ilustruje průběh adaptivního prahu při zpracování kriteriální funkce na bázi MFCC kepstrální vzdálenosti a výslednou VAD detekci. Na obrázku jsou zeleným pozadím vyznačeny úseky, které podle referenční anotace obsahují řeč. V řečových úsecích je viditelná významná změna průběhu kriteriální funkce CD_i (purpurová barva). Dále je na



Obrázek 3.2: Blokové schéma adaptivního prahu

obrázku možné vidět průběh prahové hodnoty adaptivního prahu $C_{thr,i}$ (zobrazeno černě). Z obrázku je zřetelné, že v řečových úsecích není práh aktualizován. Výsledná detekce řeči (tzn. úseky, kde hodnota kriteriální funkce překračuje prahovou hodnotu) je zobrazena červenými úseky v dolní části grafu, které v tomto případě uspokojivě korelují s referenční anotací. Na vodorovné ose je zaneseno pořadí krátkodobého segmentu v signálu (rozestup segmentů je 10 ms).



Obrázek 3.3: Zobrazení průběhu kriteriální funkce, adaptivního prahu a hodnoty detekce

- adaptivní práh procentuelně umístěný v aproximaci dynamického rozsahu hodnot kriteriální funkce:

$$C_{thr,i} = C_{min,i} + \frac{T}{100}(C_{max,i} - C_{min,i}) \tag{3.12}$$

$$C_{max,i} = q \cdot C_{max,i-1} + (1-q) \cdot C_i \tag{3.13}$$

$$C_{min,i} = q \cdot C_{min,i-1} + (1-q) \cdot C_i, \tag{3.14}$$

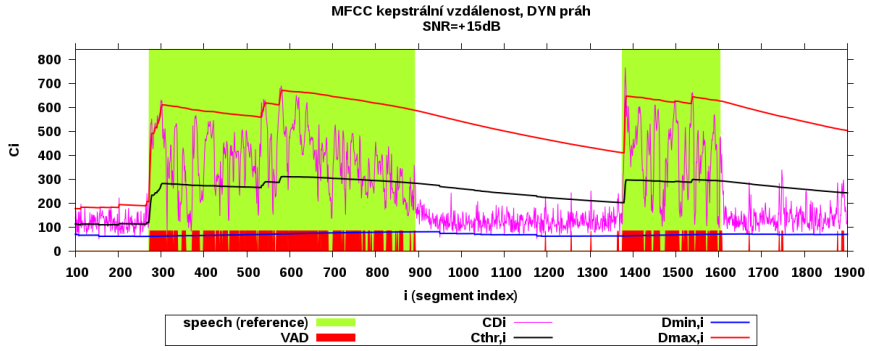
kde T vyjadřuje procentuelní umístění prahu v dynamickém rozsahu a q je parametr rychlosti aktualizace odhadu dynamického minima $C_{min,i}$ a maxima $C_{max,i}$. V implementaci VAD detektoru je možné nezávisle volit parametr q pro rychlost náběhu i poklesu dynamického maxima i minima. Parametr q je tedy v detektoru reprezentován čtyřmi nezávisle volitelnými parametry. Dále je nutné pro dynamický práh zavést parametr $C_{\Delta min}$, který vyjadřuje hodnotu minimální dynamiky. Při poklesu dynamiky $C_{\Delta,i}$ pod tuto hodnotu je aktuální segment vždy vyhodnocen jako segment bez řeči. Vztah 3.5 je tedy pro tuto metodu modifikován do podoby 3.15:

$$VAD_i = \begin{cases} 1, & \text{pokud} \quad C_i \geq C_{thr,i} \wedge C_{\Delta,i} > C_{\Delta min} \\ 0 & \text{v ostatních případech,} \end{cases} \tag{3.15}$$

kde $C_{\Delta,i}$ je aktuální hodnota aproximace dynamického rozsahu:

$$C_{\Delta,i} = C_{max,i} - C_{min,i}. \tag{3.16}$$

Na obrázku 3.4 je možné vidět chování dynamického prahu. Zobrazeny jsou průběhy hodnoty kriteriální funkce, odhadu jejího dynamického minima $C_{min,i}$ (modře) a maxima $C_{max,i}$ (červeně). Práh $C_{thr,i}$ (zobrazen černě) je v tomto případě nastaven v 40-ti procentech dynamického rozsahu. Ostatní průběhy jsou zobrazeny stejně jako na obrázku 3.3 a i zde je možné vidět, jak VAD detekce kopíruje referenční anotaci.

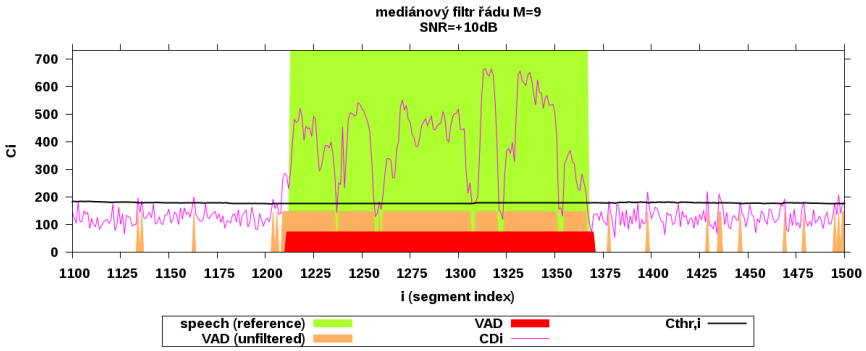


Obrázek 3.4: Zobrazení průběhu kriteriální funkce, dynamického prahu a hodnoty detekce

3.1.4 Post-processing (POST)

Na hodnotu detekce VAD_i je možné aplikovat filtraci mediánovým filtrem M-tého řádu. Princip mediánové filtrace spočívá ve zpracování M po sobě jdoucích hodnot detekce pro získání výsledné hodnoty pro jeden segment. Aktuálně vyhodnocovaný segment je uprostřed této posloupnosti. Hodnoty jsou seřazeny podle velikosti a výsledná hodnota detekce odpovídá hodnotě, která se nachází uprostřed tohoto seřazení. Filtrace je provedena pro každou hodnotu detekce VAD_i a M okolních hodnot (M je liché celé číslo, vhodná hodnota pro filtraci VAD detekce je $M = 5$ [11]). Výsledná hodnota VAD'_i tím bude vyhlazena od krátkodobých chyb typu *False Alarm* (chybně vyhodnocený neřečový segment) a *Miss* (chybně vyhodnocený řečový segment).

