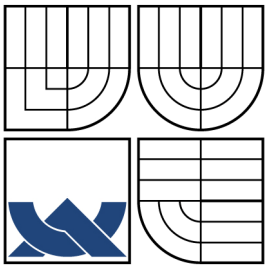
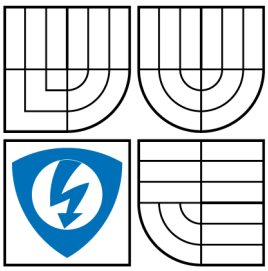


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

IDENTIFIKACE PAUZ V RUŠENÉM ŘEČOVÉM SIGNÁLU

PAUSE IDENTIFICATION IN DEGRADED SPEECH SIGNAL

Diplomová práce
MASTER´S THESIS

AUTOR PRÁCE — Bc. LENKA PODLOUCKÁ
AUTHOR

VEDOUCÍ PRÁCE — Prof. Ing. ZDENĚK SMÉKAL, CSc.
SUPERVISOR

BRNO 2008

LICENČNÍ SMLOUVA

POSKYTOVANÁ K VÝKONU PRÁVA UŽÍT ŠKOLNÍ DÍLO

uzavřená mezi smluvními stranami:

1. Pan/paní

Jméno a příjmení:Bc. Lenka Podloucká

Bytem: Revoluční 1702, Rožnov pod Radhoštěm, 756 61

Narozen/a (datum a místo): 19. února 1983, Valašské Meziříčí

(dále jen „autor“)

a

2. Vysoké učení technické v Brně

Fakulta elektrotechniky a komunikačních technologií

se sídlem Údolní 244/53, 602 00 Brno, Česká Republika,

jejímž jménem jedná na základě písemného pověření děkanem fakulty:

Prof. Ing. Radimír Vrba, CSc.

(dále jen „nabyvatel“)

Článek 1

Specifikace školního díla

1. Předmětem této smlouvy je vysokoškolská kvalifikační práce (VŠKP) *:

□ disertační práce

□ diplomová práce

□ bakalářská práce

□ jiná práce, jejíž druh je specifikován jako .....

(dále jen VŠKP nebo dílo)

Název VŠKP:	Identifikace pauz v rušeném řečovém signálu
Vedoucí/ školitel VŠKP:	Prof. Ing. Zdeněk Smékal, CSc.
Ústav:	Telekomunikací
Datum obhajoby VŠKP:	

VŠKP odevzdal autor nabyvateli v:

tištěné formě – počet exemplářů 2

elektronické formě – počet exemplářů 2

* hodící se zaškrtněte

2. Autor prohlašuje, že vytvořil samostatnou vlastní tvůrčí činností dílo shora popsané a specifikované. Autor dále prohlašuje, že při zpracovávání díla se sám nedostal do rozporu s autorským zákonem a předpisy souvisejícími a že je dílo dílem původním.
3. Dílo je chráněno jako dílo dle autorského zákona v platném znění.
4. Autor potvrzuje, že listinná a elektronická verze díla je identická.

Článek 2
Udělení licenčního oprávnění

1. Autor touto smlouvou poskytuje nabyvateli oprávnění (licenci) k výkonu práva uvedené dílo nevýdělečně užít, archivovat a zpřístupnit ke studijním, výukovým a výzkumným účelům včetně pořizovaní výpisů, opisů a rozmnoženin.
2. Licence je poskytována celosvětově, pro celou dobu trvání autorských a majetkových práv k dílu.
3. Autor souhlasí se zveřejněním díla v databázi přístupné v mezinárodní síti *
   - □ ihned po uzavření této smlouvy
   - □ 1 rok po uzavření této smlouvy
   - □ 3 roky po uzavření této smlouvy
   - □ 5 let po uzavření této smlouvy
   - □ 10 let po uzavření této smlouvy

   (z důvodu utajení v něm obsažených informací)
4. Nevýdělečné zveřejňování díla nabyvatelem v souladu s ustanovením § 47b zákona č. 111/ 1998 Sb., v platném znění, nevyžaduje licenci a nabyvatel je k němu povinen a oprávněn ze zákona.

Článek 3
Závěrečná ustanovení

1. Smlouva je sepsána ve třech vyhotoveních s platností originálu, přičemž po jednom vyhotovení obdrží autor a nabyvatel, další vyhotovení je vloženo do VŠKP.
2. Vztahy mezi smluvními stranami vzniklé a neupravené touto smlouvou se řídí autorským zákonem, občanským zákoníkem, vysokoškolským zákonem, zákonem o archivnictví, v platném znění a popř. dalšími právními předpisy.

* hodící se zaškrtněte

3. Licenční smlouva byla uzavřena na základě svobodné a pravé vůle smluvních stran, s plným porozuměním jejímu textu i důsledkům, nikoliv v tísni a za nápadně nevýhodných podmínek.
4. Licenční smlouva nabývá platnosti a účinnosti dnem jejího podpisu oběma smluvními stranami.

V Brně dne: 28. května 2008

…….. ……
Nabyvatel Autor

ANOTACE

Tato diplomová práce se zabývá identifikací pauz v rušeném řečovém signálu. Je zde popsán charakter řečového signálu a koncepce jeho zpracování.

Cílem diplomové práce bylo navrhnout metodu spolehlivého určení úseků bez řečové aktivity (pauz) jak pro řeč bez přítomnosti šumu a rušení, tak i ze směsi řeči a nežádoucího rušení. Pro identifikaci pauz bylo realizováno pět detektorů v programovém prostředí MATLAB. V časové oblasti to byl energetický detektor, ve spektrální oblasti dvoukrokový detektor využívající v prvním kroku energetické vlastnosti signálu, ve druhém výpočtu statistických veličin. V kepstrální oblasti byly realizovány tři detektory, dva s využitím integrálního algoritmu a detekce třetího byla založena na diferenciálním algoritmu.

Robustnost detektorů byla testována pro různé typy rušení a jejich úrovně odstupu signálu od šumu (Signal to Noise Ratio – SNR). Pro vyhodnocení úspěšnosti detekce byly sestaveny ROC křivky, ve kterých byl měnícím se parametrem rušivý signál.

Klíčová slova: řečový signál, detekce pauz, detektory řečové aktivity, ROC charakteristiky, odstup signálu od šumu SNR.

ABSTRACT

This diploma thesis deals with pause identification with degraded speech signal. The speech characteristics and the conception of speech signal processing are described here.

The work aim was to create the reliable recognizing method to establish speech and non-speech segments of speech signal with and without degraded speech signal. The five empty pause detectors were realized in computing environment MATLAB. There was the energetic detector in time domain, two-step detector in spectral domain, one-step integral detector, two-step integral detector and differential detector in cepstrum. The spectral detector makes use of energetic characteristics of speech signal in first step and statistic analysis in second step. Cepstral detectors make use of integral or differential algorithms.

The detectors robustness was tested for different types of speech degradation and different values of Signal to Noise Ratio. The test of influence different speech degradation was conducted to compare non-speech detection for detectors by ROC (Receiver Operating Characteristic) Curves.

Key words: speech signal, empty pause detection, voice activity detectors, ROC Curves, Signal to Noise Ratio

ČESTNÉ PROHLÁŠENÍ

Prohlašuji, že jsem tuto diplomovou práci vypracovala samostatně pod vedením Prof. Ing. ZDENEKA SMÉKALA, CSc. Všechny literární prameny a publikace, ze kterých jsem čerpala, jsou řádně uvedeny buď přímo v textu, nebo v závěrečném seznamu literatury

……..
(podpis autora)

PODĚKOVÁNÍ

Ráda bych touto cestou poděkovala panu Prof. Ing. Zdeňku Smékalovi, CSc. za odbornou i lidskou podporu a trpělivost při vypracování této diplomové práce a panu Ing. Vojtěchu Stejskalovi za pomoc při vyhledávání literárních pramenů.

Dále bych ráda poděkovala svému zaměstnavateli za snahu o vytvoření co nejlepších podmínek pro vypracování a díky také patří nejbližším za ohromnou morální podporu nejen během vypracování této práce, ale též během celého studia.

SEZNAM ZKRATEK

DD Diferenciální kepstrální detektor
DFT Diskrétní Fourierova transformace
ED Energetický detektor
FAR0 Pravděpodobnost planého poplachu
FFT Rychlá Fourierova transformace
HR0 Pravděpodobnost správné detekce segmentu řeč neobsahující
HR1 Pravděpodobnost správné detekce řečového rámce
IDFT Inverzní diskrétní Fourierova transformace
PCM Pulzní kótová modulace
ROC Receiver Operating Characteristic
SD Spektrální detektor
SNR Poměr signálu a šumu
1ID Jednokrokový integrální kepstrální detektor
2ID Dvojkrokový integrální kepstrální detektor

OBSAH

1. ÚVOD

Komunikace prostřednictvím mluvené řeči (řečového signálu) je nejdůležitějším prostředkem pro přenos informace mezi lidmi. Tento přenos obvykle začíná přípravou zprávy v mozku řečníka a končí jejím rozpoznáním posluchačem.

V dnešní době se lidstvo zabývá myšlenkou, jak vytvořit systém, který by umožňoval komunikaci mezi člověkem a strojem. Pro automatické rozpoznávání a syntézu řečového signálu se většinou používá řečový signál, který není znehodnocen hlukem, šumem nebo jiným rušením. Typ rušení se většinou identifikuje během pauz v řeči, které je nutné nalézt s dostatečnou přesností. Úspěšná detekce pauz není důležitá pouze při rozpoznávání řeči, ale i dalších aplikacích zpracování řečového signálu, jako je např. kódování nebo detekce emočních stavů mluvčího.

Cílem této diplomové práce je navrhnout účinnou metodu pro označení pauz v řeči z jednokanálového záznamu.

V jednotlivé kapitoly seznamují čtenáře s charakterem řečového signálu, koncepcí jeho zpracování, jak byla aplikována při řešení této práce a vlastními experimenty pro testování robustnosti realizovaných detektorů.

2. CHARAKTER ŘEČOVÉHO SIGNÁLU

Lidská řeč je charakterizována jistou akustickou strukturou (amplitudově-frekvenčním časovým spektrem), lingvistickou strukturou (gramatikou a skladbou) a subjektivním vlivem osobnosti řečníka (intonace, rytmus, barva hlasu atd.)

Řeč může být reprezentována buď jejím informačním obsahem nebo jako fyzikální řečový signál, který slouží coby nositel informace. Pro automatické rozpoznávání řeči (tedy i identifikaci pauz) má zásadní význam řeč ve formě řečového signálu.

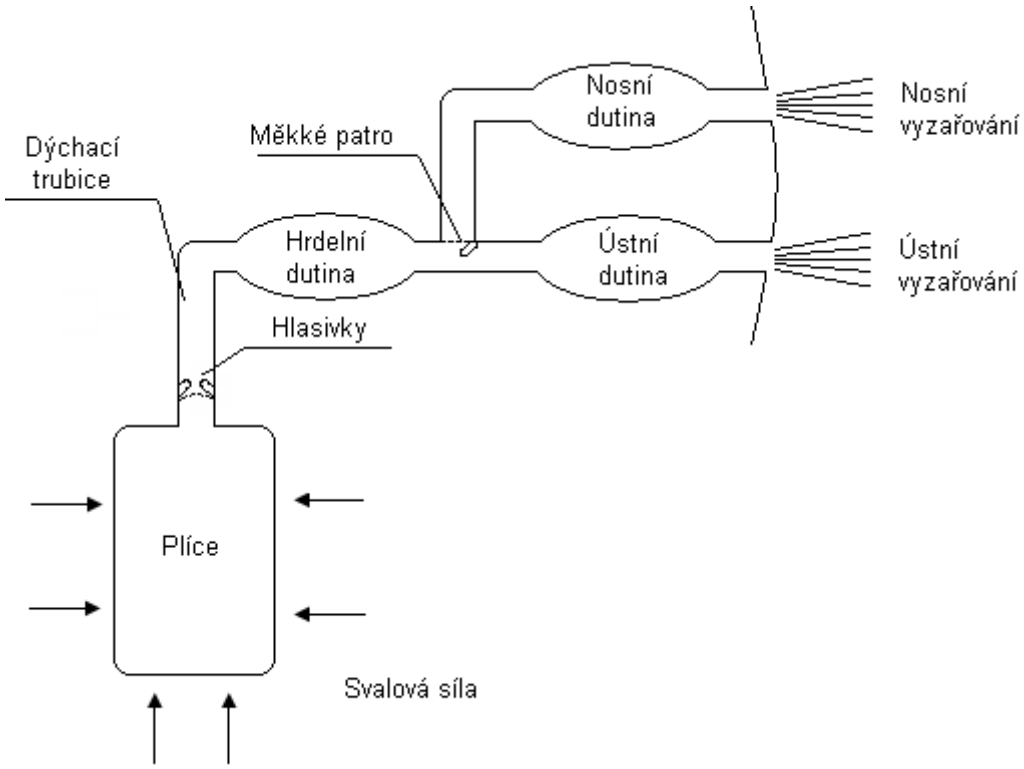
2.1 Proces vytváření řeči člověkem

Zdrojem řečových kmitů jsou lidské řečové orgány (hlasivky, dutina hrdelní, ústní a nosní, měkké a tvrdé patro, zuby a jazyk). Hnací silou celého procesu vytváření řeči je proud vzduchu dodávaný plícemi. Ten dále prochází úzkou mezerou v horní části hrtanu, tzv. hlasivkovou štěrbinou. Hlasivková štěrbina je obklopená hlasivkami, které se rozkmitají. Frekvence kmitů závisí jednak na tlaku vzduchu a jednak na svalovém napětí hlasivek. Tato frekvence je různá u dětí a dospělých, u žen a mužů. U většiny lidí se však pohybuje v rozmezí 150 až 400 Hz. Frekvence kmitů hlasivek F_0 charakterizuje základní tón lidského hlasu. Ten je přítomen při tvoření všech znělých zvuků, tj. samohlásek a znělých souhlásek.

Kmitáním hlasivek vznikají vzduchové rázy v intervalech přibližně 10 ms. Takto vytvořené signály jsou v následujícím hlasovém traktu – dutině hrdelní, ústní a nosní pohybem úst a jazyka zpracovány. Výsledkem jsou hlásky, které se dále šíří vzduchem k posluchači kulovitých plochách, proto energie řečového signálu směrem od mluvčího velmi rychle klesá. Schématicky zjednodušený model hlavních částí hlasového ústrojí podílejících se na vytváření řeči je znázorněn na obr. 2.1

2.2 Fonetický popis řeči

Na souvislou řeč lze pohlížet jako na časovou posloupnost jednotlivých zvuků. Pro potřebu zkoumání řeči z jejích různých stránek byly zavedeny fonologické jednotky v následujícím hierarchickém členění:



Obr. 2.1 Schématické znázornění lidského hlasového ústrojí

promluva → fráze → slova → slabiky → difony → fonémy → alofóny

Promluva označuje celek řeči mezi dvěma absolutními pauzami.

Fráze je obvykle menší než promluva, může být však i minimální promluvou. Z fonetického hlediska je vyznačena větnou intonací a relativními pauzami. Je nositelkou intonačních prostředků modulace řeči.

Za nejmenší jednotku řeči, která může rozlišovat jednotlivá slova, lze považovat *foném*. Fonémy lze od sebe rozlišit z několika hledisek, například podle způsobu a místa tvoření, podle artikulujícího orgánu nebo podle sluchového dojmu. Každý světový jazyk má vlastní soupis fonémů, jejich počet se pohybuje od 12 do 60, přičemž některé fonémy nalezneme shodné téměř ve všech jazycích. Český jazyk obsahuje ve spisovné podobě 36 fonémů, angličtina 44, němčina 40, ruština 40, apod.