Vliv mediánové filtrace na VAD detekci je možné vidět na obrázku 3.5. Oranžovými úseky v dolní části grafu je zobrazena původní (nefiltrovaná) detekce a červenou barvou je zobrazena VAD detekce po aplikaci mediánového filtru řádu $M = 9$. Jak je vidět, filtrace v tomto případě eliminovala falešné krátkodobé detekce v levém a pravém okolí promluvy (oblast s promluvou vyznačena zeleně) a také úspěšně doplnila chybějící řečové úseky, nesprávně vyhodnocené jako řečová pauza v důsledku krátkodobého poklesu hodnoty kriteriální funkce (zobrazena purpurovou barvou).



Obrázek 3.5: Zobrazení alikace mediánové filtrace na VAD detekci

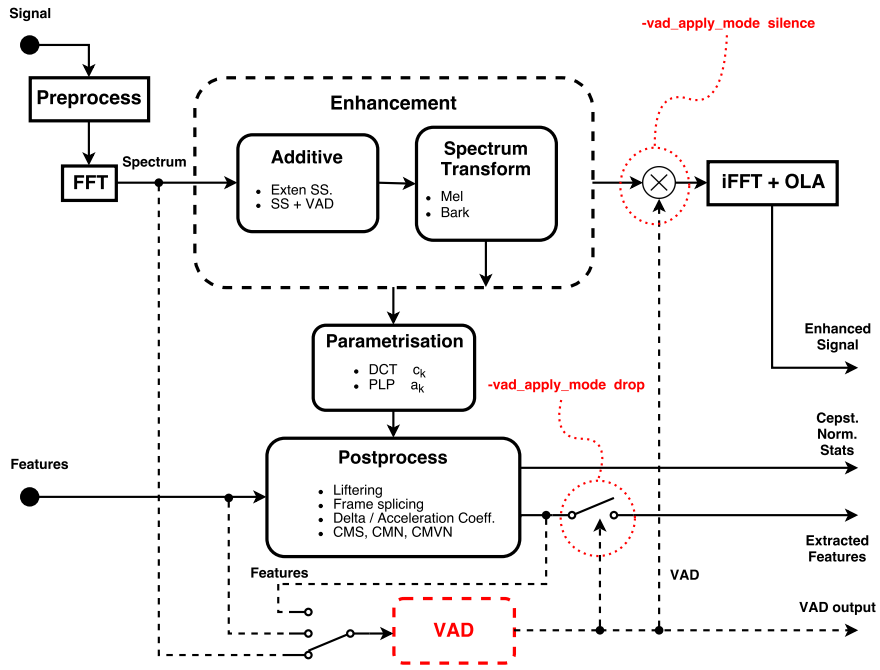
Kapitola 4

Implementace detektoru

V této kapitole bude popsána implementace VAD detektoru a začlenění do nástroje CtuCopy [18], [3]. Nástroj CtuCopy byl vyvinut v řečové laboratoři na katedře teorie obvodů pro vývoj algoritmu potlačení šumu a robutních parametrizací řečového signálu, které jsou základem pro systémy rozpoznávání řeči, identifikaci mluvčího, nebo rozpoznávání jazyka. Nástroj CtuCopy je implementován v jazyce C++, a proto byl tento jazyk použit i pro implementaci VAD detektoru.

4.1 CtuCopy

Nástroj CtuCopy umožňuje aplikovat algoritmy zvýrazňování řeči na vstupní signál nebo extrakci řečových příznaků. Z pohledu uživatele je nástroj volán z příkazové řádky a je podporován operačními systémy Linux a Windows. Nástroj umožnuje číst vstupní signál ze standardní konzole `std::in` anebo ze souborů různých formátů a následně zapisovat zvýrazněný signal/příznaky do standardní konzole `std::out` či do souboru. Zjednodušený diagram funkčních modulů je zobrazen na obr. 4.1. Jak je vidět z toho obrázku, jednotlivé části nástroje jsou implementovány jako nezávislé funkční bloky, které je možné vzájemně řetězit a kombinovat. Na obrázku je dále možné vidět začlenění VAD modulu do struktury CtuCopy. Detailní popis VAD modulu je popsán v následující části.



Obrázek 4.1: Blokové schéma programu CtuCopy s vyznačeným začleněním modulu VAD

4.2 Implementace VAD detektoru

Jak je uvedeno výše, VAD detektor byl naprogramován jako funkční modul do existujícího softwarového nástroje CtuCopy. Začlenění detektoru do nástroje CtuCopy je zobrazeno na obrázku 4.1. Z blokového schématu je vidět, že vstupními daty VAD modulu může být spektrum signálu, anebo řečové příznaky z parametrizařního modulu, případně i ze vstupního souboru. Výstup detekce je potom možné uložit do souboru (*VAD Output*), použít pro vyřazování neřečových příznaků z ukládání, případně použít ke klíčování výstupního signálu (neřečové segmenty budou nahrazeny tichou pasáží). Červeně jsou na obrázku vyznačeny konfigurační parametry, které ovlivňují spolupráci VAD modulu s ostatními částmi programu CtuCopy.

Samotný VAD detektor pracuje s kriteriální funkcí na bázi energie nebo kepstrální vzdálenosti, kde rozhodnutí o přítomnosti řeči je prováděno na základě různých způsobů prahování (adaptivní práh, práh na bázi sledování dynamiky). Detekci je možné následně vyhladit pomocí mediánové filtrace. V rámci implementace detektoru byla implementována interní třída pro výpočet LPC kepstrálních koeficientů, třídy pro výpočet kepstrální vzdálenosti a výpočet krátkodobé energie signálu a třídy pro stanovení prahu pro všechny

popsané metody.

Schéma detektoru je možné vidět na obrázku 4.2. Detektor využívá stávající moduly nástroje CtuCopy pro načtení signálu, předzpracování (segmentaci na krátkodobé segmenty, váhování), výpočet spektra, moduly potlačování šumu a parametrizace (výpočet krátkodobých řečových příznaků).