Jednotlivé fonémy jsou v mluvené řeči spojovány do posloupností mluvených celků, v nichž lze nalézt další stavební jednotku – *slabiku*. Slabiku je možné již přesně srovnávat s psanou formou. Každé *slovo* je poté zastoupeno určitou

kombinací slabik, přičemž jejich počet tvoří vždy celé číslo. Evropské jazyky používají 2005 – 3005 slabik a 45 000 – 50 000 slov.

Difony představují přechodné zvuky a lze s nimi popisovat koartikulační vliv dané hlásky na hlásku předchozí a následující. Difony tedy lépe popisují akustický signál nežli fonémy, avšak z průměrného počtu 40 fonémů vzniká přibližně 1000 difonů, přičemž se reálně nevyskytují všechny kombinace fonémů.

Alofóny jsou nejmenší zvukově rozlišitelné jednotky řeči. Slouží k rozpoznání jednotlivých fonémů v promluvách. Bývají charakteristické pro danou situaci (slavnostní projev, běžný hovor) nebo pro mluvu jednotlivce (oblastní varianty, emotivní varianty, sociální varianty) apod. [2]

2.3 Základní řečové jednotky v oblastech počítačového zpracování řeči

Při automatickém zpracování řečového signálu se vedle výše zmíněných fonetických jednotek používají i další řečové jednotky. Tyto jednotky bývají většinou navrženy tak, aby co nejlépe vyhovovaly potřebám automatického zpracování řeči počítačem. Volba jednotky tedy záleží na typu řešené úlohy, je-li při zpracovávání řečového signálu informace, kterou signál nese, podstatná či ne.

V případě rozpoznávání a syntézy plynulé řeči se často používají jednotky, které dobře reprezentují závislosti jednotek i na okolním řečovém kontextu, těmi jsou vedle difonů například *trifony*.

Je-li cílem automatického zpracování řečového signálu kódování nebo identifikace pauz, kdy se zpracovává řečový signál bez ohledu na informaci, kterou nese, volí se jednotky nezávislé na fonetické reprezentaci. Jsou to krátké úseky signálu, nazývané *segmenty* či *mikrosegmenty*. Typická délka mikrosegmentu se pohybuje v rozmezí 20 – 30 ms.

2.4 Pauzy

Diplomová práce se zabývá identifikací pauz. Proto je tedy vhodné v rámci kapitoly zabývající se charakterem řečového signálu charakterizovat pojem pauza. Pauzy se dají považovat za jednu z prozodických vlastností řeči.

Prozodii lze studovat z několika hledisek, úrovní reprezentace prozodických jevů (akustická, percepční, lingvistická a artikulační úroveň reprezentace prozodických jevů). Akustická úroveň představuje akustickou realizaci prozodických jevů, parametry, které se k ní vztahují jsou frekvence základního hlasivkového tónu, intenzita (amplituda, energie) a doba trvání (časování, se kterým pauzy právě úzce souvisí). Percepční úroveň reprezentuje prozodické vlastnosti řeči tak, jak je vnímá průměrný posluchač. Jevy se zde nazývají výška hlasu (melodie), hlasitost a délka segmentů řeči. Lingvistická úroveň představuje prozodii promluvy jako posloupnost abstraktních jednotek, tuto úroveň nelze měřit, charakteristiky této úrovně se vztahují k tónu, intonaci a přízvuku. Artikulační úroveň prozodie reprezentuje prozodii jako posloupnost fyzikálních pohybů mluvidel. [1]

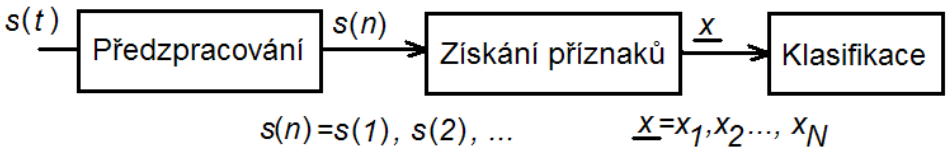
Pauzy jsou části řečového signálu, které ovlivňují časování i srozumitelnost řeči. Rozdělují jednotlivé větné úseky či celé věty. V těchto úsecích se řečník obvykle nadechne, nádech se tedy projeví jako pauza. Nádech může být slyšitelný jako šumový zvuk, který může doprovázet polknutí, mlasknutí, atd. nebo může být tichý. Délka pauz rozdělujících větné úseky se může lišit v závislosti na syntaktické vazbě, jejich počet na emočním stavu řečníka. Délka se může nabývat velikosti od 500 ms výše. [3]

Při rozpoznávání řeči je pauzou považován časový úsek, který v řeči odděluje jednotlivé difony či trifony. Může odpovídat času, který je potřeba ke změně hlasového ústrojí pro vytvoření fonému, čí délce mezi dvěma slabikami. Délka pauzy je zde výrazně menší než u pauz rozdělujících větné úseky. Může se pohybovat již okolo 30 ms. Při řešení diplomové práce byla za minimální délku pauzy považována doba 100 ms, kterou uvádí [3].

Různé může být i zastoupení pauz v mluvené řeči. Zatímco pauzy tvoří až polovinu spontánní konverzační řeči, u čtené řeči je to asi 20%. Jestliže člověk mluví pomaleji než normálně, počet pauz a jejich prodlužování se zvyšuje. Naopak při rychlé řeči se pauzy zkracují a klesá jejich počet. [1]

3. KONCEPCE ZPRACOVÁNÍ ŘEČOVÉHO SIGNÁLU

Systémy zpracovávající řečový signál (např. systémy pro identifikaci pauz, filtraci rušivých signálů ze zašumnělé řeči, jejího rozpoznávání, rozpoznávání mluvčích, jejich emočních stavů, případně rozpoznávání dalších jevů) odpovídají určitým zákonitostem obecného procesu rozpoznávání signálů. Tento proces je znázorněn na obr. 3.1.



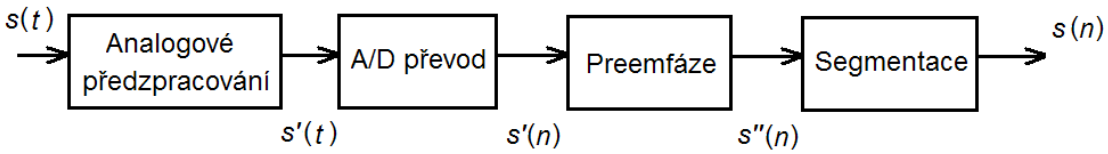
Obr. 3.1 Koncepce zpracování řečových signálů

Prvním blokem je předzpracování. Pod tímto blokem si lze představit převod akustického signálu do digitální podoby a několik standardních operací pro úpravu navzorkovaného signálu. Navzorkovaný signál $s(n)$ však stále nese spoustu „nezajímavých“ informací pro další zpracování, také různost a počet vzorků je značně vysoká, je tedy potřeba získat jen několik důležitých příznaků $\underline{x}$ pro dané zpracování. Úkolem bloku klasifikace je na základě příznaků signál roztřídit dle požadovaného zadání. Funkce jednotlivé bloky, které byly využity při řešení této diplomové práce, budou popsány v následujících kapitolách.

3.1 Předzpracování

Řečový signál je značně variabilní a prakticky není možné vyslovit jedno slovo dvakrát naprosto stejně, tzn. dodržet přízvuk, výšku tónu, hlasitost, či rychlost promluvy. Důležitým parametrem ovlivňujícím kvalitu řečového signálu je rušení, okolní zvuky a zkreslení způsobeném následným zpracováním nebo přenosem signálu (např.: kmitočtové charakteristiky mikrofonů, filtrů a zesilovačů či přenosových cest).

Všechny tyto vlivy snižují celkovou úspěšnost rozpoznávání a analýzy řeči. Proto je účelné některé vlivy potlačit vhodným předzpracováním na počátku celého procesu.[2] Blokový diagram předzpracování řečového signálu je uveden na obr. 3.2.



Obr. 3.2 Blokový diagram operací předzpracování řeči

3.1.1 Analogové předzpracování

Analogovým předzpracováním se rozumí manipulace s řečovým signálem do té doby, než bude reprezentován sledem vzorků. Nejprve se řeč snímá mikrofonem (převod změn akustického tlaku na elektrický signál). Mikrofon by měl zajistit velmi dobrý odstup řeči od zvuků v pozadí. Takto získaný elektrický signál se nachází v rozsahu pouze několika mV a proto je zapotřebí jej zesílit. Dále je nutné omezit zesílený řečový signál dolní propustí, jejíž kmitočet musí být maximálně poloviční než je vzorkovací kmitočet A/D převodníku (dodržení Shannovova teorému).

3.1.2 Analogově-digitální převod

Pro další zpracování je zapotřebí analogové kmity převést do číslicového tvaru tak, aby spojitý signál byl reprezentován posloupností číselných údajů. Tento proces, nazývaný také *pulzní kódová modulace* (PCM) nebo digitalizace, se skládá ze dvou kroků – vzorkování a kvantování s následným kódováním. V dnešních počítačích se využívá PCM pro uložení nejjednodušších zvukových souborů typu `*.wav`, které byly vstupem pro praktickou část řešení této práce.

Vzorkování

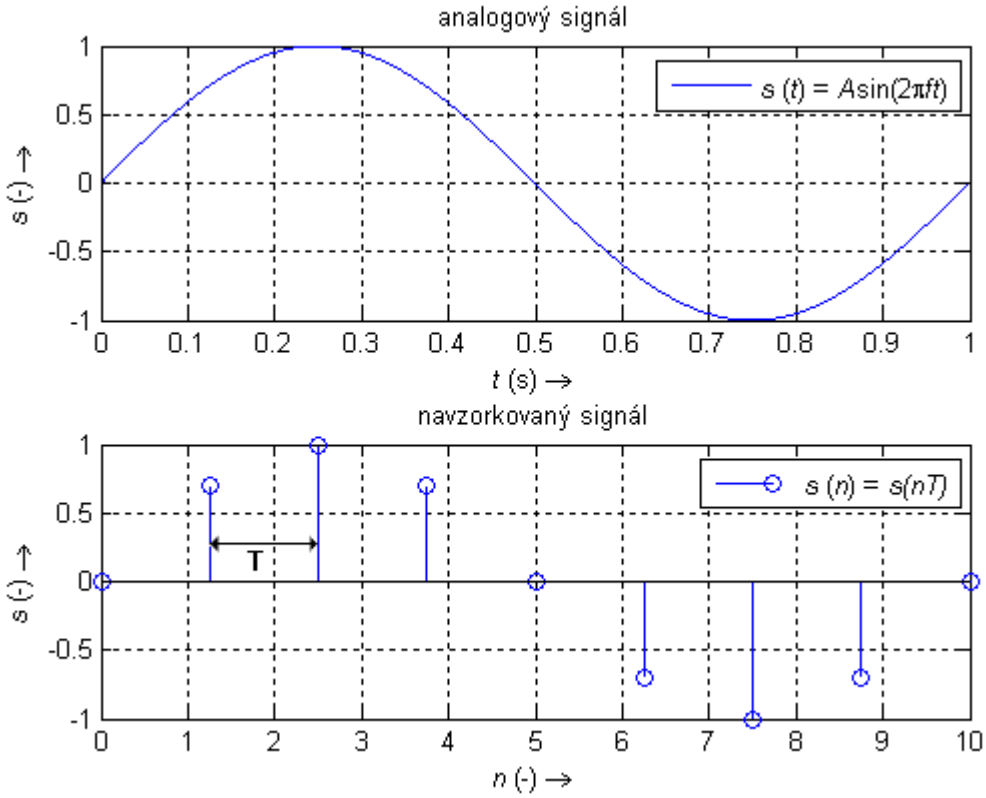
Vzorkováním se rozumí transformace signálu, kdy z původního *s‘(t)* spojitého v čase jsou získávána posloupnost vzorků *s‘*(*n*) diskrétních v čase. Vzorkování probíhá v přesně definovaných časových okamžicích t_n (3.1)

$$t_n = n \cdot T \,, \tag{3.1}$$

kde T je perioda vzorkování a n=0,…,∞. Vzorkovací kmitočet f_{vz} je poté definován jako převrácená hodnota vzorkovací periody T. Na Obr. 3.3 je zachycen signál spojitý v čase a navzorkovaný signál.

Při vzorkování analogového signálu s maximální frekvencí f_m musí být dodržen Shannonův teorém (3.2). Kdy hodnota vzorkovacího kmitočtu f_{vz} by měla být nejméně dvojnásobná oproti meznímu kmitočtu f_m vstupního signálu

$$f_{vz} \geq 2 \cdot f_m . \tag{3.2}$$



Obr. 3.3 Vzorkování – signál spojitý v čase a navzorkovaný signál pro amplitudu A=1

V opačném případě nastane tzv. aliasing (překrytí spekter), neboť krok vzorkování je provázen periodizací spektra původního signálu s periodou f_{vz} a v případě nedodržení vztahu (3.2) dojde k překrytí spektrer navzorkovaného signálu a v oblastech překrytí se tato spektra komplexně sčítají. Tento jev nám pak znemožní zpětně získat původní analogový signál ze série vzorků.

V praxi se před vzorkovač zařazuje speciální dolní propust, antialiasingový filtr. Ten odřezává spektrální složky signálu $s(t)$ ležící nad úhlovým kmitočtem $f_{vz}/2$. Reálné neideální vlastnosti antialiasingových filtrů jsou jedním z důvodů, proč v praxi dodržujeme podmínku (3.2) s jistou rezervou.

Kvantování a kódování

Kvantizace s následným kódováním je proces, při kterém je analogové hodnotě signálového vzorku přiřazena číselná hodnota. Každý vzorek je tedy vyjádřen *N*-bitovým slovem v některém z dvojkových kódů. Počet kvantovacích hladin *s*, který je definován podle vztahu (3.3)

$$s = 2^N, \tag{3.3}$$

kde *N* je délka převedeného bitového slova.

Při kvantizačním procesu dochází k určité ztrátě informace vlivem zaokrouhlování okamžitých velikostí signálu. Tato ztráta se nazývá kvantizační šum. Odstup signálu od kvantizačního šumu *SNR* se udává v decibelech pro *N*-bitový převod dán vztahem (3.4)

$$SNR = 6N - 7{,}24, \tag{3.4}$$

kde *N* je délka převedeného bitového slova.