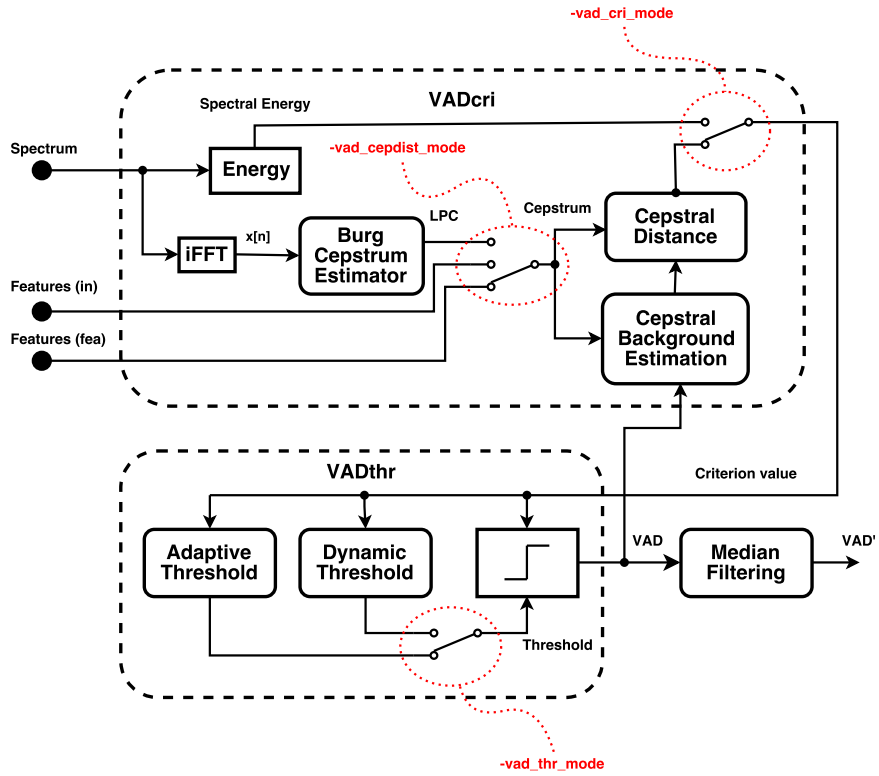
Hlavní třídou detektoru je třída `VAD` a čistě virtuální třídy `VADcri` a `VADthr`. Třída `VADcri` slouží jako prototyp pro třídy k výpočtu hodnoty kriteriální funkce. Tyto třídy jsou:

- `VADcri_energy` ... spektrální energie jako kriteriální funkce
- `VADcri_cepdist` ... kepstrální vzdálenost jako kriteriální funkce (vztah 3.3). Konfigurací je možné zvolit, co má být zdrojovým vektorem příznaků pro výpočet kepstrální vzdálenosti. To může být:
  - LPC kepstrum vypočítané přímo uvnitř `VADcri_cepdist` třídy ze vstupního signálu
  - vektor příznaků převzatý z `in` třídy (možné pouze pokud je vstupním souborem HTK soubor s příznaky)
  - vektor příznaků převzatý z `fea` třídy (pro tuto volbu je nutné, aby byl program CtuCopy nastaven k extrakci příznaků a ukládání HTK souboru)

Třída `VADthr` je prototyp třídy pro výpočet prahové hodnoty kriteriální funkce a pro určení hodnoty VAD detekce. To jsou třídy:

- `VADthr_absolute` ... fixní práh daný absolutním číslem (vztah 3.6)
- `VADthr_perc` ... práh daný procentuelním umístěním v rozsahu minimální a maximální hodnoty kriteriální funkce (vztah 3.7)
- `VADthr_adapt` ... adaptivní práh na bázi sledování změn v charakteristikách pozadí (vztah 3.8)
- `VADthr_dyn` ... adaptivní práh na bázi sledování dynamiky kriteriální funkce (vztah 3.12)

Hlavní třída `VAD` obsahuje instanční proměnné `VADcri* vadcri;` a `VADthr* vadthr;` s ukazateli na objekty typu `VADcri` a `VADthr`. Konkrétní typ objektu je vytvořen v závislosti na konfiguraci.



Obrázek 4.2: Blokové schéma implementace VAD modulu

4.3 Konfigurace a použití detektoru

Spouštění nástroje CtuCopy bylo rozšířeno o řadu parametrů příkazové řádky, sloužících ke konfiguraci detektoru. Základní parametry pro použití detektoru jsou:

- `-vad_out_mode <mode>` ... volba způsobu ukládání výstupu VAD detekce. Možné hodnoty `<mode>` jsou:
  - `none` ... VAD neaktivní
  - `vad` ... bude uložen VAD soubor s detekcí
  - `debug` ... bude uložen VAD soubor a dodatečné soubory s průběhy mezivýpočtů pro zvolenou metodu detekce. Vhodné pro vizualizaci a ladění parametrů.
- `-vad_apply_mode <mode>` ... volba způsobu aplikace VAD detekce na výstupní data. Možné hodnoty `<mode>` jsou:

- `none` ... VAD neaktivní
- `silence` ... neřečové segmenty budou ve výstupním signálu nahrazeny tichou pasáží. Tato volba je aktivní pouze v případě, že je program CtuCopy nastaven pro ukládání signálu.
- `drop` ... neřečové segmenty budou vyřazeny z ukládání. Vhodné při extrakci řečových příznaků a ukládání HTK souboru.

Pro aktivaci detektoru je nutné, aby byl alespoň jeden z parametrů `-vad_out_mode` a `-vad_apply_mode` nastaven na hodnotu jinou než `none`. Pokud je zapnuto ukládání detekce pomocí parametru `-vad_out_mode`, je nutné navíc specifikovat výstupní soubor `<file>` pomocí parametru `-vad_out <file>`.

Ostatní parametry detektoru jsou automaticky nastaveny na výchozí hodnoty a není nezbytně nutné je přenastavovat. Zde je jejich úplný seznam:

- `-vad_cri_mode <mode>` ... volba metody výpočtu kriteriální funkce. Možné hodnoty `<mode>` jsou:
  - `energy` ... krátkodobá energie signálu (3.1). Je možné zvolit, zda bude hodnota energie normalizována v decibelech (3.2) pomocí parametru `-vad_energy_db <on|off>`
  - `cepdist` ... kepstrální vzdálenost (vztah 3.3). Výpočet kepstrální vzdálenosti je možné ovlivnit následujícími parametry:
    - `-vad_cepdist_init <N>` ... délka N inicializační části (počet zaručeně neřečových segmentů) na začátku signálu pro inicializaci odhadu kepstra pozadí (vztah 3.4)
    - `-vad_cepdist_p <p>` ... parametr zapomínání při odhadech kepstra pozadí. Odpovídá koeficientu p ve vztahu 3.4
    - `-vad_cepdist_mode <mode>` ... volba vstupního vektoru pro výpočet kepstrální vzdálenosti. Možné hodnoty `<mode>` jsou:
      - `fea` ... vektor příznaků převzatých z tříd pro extrakci příznaků. Tuto volbu je možno použít pouze v případě, že program CtuCopy je nastaven k extrakci řečových příznaků a k ukládání HTK souboru pomocí volby `-format_out htk`.
      - `in` ... vektor příznaků ze vstupního souboru. Lze použít pouze pokud je vstupním souborem HTK soubor
      - `lpc` ... interní třída VAD detektoru pro výpočet LPC kepstrálních koeficientů. Je možné zvolit počet kepstrálních koeficientů L (vztah 3.3) pomocí parametru `-vad_lpc_coefs <L>`.
- `-vad_thr_mode <mode>` ... volba metody stanovení prahu. Možné hodnoty `<mode>` jsou:

- `absolute` ... práh nastaven na fixní, absolutní hodnotu (vztah 3.6), danou parametrem `-vad_absolute_thr <threshold>`. Vhodné pro ladění ostatních parametrů
- `perc` ... práh procentuelně nastaven v rozsahu minimální a maximální nalezené hodnoty kriteriální funkce, jak vyjadřuje vztah 3.7. Nastavení prahu T je možné zadat parametrem `-vad_perc_thr <T>`
- `adapt` ... adaptivní práh daný vztahem 3.8. Adaptivní práh je parametrizován následujícími parametry:
  - `-vad_adapt_init <N>` ... délka N inicializační části (počet zaručeně neřečových segmentů) na začátku signálu pro nastavení prahu
  - `-vad_adapt_q <q>` ... parametr zapomínání q, platný pro vztahy 3.9 a 3.10
  - `-vad_adapt_za <za>` ... hodnota koeficientu z_α, platná pro vztah 3.8
- `dyn` ... adaptivní dynamický práh podle vztahu 3.12. Parametrizován následujícími parametry:
  - `-vad_dyn_perc <T>` ... procentuelní umístění prahu v dynamickém rozsahu, odpovídá proměnné T ve vztahu 3.12
  - `-vad_dyn_min <min>` ... hodnota minimální dynamiky $C_{\Delta min}$
  - `-vad_dyn_qmaxinc <q>` ... parametr q ze vztahu 3.12 pro případ nárůstu dynamického maxima $C_{max,i}$
  - `-vad_dyn_qmaxdec <q>` ... parametr q pro případ poklesu dynamického maxima
  - `-vad_dyn_qmindec <q>` ... parametr q pro případ poklesu dynamického minima $C_{min,i}$
  - `-vad_dyn_qmininc <q>` ... parametr q pro případ nárůstu dynamického minima

- `-vad_filter_order <M>` ... velikost M mediánového filtru, aplikovaného na výstup VAD detektoru (viz. sekce 3.1.4)

Kapitola 5

Experimenty

V experimentální části bakalářské práce bylo provedeno testování správného chování vytvořeného detektoru. Kapitola prezentuje výsledky experimentů, které byly zaměřeny na chování detektoru v různých akustických prostředích. Pro tyto experimenty byla použita databáze `QUT-NOISE-TIMIT` [2], která byla vytvořena pro potřeby evaluace VAD detektorů v řečové komunitě. První sada experimentů se zaměřuje na porovnávání různých řečových příznaků (MFCC, PLP, LPC), standardně používaných v úloze rozpoznávání řeči pro detekci řečové aktivity pomocí kepstrálního VAD detektoru. Druhá část experimentů zkoumá vliv mediánové filtrace na výstup detektoru.

5.1 Databáze QUT-NOISE-TIMIT

Pro experimenty byla použita signálová databáze `QUT-NOISE-TIMIT` [2], která vzniká smícháním čistých řečových signálů z anglicky mluvené databáze `TIMIT` [1] s databází `QUT-NOISE`, obsahující nahrávky šumového pozadí z různých akustických prostředí. Databáze `QUT-NOISE-TIMIT` obsahuje signály s šumovým pozadím pěti různých typů: `CAFE`, `HOME`, `STREET`, `CAR` a `REVERB`. Pro každý typ šumu jsou obsaženy dvě různé nahrávací lokace (například lokace `KITCHEN` a `LIVINGB` pro typ šumu `HOME`). Podle těchto lokací je databáze rozdělena na dvě podmnožiny, označené jako *Group A* a *Group B*. V každé lokaci byla uskutečněna dvě nahrávací sezení, označená příponami `-1` a `-2`. Přehled množin signálů použitých k ladění parametrů a k evaluaci zachycují tabulky 5.1 a 5.2. Každá testovaná množina obsahuje 200 signálů, každý o délce 60 sekund.

označení testované množiny	obsažená šumová pozadí (lokace, nahrávací sezení)	podmnožiny podle SNR	zastoupení řeči (%)
CAFE	CAFE-FOODCOURTB-1 CAFE-FOODCOURTB-2	+10/+15 dB	81
		+5/0 dB	80
		-5/-10 dB	80
HOME	HOME-KITCHEN-1 HOME-KITCHEN-2	+10/+15 dB	78
		+5/0 dB	73
		-5/-10 dB	77
STREET	STREET-CITY-1 STREET-CITY-2	+10/+15 dB	82
		+5/0 dB	78
		-5/-10 dB	81
CAR	CAR-WINDOWNB-1 CAR-WINDOWNB-2	+10/+15 dB	70
		+5/0 dB	75
		-5/-10 dB	77
REVERB	REVERB-POOL-1 REVERB-POOL-2	+10/+15 dB	76
		+5/0 dB	77
		-5/-10 dB	72

Tabulka 5.1: Přehled množin signálů použitých pro ladění parametrů (skupina *Group A*)

označení testované množiny	obsažená šumová pozadí (lokace, nahrávací sezení)	podmnožiny podle SNR	zastoupení řeči (%)
CAFE	CAFE-CAFE-1 CAFE-CAFE-2	+10/+15 dB	71
		+5/0 dB	79
		-5/-10 dB	76
HOME	HOME-LIVINGB-1 HOME-LIVINGB-2	+10/+15 dB	82
		+5/0 dB	81
		-5/-10 dB	77
STREET	STREET-KG-1 STREET-KG-2	+10/+15 dB	85
		+5/0 dB	76
		-5/-10 dB	77
CAR	CAR-WINUPB-1 CAR-WINUPB-2	+10/+15 dB	78
		+5/0 dB	84
		-5/-10 dB	80
REVERB	REVERB-CARPARK-1 REVERB-CARPARK-2	+10/+15 dB	71
		+5/0 dB	81
		-5/-10 dB	76

Tabulka 5.2: Přehled množin signálů použitých pro evaluaci (skupina *Group B*)

Pro všechny uvedené množiny jsou k dispozici signály s danými úrovněmi SNR: +15, +10, +5, 0, -5 a -10 dB. Při experimentech v této práci byla data podle úrovně SNR seskupena do tří skupin:

- nízká úroveň šumu (SNR +15 a +10 dB)
- střední úroveň šumu (SNR +5 a 0 dB)
- vysoká úroveň šumu (SNR -5 a -10 dB)

Databáze `QUT-NOISE-TIMIT` obsahuje další členění související s poměrem řečových a neřečových segmentů ve výsledném signálu. Tomuto členění odpovídají podmnožiny označené jako `sA` a `sB`. Pro experimenty v rámci této bakalářské práce byla použita pouze podmnožina `sA`. Dále byla data omezena pouze na signály o délce 60 sekund (nebyly použity signály s délkou 120 sekund). Z celkové databáze tedy byla použita právě čtvrtina všech signálů. Tento výběr odpovídá výběru signálů v experimentální části článku [23].