Při zpracování či analýze řečového signálu je postačující vzorkovací frekvence 8 – 12 kHz a 8 – 10 bitů pro kvantování [2]. Pro vysoce kvalitní záznam řečového signálu se doporučuje 16 bitový převod a vzorkovací frekvence 16 - 22 kHz. V dnešní době se ustálily standardy vzorkovacích frekvencí 11 025 Hz, 22 050 Hz, 44 100 Hz (vzorkovací frekvence PC zvukových karet), 8 000 Hz (telefonní karty), 16 000 Hz (zvukové karty některých pracovních stanic pracujících pod operačním systémem Unix). [1]

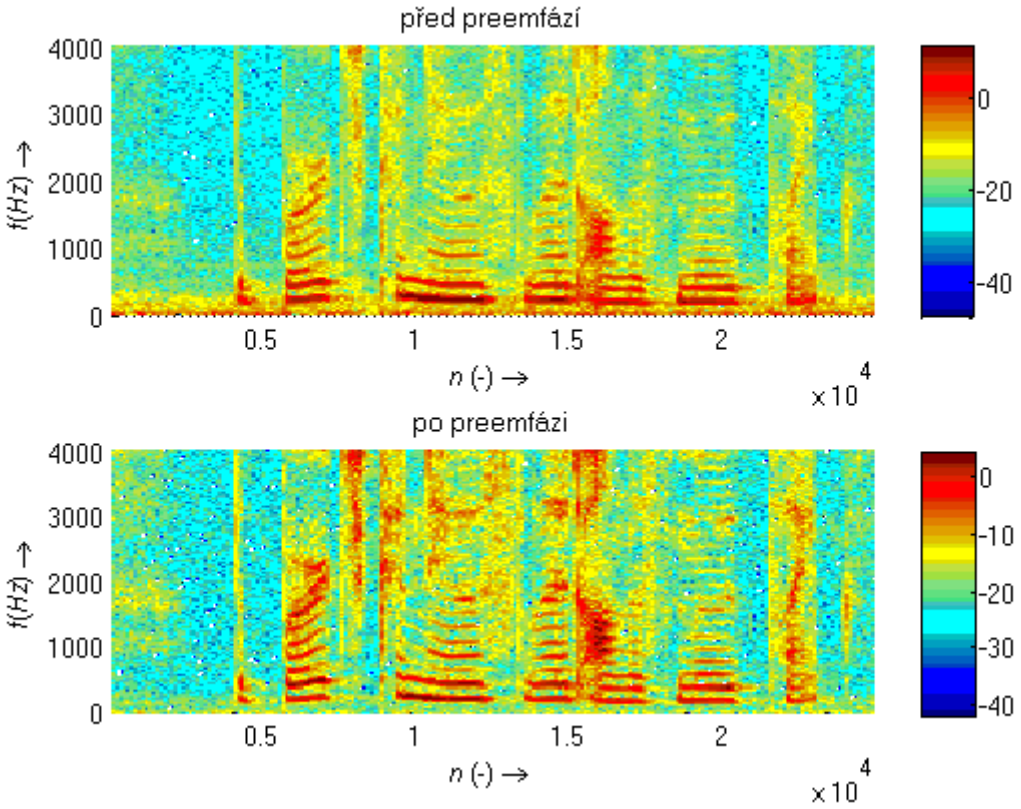
3.1.3 Preemfáze

Podstatná část energie řečového signálu je pod hranicí 300 Hz, ačkoliv užitečné informace v signálu jsou téměř kompletně obsaženy v pásmu nad 300 Hz. Vezmeme-li v úvahu, že kvantizační šum vykazuje rovnoměrné spektrum, je jeho negativní vliv podstatně větší na energeticky slabší, ale důležitější složky spektra řečového signálu. Filtrace prováděná před vážením segmentu na zdůraznění vyšších kmitočtů se nazývá preemfáze. Jedná se o filtraci prvního řádu a v časové oblasti je vyjádřena rovnicí (3.5)

$$s''(n) = s'(n) - \kappa s'(n-1), \tag{3.5}$$

kde *s*''(*n*) je signál po filtraci preemfázovým filtrem, *s*'(*n*) signál po analogově-digitálním převodu a konstanta κ nabývá hodnot od 0.9 do 1.

Vliv preemfáze na změnu energie ve spektru je znázorněn spektrogramy na obr. 3.4



Obr. 3.4 Spektrogramy řečového signálu před a po preemfázi

3.1.4 Segmentace pomocí oken

Vzhledem ke své povaze je řečový signál zpracováván metodami tzv. krátkodobé analýzy. Tyto metody vycházejí z předpokladu, že vlastnosti řečového signálu se v čase mění „pomalu“. Signál je tedy rozdělen na ekvidistantní časové úseky – segmenty o délce N vzorků, kde je řečový signál považován za stacionární.

Délka segmentu (rámce) musí být taková, aby bylo možné segment aproximovat, ale dostatečně dlouhý, aby bylo zaručeno, že požadované parametry budou bezchybně změřeny. Těmto požadavkům vyhovuje délka segmentu 20 až 30 ms, což souvisí se změnami nastavení lidského hlasového ústrojí, které probíhají v nejkratším časovém intervalu 10 až 25 ms.

Řečový segment $s(n)$ o N vzorcích je z řečového signálu $s''(n)$ po preemfázi vytvořen pomocí váhové posloupnosti tzv. okna $w(n)$ (3.6)

$$s(n) = s''(n) \cdot w(n)\,, \tag{3.6}$$

Úkolem okna je vybrat po úsecích příslušné vzorky signálu a přidělit jim určitou váhu. Váhová funkce $w(n)$ určuje typ okna. Nejčastěji se při zpracování řečového signálu používají obdélníkové a Hammingovo okno, pro které platí:

pravoúhlé	$w(n) = 1$	pro n = 1,2,…, N
	$w(n) = 0$	pro ostatní n
Hammingovo	$w(n) = 0{,}54 - 0{,}46\cos(2\pi n / N)$	pro n = 1,2,…, N
	$w(n) = 0$	pro ostatní n

kde π je Ludolfovo číslo.

Přestože pravoúhlé okénko je jednoduší, často se upřednostňuje použití Hammingova okna, vzhledem k tomu, že potlačuje vzorky na okrajích segmentů, čímž se zvyšuje stabilita některých výpočtů.

Při dělení signálu na segmenty se mohou jednotlivé segmenty překrývat. Malé nebo žádné překrytí zaručuje rychlý časový posun v signálu, menší nároky na paměť i procesor automatického zpracování, hodnoty parametrů jednotlivých rámců se mohou od sousedních výrazně měnit. Velké překrytí znamená pomalý časový posuv po signálu, vyšší nároky na paměť a procesor, průběhy parametrů se mění pozvolně, avšak mohou být pro sousední rámce velice podobné. Volí se tedy kompromis mezi velkým a malým překrytím. Typická délka je 10 ms, což při vzorkovací frekvenci 8 kHz znamená 100 segmentů za vteřinu.

3.2 Získání příznaků

Druhým blokem automatického zpracování řečového signálu je získání příznaků. Úkolem této části je oddělit a vyzvednout parametry řeči nesoucí vhodné charakteristické vlastnosti pro následnou klasifikaci, tedy získat určitý zjednodušený popis – příznaky.

Existují tři oblasti zobrazení charakteru signálu, které se používají při zpravování řeči. Těmi oblastmi jsou časová oblast, spektrum a kepstum. Cílem této práce bylo navrhnout několik detektorů v různých oblastech a porovnat jejich vlastnosti. V každé oblasti byl realizován alespoň jeden detektor, v následujících podkapitolách jsou popsány právě ty příznaky, na jejichž základě realizované detektory pauzy identifikují.

3.2.1 Zpracování v časové oblasti

V časové oblasti je řečový signál reprezentován posloupností vzorků. Většinu metod krátkodobé analýzy pro daný segment lze vyjádřit dle vztahu (3.7)

$$Q_n = \sum_{n=1}^{N} \tau(s[n]), \tag{3.7}$$

kde Q_n je krátkodobá charakteristika, $s[n]$ značí vzorek nasegmentovaného časového signálu v čase n, N je délka segmentu a $\tau(.)$ vyjadřuje příslušnou transformační funkci.

Nejčastějšími užívanými krátkodobými analýzami v časové oblasti pro zpracování řeči jsou krátkodobá energie (intenzita), krátkodobá funkce středního počtu průchodů signálu nulou a krátkodobá autokorelační funkce.