Vzhledem k tomu, že použité algoritmy vyžadují inicializaci, byl pro účely této práce skript generující databázi `QUT-NOISE-TIMIT` upraven tak, aby

výsledné signály vždy zaručeně obsahovaly inicializační (neřečovou) část o délce minimálně 2 sekundy. Požitá databáze tedy není identická s originální databází `QUT-NOISE-TIMIT` a při porovnávání výsledků s jinými autory je nutné toto brát v úvahu. Úprava kódu, která tuto modifikaci umožňuje, je dostupná na přiloženém CD.

5.2 Použité metriky

K vyhodnocení úspěšnosti detektoru byl použit existující nástroj *vadcrit*, který umožňuje vypočítat úspěšnost detekce VAD_i porovnáním s referenční anotací REF_i. Byla sledována kritéria FAR (False Alarm Rate), MR (Miss Rate) a HTER (Half Total Error Rate), popsaná například v [23].

$$\mathrm{MR} = 100 \cdot \frac{\sum_{i=0}^{N-1} |\mathrm{REF}_i - \mathrm{VAD}_i| \cdot \mathrm{REF}_i}{\sum_{i=0}^{N-1} \mathrm{REF}_i} \quad [\%] \tag{5.1}$$

$$\mathrm{FAR} = 100 \cdot \frac{\sum_{i=0}^{N-1} |\mathrm{REF}_i - \mathrm{VAD}_i| \cdot (1 - \mathrm{REF}_i)}{\sum_{i=0}^{N-1} (1 - \mathrm{REF}_i)} \quad [\%] \tag{5.2}$$

Aritmetickým průměrem z hodnot MR a FAR lze získat hodnotu kritéria HTER (vztah 5.3), standardně používaného pro evaluaci VAD detektorů ([2], [23]), které bylo použito pro konečné srovnávání úspěšnosti detektorů.

$$\mathrm{HTER} = \frac{\mathrm{MR} + \mathrm{FAR}}{2} \quad [\%] \tag{5.3}$$

5.3 Nastavení experimentů

Vzorkovací frekvence všech testovaných signálů byla $f_s = 16$ kHz. Pro délku krátkodobého segmentu w a rozestup krátkodobých segmentů s byly použity hodnoty obvyklé pro zpracování řečových signálů: $w = 25$ ms a $s = 10$ ms. Algoritmy pro výpočet řečových příznaků byly nastaveny na níže uvedené konfigurace (obvyklé parametry pro extrakci řečových příznaků).

- Nastavení CtuCopy pro extrakci MFCC řečových příznaků:

```
-fea_ncepcoefs 12
-fea_kind dctc
-fea_c0 on
-fea_E off
-fea_lifter 22
-fb_scale mel
-fb_shape triang
-fb_power on
-fb_eqld off
-fb_inld off
-fb_definition '100-7940Hz:1-30/30filters'
-dither 0.1
-preem 0.97
```

- Nastavení CtuCopy pro extrakci PLP řečových příznaků:

```
-fea_ncepcoefs 12
-fea_kind lpc
-fea_c0 on
-fea_E off
-fea_lifter 22
-fea_lporder 12
-fb_scale bark
-fb_shape trapez
-fb_power on
-fb_eqld on
-fb_inld on
-fb_definition '100-8000Hz:1-22/22filters'
-dither 0.1
-preem 0
```

- Při výpočtu LPC kepstra bylo použito prvních 12 koeficientů vypočítaných pomocí třídy `BurgCepstumEstimator`.

Pro každou testovanou množinu signálů (`CAFE`, `HOME`, ...) byly na podmnožině *Group A* (tabulka 5.1) nalezeny vhodné parametry prahovacích algoritmů a s těmito parametry byla na podmnožině *Group B* (tabulka 5.2) vyhodnocena úspěšnost detektoru. Ladění parametrů tedy vždy probíhalo v odlišném akustickém prostředí než výsledná evaluace.

5.3.1 Zkoumané algoritmy

Zkoumán byl detektor s kriteriální funkcí na principu kepstrální vzdálenosti (vztah 3.3). Podle použité metody pro výpočet kepstra a prahovacího algoritmu dostáváme šest různých detektorů:

- MFCC kepstrum (odstavec 2.4.1), DYN práh (vztah 3.12)
- MFCC kepstrum, ADAPT práh (vztah 3.8)
- PLP kepstrum (odstavec 2.4.2), DYN práh
- PLP kepstrum, ADAPT práh
- LPC kepstrum (odstavec 2.3.2), DYN práh
- LPC kepstrum, ADAPT práh

5.4 Dosažené výsledky

V této sekci budou prezentovány a diskutovány výsledky detektoru při výše uvedených experimentech.

5.4.1 Porovnání řečových příznaků pro VAD detekci

Dosažené hodnoty HTER pro všechny zkoumané detektory a všechna akustická prostředí shrnuje graf 5.1. Výsledky jsou rozčleněny do tří kategorií podle úrovně šumu tak, jak bylo uvedeno v sekci 5.1. V tabulce 5.3 jsou obsaženy hodnoty získané výpočtem aritmetického průměru z výsledků pro jednotlivá akustická prostředí (typy šumu). V posledním sloupci je aritmetický průměr z hodnot pro jednotlivé úrovně šumu a zvýrazněn je detektor, který celkové dosáhl nejlepšího výsledku. Kromě hlavního kritéria HTER zobrazuje tabulka 5.3 hodnoty kritérií MR a FAR, která udávají relativní chybovost detekce zvlášť na řečových a neřečových segmentech.

Pro danou aplikaci, pro kterou má být detektor použit, může být důležitější jedno z těchto kritérií než celková úspěšnost detekce. Například v případě