Energie signálu

Jedním z parametrů v časové oblasti, který byl využit při řešení, je energie E diskrétního řečového signálu $s(n)$, definovaná na jednom segmentu o délce N vztahem (3.8)

$$E = \frac{1}{N} \sum_{n=1}^{N} s^2[n], \tag{3.8}$$

Energie řeči je výrazně vyšší než energie signálu v pauzách. Proto může být použita jako parametr pro identifikaci řečové aktivity či pauz. Avšak segmenty řečového signálu, které obsahují neznělé hlásky, mohou mít rovněž nízkou energii, a z tohoto důvodu energetickým detektorem označeny jako části řečovou aktivitu neobsahující. Energie segmentu signálu je nejjednodušším kritériem pro identifikaci řečové přítomnosti od řečového klidu u signálů neobsahujících šum či jiné rušení.

3.2.2 Zpracování ve spektru

Spektrem signálu je nazýváno zobrazení signálu v kmitočtové oblasti, a to ve dvou rovinách (úhlový kmitočet - amplituda) a (úhlový kmitočet – počáteční fáze). Existuje velké množství typů spekter. Některá spektra dávají úplnou informaci o signálu, jiná jen částečnou. Pro výpočet spektra diskrétního signálu je se užívá diskrétní Fourierovy transformace (DFT), která přiřazuje posloupnosti délky N

navzorkovaného signálu *s*(*n*) jinou posloupnost *S*(*k*) stejné délky. [5] Výpočet spektra je uveden ve vztahu (3.9)

$$S(k) = \sum_{n-0}^{N-1} s(n) \exp\left(- j \frac{2\pi}{N} kn \right). \tag{3.9}$$

Pro získání posloupnosti *s*(*n*) z posloupnosti *S*(*k*) se využívá zpětné (inverzní) diskrétní Fourierovy transformace (IDFT), která se dá zapsat vztahem (3.10)

$$s(n) = \frac{1}{N} \sum_{k-0}^{N-1} S(k) \exp\left(j \frac{2\pi}{N} nk \right). \tag{3.10}$$

Veličina *k* v obou vztazích je pořadové číslo spektrální složky. Normovaný úhlový kmitočet ω_k k této složce je dán vztahem (3.11)

$$\omega_k = \frac{2\pi}{N} k \,. \tag{3.11}$$

Vztah mezi úhlovým kmitočtem ω a frekvencí *f* je dán (3.12)

$$\omega = 2\pi f \,. \tag{3.12}$$

Aplikací vztahu (3.12) v rovnici (3.11) a následnou úpravou je získán vztah mezi pořadovým číslem spektrální složky *k* a normovanou frekvencí k této složce f_k (3.13)

$$f = \frac{k}{N} \,. \tag{3.13}$$

Při numerickém výpočtu spektrálních složek se používá algoritmus FFT (Fast Fourier Transform – rychlá Fourierova transformace). Při použití tohoto algoritmu jsou získány další příznaky popisující řečový signál ve spektru – spektrální koeficienty.

3.2.3 Zpracování v kepstru

Kepstální analýza patří do homomorfní skupiny metod nelineárního zpracování řečových signálů, které jsou založeny na využití obecného principu superpozice. Tyto analýzy se hodí pro oddělování signálů, které vznikly konvolucí či násobením dvou nebo více složek. Parametry získané kepstrální analýzou mohou sloužit k identifikaci pauz, k rozpoznávání řeči, ale jsou zároveň cenným materiálem pro rozpoznávání mluvčích.

Slovo „kepstrum“ vzniklo přesmyčkou písmen ve slově spektrum a má se spektrem signálu i úzký vztah. Kepstrum je nelineárně upravené spektrum, kde

užívanou nelineární funkcí je přirozený logaritmus. Kepstrum je tedy definováno rovnicí (3.14)

$$\ln S(k) = \sum_{n=-\infty}^{+\infty} c(n) e^{j2\pi fn}, \tag{3.14}$$

kde hodnoty $c(n)$ jsou kepstální koeficienty a $S(k)$ spektrum diskrétního signálu. Předpokládáme-li, že funkce $S(k)$ je sudá a koeficienty $c(n)$ jsou reálné a platí pro ně (3.15)

$$c(n) = c(-n), \tag{3.15}$$

je suma v rovnici (3.14) definicí diskrétní Fourierovy transformace, a proto je možno kepstrální koeficienty vypočítat dle (3.16)

$$c(n) = F^{-1}[\ln(S(k))] = F^{-1}\{\ln|F[s(n)]|^2\}, \tag{3.16}$$

kde F je diskrétní Fourierova transformace, F^{-1} funkcí inverzní k diskrétní Fourierově transformaci a $s(n)$ navzorkovaný vstupní signál. [1]

Dalšími užívanými parametry zpracování řeči při homomorfní analýze jsou vedle výše definovaných kepstrálních koeficientů melovské kepstální koeficienty (kompenzace nelineárního vnímání frekvencí lidským uchem rozložením frekvenční oblasti do melovské frekvenční škály) a dynamické koeficienty (vyjadřují dynamiku-derivaci časové změny vektorů výše uvedených kepstálních příznaků).

3.3 Klasifikace

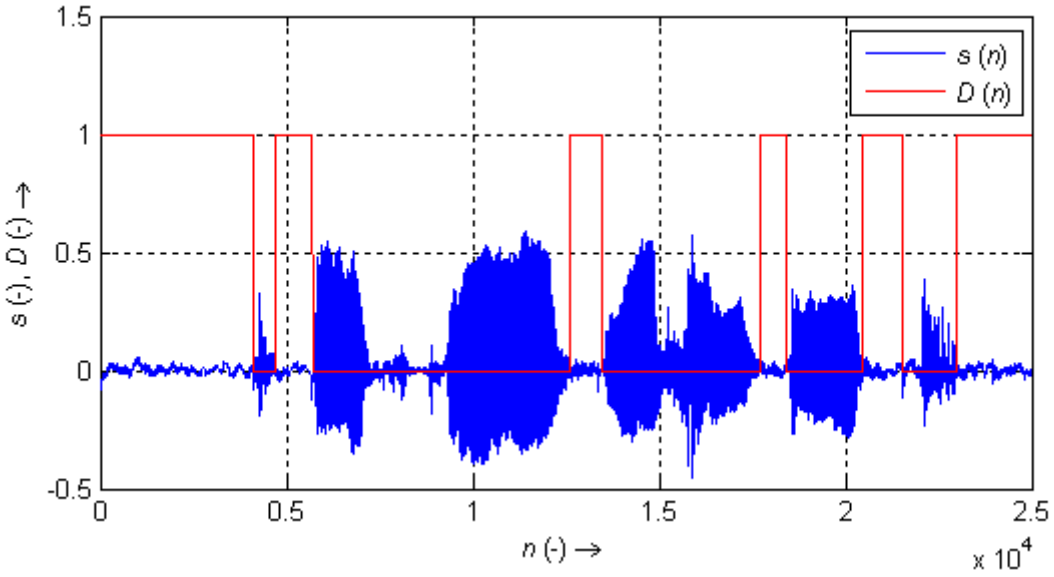
V předchozí podkapitole byly popsány jednotlivé příznaky, které byly při realizaci této diplomové práce použity pro klasifikaci - identifikaci pauz. Klasifikace příznaků je tedy synonymem detekce pauz, z tohoto důvodu čtenář v této podkapitole nalezne popisy detektorů, které byly v rámci diplomové práce realizovány.

3.3.1 Ideální detektor

Ideálním způsobem detekce pauz se rozumí označení úseků bez řečové aktivity ručně, dle následující logiky identifikační funkce $D(n)$:

$$D(n) = \begin{cases} 0 & \text{pro signál s řečovou aktivitou} \\ 1 & \text{pro signál bez řečové aktivity} \end{cases}, \tag{3.17}$$

Pro velké množství dat je tento způsob velmi náročný a pro automatické rozpoznávání řečového signálu nerealizovatelný. Na obr. 3.5 je znázorněna detekce řečového signálu ideálním detektorem.



Obr. 3.5 Ideální detekce pauz u prvních 25 000 vzorků zkoumaného řečovém signálu

Ideální detektor byl použit jako reference, ke které byly vztahovány výsledky identifikace realizovaných automatických detektorů.