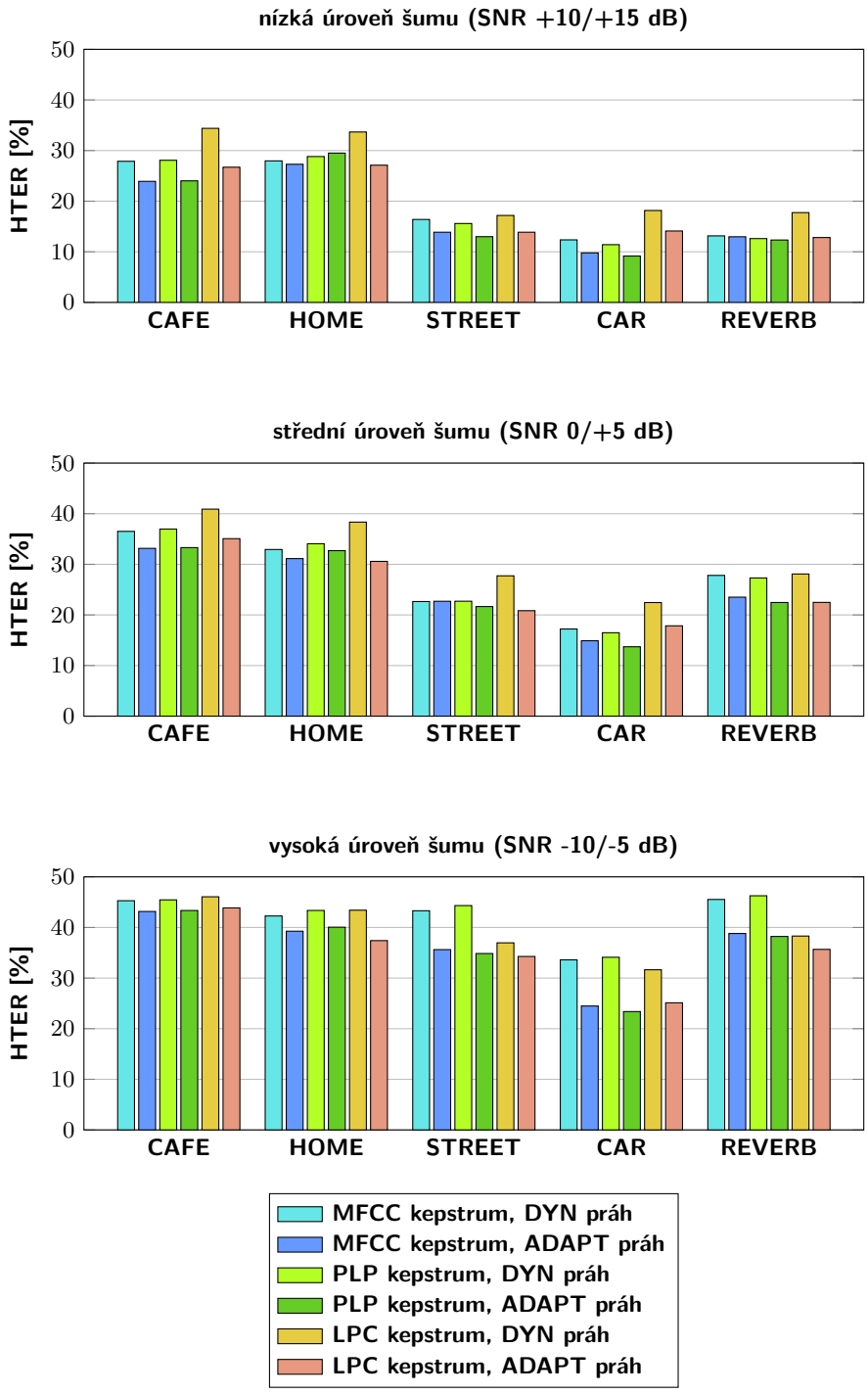
extrakce řečových příznaků pro identifikaci mluvčího je žádoucí, aby výsledná data obsahovala zaručeně jen řečové segmenty a chybějící řečové segmenty nejsou zásadním problémem při identifikaci. V tom případě je nutné zvolit konfiguraci detektoru s co nejnižší hodnotou FAR (False Alarm Rate). Naopak pro úlohu rozpoznávání řeči je nutné, aby ve výsledných datech chybělo co nejméně řečových segmentů. V tomto případě bude zajímavé hlavně kritérium MR.

Jak je vidět z tabulky 5.3, celkově nejlepších výsledků dosáhl detektor využívající parametrizaci PLP a adaptivní práh. Adaptivní práh byl obecně (ve všech prostředích) úspěšnější než práh na bázi sledování dynamiky průběhu kriteriální funkce. Tento rozdíl byl významný zejména v prostředích s vysokou úrovní šumu (viz. spodní část grafu 5.1).

Z uvedených výsledků je dále možné pozorovat, že nejproblematičtějšími prostředími byly pro detektor prostředí CAFE a HOME, a to i při nízké úrovni šumu (vysokém SNR).

typ detektoru	kritérium	SNR [dB] +15/+10	SNR [dB] +5/0	SNR [dB] -5/-10	prům. hodnota
MFCC kepstrum DYN práh	HTER [%]	19,55	27,42	42,00	29,66
	MR [%]	18,55	21,18	37,00	25,57
	FAR [%]	20,55	33,67	47,00	33,74
MFCC kepstrum ADAPT práh	HTER [%]	17,57	25,08	36,27	26,30
	MR [%]	18,95	29,17	46,24	31,45
	FAR [%]	16,19	20,98	26,29	21,15
PLP kepstrum DYN práh	HTER [%]	19,30	27,51	42,70	29,84
	MR [%]	18,74	21,80	36,63	25,72
	FAR [%]	19,87	33,21	48,77	33,95
PLP kepstrum ADAPT práh	HTER [%]	18,60	24,77	36,00	26,11
	MR [%]	30,87	31,89	46,42	33,06
	FAR [%]	14,33	17,64	25,52	19,16
LPC kepstrum DYN práh	HTER [%]	24,23	31,50	39,27	31,67
	MR [%]	23,09	32,91	37,04	31,01
	FAR [%]	25,37	30,09	41,51	32,33
LPC kepstrum ADAPT práh	HTER [%]	18,92	25,36	35,26	26,51
	MR [%]	16,83	25,03	37,73	26,52
	FAR [%]	21,02	25,69	32,80	26,50

Tabulka 5.3: Souhrnné výsledky s nastavením optimálním pro extrakci řečových příznaků (průměr z dílčích výsledků pro jednotlivé typy šumu).



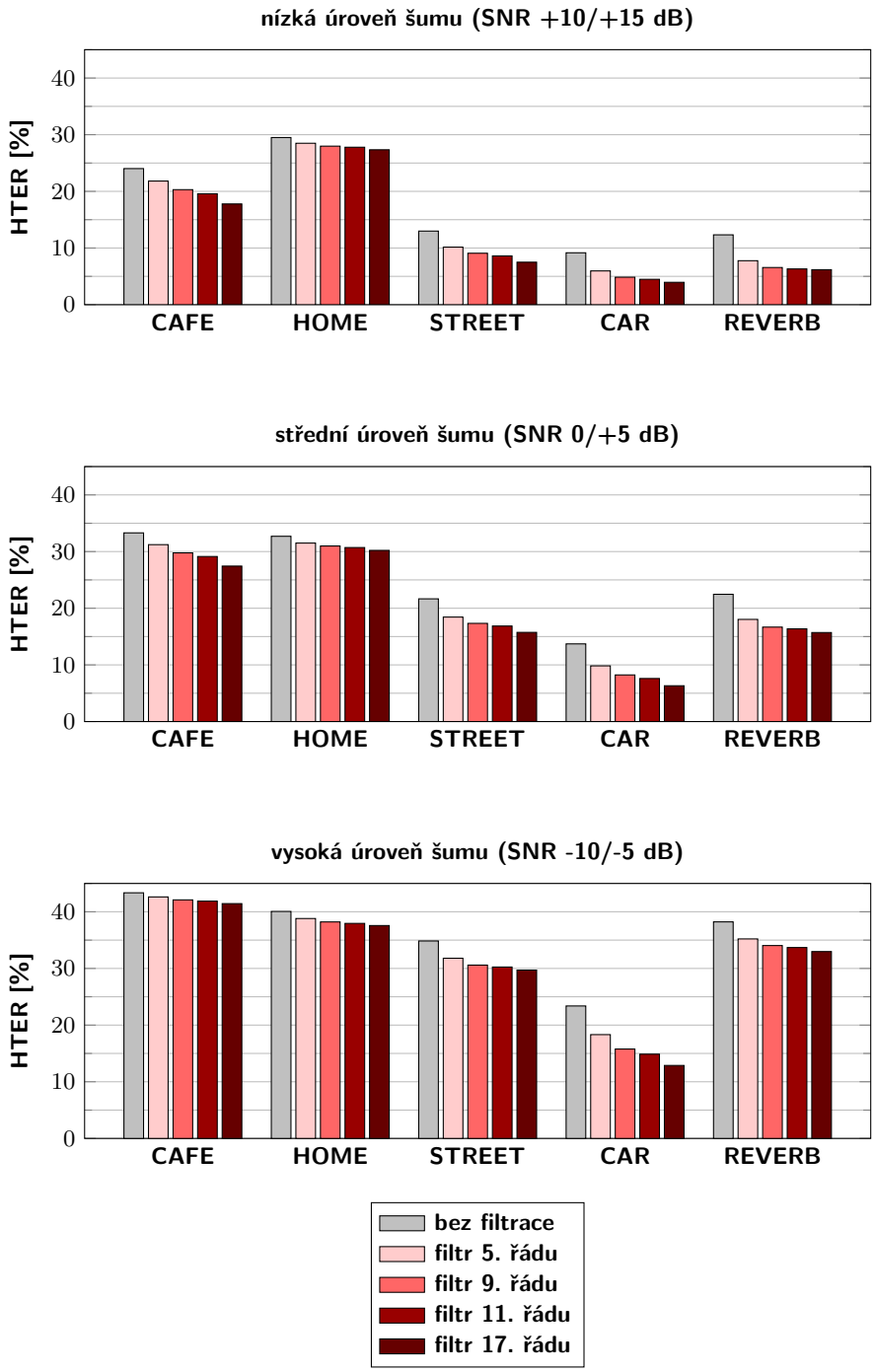
Obrázek 5.1: Dosažené výsledky detektoru při hodnotách parametrů optimálních pro extrakci řečových příznaků

5.4.2 Vliv mediánové filtrace na VAD detekci

Při druhém experimentu byl zkoumán vliv mediánové filtrace (viz. sekce 3.1.4) filtrem různého řádu na úspěšnost výsledné detekce. Pro tento experiment byl zvolen detektor na bázi kepstrální vzdálenosti s PLP řečovou parametrizací a adaptivním prahem, který v prvním experimentu dosáhl celkově nejlepších výsledků. Byly porovnávány výsledky detektoru bez filtrace s výsledky s mediánovou filtrací 5., 9., 11. a 17. řádu (při stejných hodnotách všech ostatních parametrů detektoru). Výsledky na úrovni HTER jsou zobrazeny v grafu 5.2. Z uvedených výsledků je zřejmé, že velikost mediánového filtru má velký vliv na úspěšnost detekce zejména v případech nižší chybovosti původní (nefiltrované) detekce. V případech vyšší chybovosti původní detekce (HTER = 30 % a výše) dochází spíše k stabilizaci výsledků a není účinné zařazovat filtr vyššího řádu. V tabulce 5.4 jsou zobrazeny souhrnné výsledky získané průměrem z hodnot pro všechna prostředí. Z uvedených čísel je zřejmé, že mediánová filtrace byla účinná zejména na úrovni kritéria FAR, zatímco hodnota kritéria MR byla filtrem velmi málo ovlivněna. V prostředích s vysokou úrovní šumu zařazení mediánové filtrace výsledek na úrovni kritéria MR dokonce zhoršilo.

řád mediánového filtru	kritérium	SNR [dB] +15/+10	SNR [dB] +5/0	SNR [dB] -5/-10	prům. hodnota
bez filtrace	HTER [%]	17,60	24,76	35,97	26,11
	MR [%]	20,87	31,90	46,42	33,06
	FAR [%]	14,33	17,64	25,52	19,17
filtr 5. řádu	HTER [%]	14,84	21,80	33,35	23,33
	MR [%]	20,25	31,59	48,55	33,46
	FAR [%]	9,42	12,02	18,15	13,19
filtr 9. řádu	HTER [%]	13,75	20,60	32,14	22,17
	MR [%]	19,13	30,43	48,52	32,70
	FAR [%]	8,39	10,77	15,77	11,64
filtr 11. řádu	HTER [%]	13,35	20,13	31,73	21,74
	MR [%]	18,54	29,76	48,32	32,21
	FAR [%]	8,16	10,51	15,14	11,27
filtr 17. řádu	HTER [%]	12,55	19,08	30,92	20,85
	MR [%]	17,20	28,08	47,74	31,01
	FAR [%]	7,89	10,09	14,09	10,69

Tabulka 5.4: Souhrnné výsledky detektoru s mediánovým filtrem (průměr z dílčích výsledků pro jednotlivé typy šumu).



Obrázek 5.2: Srovnání výsledků MFCC detektoru s ADAPT prahem pro různé velikosti mediánového filtru

Kapitola 6

Závěr

V rámci této bakalářské práce byl do nástroje CtuCopy přidán modul detektoru řečové aktivity pracující na bázi kepstrální vzdálenosti jako kriteriální funkce s možností volby z několika různých řečových parametrizací, dvou různých metod heuristického prahování a s možností následné filtrace detekce mediánovým filtrem. Tím bylo umožněno programu CtuCopy ukládat při extrakci řečových příznaků data odpovídající pouze řečovým segmentům signálu. Pro ověření správné funkčnosti detektoru bylo jeho chování zkoumáno v různých akustických prostředích s různými úrovněmi SNR s využitím upravené anglicky mluvené signálové databáze `QUT-NOISE-TIMIT`. Výsledky různých konfigurací detektoru byly porovnávány na základě kritéria HTER standardně používaného při evaluaci VAD detektrorů. Celkově nejlepších výsledků při téměř všech úrovních SNR dosáhl detektor na bázi kepstrální vzdálenosti z PLP řečové parametrizace a adaptivního prahování sledujícího změny v charakteristikách pozadí. S touto konfigurací byl zkoumán vliv mediánové filtrace na úspěšnost detekce při různé velikosti mediánového filtru. Bylo zjištěno, že mediánová filtrace má pozitivní vliv na úspěšnost detekce zejména v případech nižší chybovosti původní (nefiltrované) detekce, a to až do mediánového filtru řádu 17.