3.3.2 Energetický detektor

V časové oblasti zobrazení signálu byl realizován energetický detektor, který je založen na rozdílných energetických vlastnostech segmentů řečového signálu řeč obsahujících a neobsahujících. Je tedy nejjednodušším kritériem pro rozlišení přítomnosti řeči od řečového klidu.

Blokové schéma energetického detektoru je na obr. 3.6. Detektor stanovuje práh na základě sledování krátkodobé energie signálu. Střední hodnota energie $\hat{\mu}_E(i)$ signálu je odhadována exponenciálním zapomínáním zvlášť v segmentech s řečovou aktivitou a v segmentech bez řečové aktivity. V segmentech s řečovou aktivitou se odhad střední energie *i*-tého segmentu aktualizuje s konstantou $\lambda_2 = 0{,}99$.

$$\hat{\mu}_E(i) = \lambda_2 \hat{\mu}_E(i-1) + (1-\lambda_2) E_x(i)\,, \tag{3.18}$$

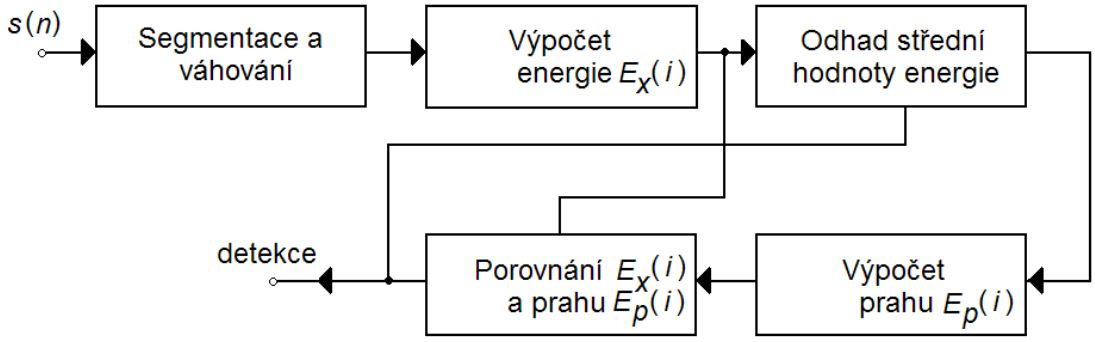
Kde $E_x(i)$ je energie akutního *i*-tého segmentu. V segmentech bez řečové aktivity, pak rychleji s konstantou zapomínání $\lambda_1 = 0{,}5$ je energie šumu *i*-tého segmentu

$$\hat{\mu}_E(i) = \lambda_1 \hat{\mu}_E(i-1) + (1-\lambda_1) E_x(i), \tag{3.19}$$

V každém segmentu se aktualizuje hodnota prahu $E_p(i)$ pro určení řečové aktivity. Je-li energie aktuálního segmentu $E_p(i)$ menší než práh, je segment považován za pauzu. Velikost *i*-tého segmentu je dána empiricky jako (3.20):

$$E_P(i) = z\hat{\mu}_E(i), \tag{3.20}$$

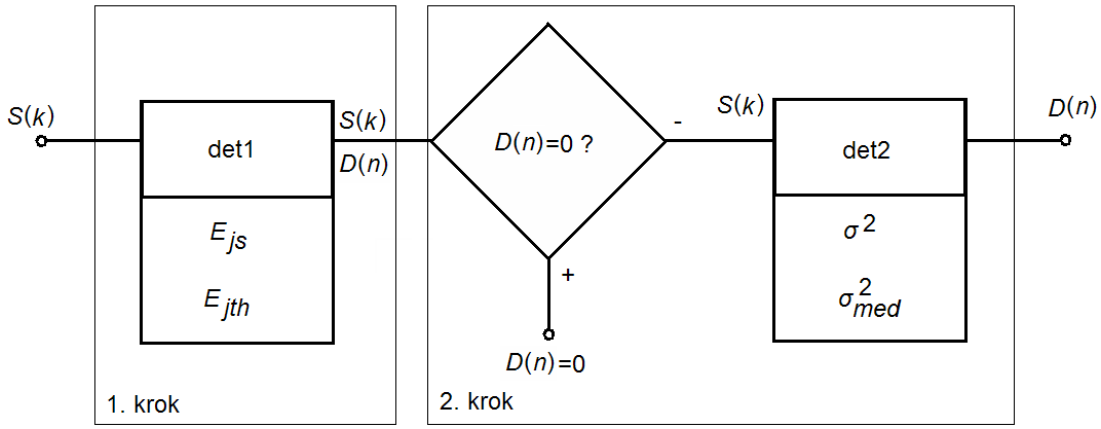
kde z = 1,3 je empiricky stanovená konstanta [6].



Obr. 3.6 Blokové schéma energetického detektoru

3.3.3 Spektrální detektor

Ve spektrální oblasti byl realizován dvojkrokový spektrální detektor, jehož blokové schéma je znázorněno na obr. 3.7. První krok vychází z úvahy rozdílných energetických vlastností v jednotlivých dílčích pásmech spekra, které lze sledovat na spektrogramu, viz. obr. 3.3. Ve druhém kroku spektrálního detektoru je aplikována jednoduchá statistika.



Obr. 3.7 Blokové schéma spektrálního detektoru

V prvním kroku se provede pásmová filtrace segmentu. Spektrum je tedy rozděleno do 4 stejně velkých sub-pásem. Při vzorkovacím kmitočtu 8 kHz je velikost jednoho pásma 1 kHz. Pro každé dílčí pásmo je vypočtena spektrální energie E_{js} (3.21)

$$E_{js} = \sum_{k=j}^{j\frac{N}{8}} S(k)^2 , \tag{3.21}$$

kde $S(k)$ je spektrum segmentu řečového signálu, j pořadové číslo dílčího pásma a N délka segmentu.

Pro každé sub-pásmo je vypočtena prahová hodnota energie E_{jth} jako aritmetický průměr prvních 30 segmentů, o kterých se předpokládá, že řeč neobsahují (3.22)

$$E_{jth} = z \frac{\sum_{i=1}^{30} E_{js}(i)}{30} , \tag{3.22}$$

kde i představuje pořadové číslo segmentu a empirická konstanta $z = 1$ pro $j = 1,2,3$ a $z = 0.6$ pro $j = 4$.

V prvním kroku je pauza detekována, je-li prahová hodnota energie překročena v prvním sub-pásmu nebo alespoň ve dvou dalších ze třech ostatních sub-pásmem. Detekce pauzy $D(n)$ je opět definována dle rovnice (3.17).

Výpočet druhého kroku je realizován pouze tehdy, je-li v prvním kroku pro daný segment detekována pauza. Ve druhém kroku je vypočten rozptyl σ^2 spektrálního vektoru pro daný segment (3.23)

$$\sigma^2 = \frac{\sum_{k=1}^{N} |S(k)|^2 - \frac{\left|\sum_{k=1}^{N} S(k)\right|^2}{N}}{N-1} , \tag{3.23}$$

kde N je délka spektrálního vektoru pro daný segment. Z důvodu kolísání posloupnosti σ^2 a tím způsobených špatných rozhodnutí, je posloupnost σ^2 dále ještě vyhlazena mediánovým filtrem řádu m (3.24)

$$\sigma^2_{med} = \mathrm{med}(\Delta c_N[n], m) , \tag{3.24}$$

Prahová hodnota pro druhý krok spektrálního detektoru byla nastavena jako aritmetický průměr prvních třiceti prvků posloupnosti σ^2_{med}.