Literatura

[1] DARPA TIMIT. Acoustic-phonetic continuous speech corpus cd-rom. In *Document NISTIR 4930, NIST Speech Disk 1-1.1.*

[2] DEAN, D., SRIDHARAN, S., VOGT, R., MASON, M. The QUT-NOISE-TIMIT Corpus for the evaluation of voice activity detection algorithms. In *Proceedings of Interspeech 2010, 26-30 September 2010.* Makuhari Messe International Convention Complex, MAkuhari, Japan.

[3] FOUSEK, P. *Předzpracování řeči s šumovým pozadím pro účely komunikace a rozpoznávání.* Diplomová práce. Praha, 2002. České vysoké učení technické v Praze.

[4] FRIGO, M., JOHNSON, S. G. *FFTW. Manual for FFTW version 3.3.3, 25 November 2012.*. Massachusetts, 2012. Massachusetts Institute of Technology.

[5] HERMANSKY, H. Perceptual linear predictive (PLP) analysis of speech. In *The Journal of the Acoustical Society of America 87, 1738 (1990).* 1990. Acoustical Society of America.

[6] KOSEK, M., MIZERA, P. Analysis of Cepstral Voice Activity Detector in various acoustic conditions. In *Proceedings of the International Student Scientific Conference Poster – 21/2017.* Praha 2017. České vysoké učení technické v Praze.

[7] MIZERA, P. *Pitch-synchronní segmentace řečového signálu.* Diplomová práce. Praha, 2012. České vysoké učení technické v Praze.

[8] POLLÁK, P. *Criteria for VAD classification. Internal research report #R02-1.* Praha, 2002. České vysoké učení technické v Praze.

[9] POLLÁK, P. *A2M31ZRE - Zpracování řeči.* Studijní materiály k předmětu. [online]. Dostupné z: https://moodle.fel.cvut.cz/course/view.php?id=1807

[10] POLLÁK, P., RAJNOHA, J. *Long Recording Segmentation Based on Simple Power Voice Activity Detection with Adaptive Threshold and Post-Processing.* St. Petersburg, 2009. SPECOM'2009.

[11] PSUTKA, J. *Mluvíme s počítačem česky.* Praha: Academia, 2006. Česká matice technická (Academia). ISBN 80-200-1309-1.

[12] RABINER, L. R., SCHAFER, R. W. *Introduction to digital speech processing.* Boston, Mass.: Now, c2007. ISBN 978-1-60198-070-0.

[13] RAJNOHA, J., POLLÁK, P. Detektory řečové aktivity na bázi perceptivní kepstrální analýzy. In *Technical Computing Prague 2008 [CD-ROM].* Praha: Humusoft, 2008, díl 1, s. 1-9. ISBN 978-80-7080-692-0.

[14] RAJNOHA, J. *Rozpoznávání řeči v reálných podmínkách na platformě standardního PC.* Diplomová práce. Praha, 2006. České vysoké učení technické v Praze.

[15] RAMIREZ, J., GORRIZ, J. M., SEGURA, J. C. *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness, Robust Speech Recognition and Understanding, Michael Grimm and Kristian Kroschel (Ed.).* 2007. I-Tech Education and Publishing. ISBN: 978-3-902613-08-0.

[16] ROSCA, J., BALAN, R., FAN, N.P., BEAUGEANT, C., GILG, V. Multichannel Voice Detection in Adverse Environments. In *Proceedings of European Signal Processing Conference.* Toulouse, France, 2002. ISSN 2219-5491.

[17] SOHN, J., KIM, N. S., SUNG, W. A statistical model-based voice activity detection. In *IEEE signal processing letters, vol. 6, no. 1, 1–3, 1999.*

[18] *Speech Processing and Signal Analysis Group* [online]. [cit. 2017-03-23]. Dostupné z: http://noel.feld.cvut.cz/speechlab/start.php?page=download&lang=en

[19] TATARINOV, J. *Detektory řečové aktivity na bázi skrytých Markovových modeløu.* Disertační práce. Praha, 2010. České vysoké učení technické v Praze.

[20] TSIARTAS, A., CHASPARI, T., KATSAMANIS, N., GHOSH, P. K., LI, M., VAN SEGBROECK, M., POTAMIANOS, A., NARAYANAN, S. Multi-band long-term signal variability features for robust voice activity detection. In *Proceedings of INTERSPEECH 2013.* 2013, 718-–722.

[21] UHLÍŘ, J. *Technologie hlasových komunikací.* Praha: Nakladatelství ČVUT, 2007. ISBN 978-80-01-03888-8.

[22] VONDRÁŠEK, M. POLLÁK, P. *Methods for Speech SNR estimation: Evaluation Tool and Analysis of VAD Dependency.* Praha, 2005. České vysoké učení technické v Praze.

[23] WISDOM, S. OKOPAL, G. ATLAS, L. PITTON, J. *Voice Activity Detection Using Subband Noncircularity.* 4505-4509. 10.1109/ICASSP.2015.7178823.

[24] WU, J., ZHANG, X.-L. Efficient multiple kernel support vector machine based voice activity detection. In *IEEE Signal Processing Letters, vol. 18, no. 8, 466–469, 2011.*

[25] ZÁRUBA, M. *Moderní metody rozpoznávání mluvčího na bázi GMM a DNN.* Diplomová práce. Praha, 2017. České vysoké učení technické v Praze.

Příloha A

Obsah přiloženého CD

Na přiloženém CD jsou k dispozici následující data:

- Text této bakalářské práce ve formátu PDF:

  `Implementace_kepstralniho_detektoru_recove_aktivity_pri_vypoctu_recovych_priznaku.pdf`

- Zdrojové soubory poslední verze programu CtuCopy:

  `ctucopy.tgz`
  `ctucopy.tgz.md5`

  Aktuální verze programu CtuCopy je dostupná v následujícím GIT repozitáři:

  `https://github.com/pmizera/ctucopy.git`

- Zdrojové soubory databáze QUT-NOISE-TIMIT s úpravou umožňující generovat signály obsahující inicializační neřečovou fázi:

  `QUT-NOISE-speechstartrange.tgz`
  `QUT-NOISE-speechstartrange.tgz.md5`

  Dostupné též na branchi `speechstartrange` v následujícím GIT repozitáři:

  `https://github.com/mchlksk/QUT-NOISE.git`