3.3.4 Jednokrokový integrální kepstrální detektor

Jednoduchým algoritmem pro identifikaci pauz v kepstální oblasti je jednokrokový kepstrální detektor [7]. Blokové schéma je znázorněno na obr. 3.8 – blok 1. krok. Jednokrokový kepstální detektor je založen na výpočtu rozdílu (vzdálenosti) aktuálního kepstrálního $c(n)$ vektoru a krátkodobého průměru kepstra pozadí $\bar{c}(n)$. Integrální kepstrální vzdálenost je definována vztahem (3.25), pro jednoduchost byly indexy (n) vynechány.

$$\Delta c = 4{,}3429\sqrt{(c_0 - \bar{c}_0)^2 + 2\sum_{k=1}^{p}(c_k - \bar{c}_k)^2}\ , \tag{3.25}$$

Kde c_0, c_0, c_k a c_k jsou kepstrální koeficienty a p délka kepstálního vektoru.

Kepstrální vektor pozadí $\bar{c}(n)$ je aktualizován pouze v pauzách (3.26), kde parametr λ specifikuje časovou konstantu krátkodobého zapomínání.

$$\bar{\mathbf{c}}(n+1) = (1-\lambda)\bar{\mathbf{c}}(n) + \lambda\mathbf{c}(n)\ . \tag{3.26}$$

Rozhodnutí, zda aktuální segment řečového signálu obsahuje řeč, nebo jedná-li se o pauzu, je založeno na jednoduché statistice. V několika prvních segmentech řečového signálu, o kterých se předpokládá, že řeč neobsahují, a jejich celkový počet je větší než 30, je vypočtena dlouhodobá střední hodnota $\overline{\Delta c_N}$ a směrodatná odchylka vektoru Δc dv(Δc_N). Z těchto dvou hodnot je určena rozhodovací úroveň (3.27)

$$\Delta c_{th} = \overline{\Delta c_N}(n) + z \cdot \mathrm{dv}(\Delta c_N(n))\ , \tag{3.27}$$

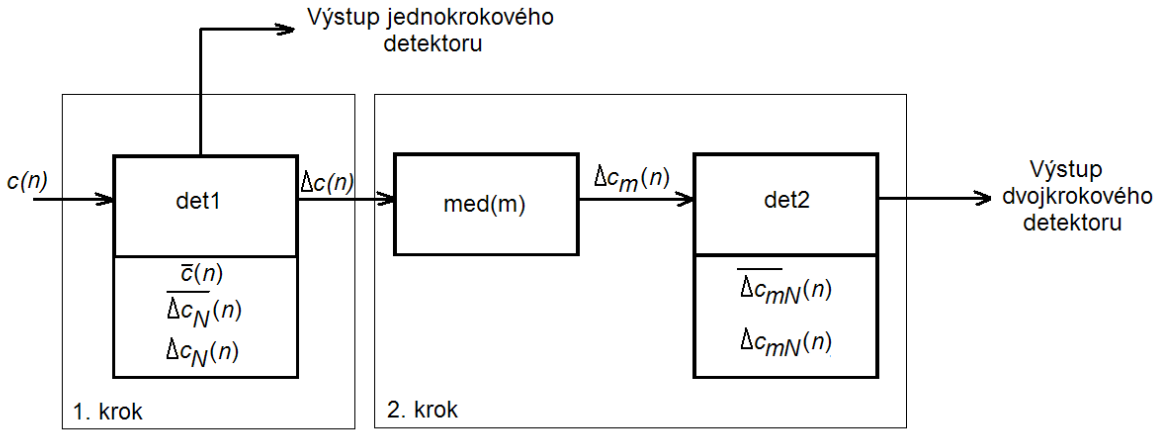
kde s pomocí konstanty z se řídí počet chybných detekcí. Pauza je poté detekována, je-li aktuální kepstální vzdálenost menší než prahová hodnota. Detekce pauzy $D(n)$ je definována dle rovnice (3.17)

3.3.5 Dvojkrokový integální kepstrální detektor

Kvůli častým špatným rozhodnutím jednokrokového detektoru, která jsou způsobena kolísáním charakteristik šumu v pozadí, byl realizován dvojkrokový detektor. Posloupnost Δc_N je vyhlazena mediánovým filtrem řádu m (3.28)

$$\Delta c_m(n) = \mathrm{med}(\Delta c_N(n), m)\ , \tag{3.28}$$

kde Δc_m je vyfiltrovaná posloupnost. Rozhodovací úrovně pro detekci jsou opět stanoveny dle rovnice (3.27). Blokové schéma dvojkrokového detektoru je znázorněno na obrázku 3.8.



Obr. 3.8 Integrální kepstální detektor

Velký vliv na správnou detekci dvojkrokového kepstrálního detektoru má v řád *m* mediánového filtru. Vysokou hodnotou parametru se eliminují špatné detekce, avšak příliš vysoká hodnota způsobí špatné určení počátku a konce detekovaných pauz. Volí se tedy kompromis, při řešení této diplomové práce byl použit mediánový filtr 6. řádu.

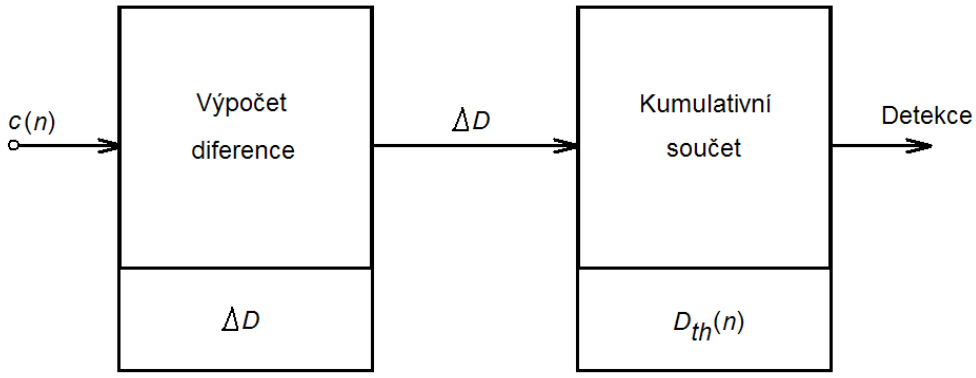
3.3.6 Diferenciální kepstální detektor

Diferenciální kepstální detektor je dalším detektorem, který byl v rámci diplomové práce realizován. Rovnice definující kepstální vzdálenost Δc (3.25) může být přepsána do tvaru (3.29)

$$\Delta D(n) = \sqrt{\delta_0^2(n) + 2\sum_{k=1}^{p} \delta_k^2(n)}\,, \tag{3.29}$$

kde ΔD je diferenciální kepstrální vzdálenost, δ_k je časovou derivací kepsrálních koeficientů c_k. Časová derivace při numerických výpočtech je nahrazena diferencí. Diference je rozdíl mezi dvěmi sousedními hodnotami vzorků (zde mezi dvěma sousedními kepstálními koeficienty). Diference je dále vyhlazena kumulativním součtem. Rozhodovací úrovně jsou opět definovány podle rovnice (3.27).

Blokové schéma diferenciálního detektoru čtenář nalezne na obr. 3.9.



Obr. 3.9 Diferenciální kepstální detektor

4. EXPERIMENTY

V předchozí kapitole byla popsána koncepce zpracování řečového signálu, jak byla aplikována na konkrétní řešení zadání. V rámci diplomové práce byly výše popsané detektory realizovány v programovém prostředí MATLAB. V této kapitole čtenář nalezne rozbor realizovaných experimentů, jejich vyhodnocení a popis jednotlivých skriptovacích souborů a funkcí `*.m` v MATLABu.

4.1 Vstupní data pro realizované detektory

Robustnost realizovaných detektorů byla testována především vůči rušení. Zároveň byl testován vliv preemfáze na výslednou detekci. Pro nezkreslení výsledných vyhodnocení vlivem robustnosti detektorů na různé druhy promluv, byla vstupními daty právě jedna promluva (řečový signál), která byla rušena různými úrovněmi bílého šumu a několika reálných hluků. Standardním kritériem pro měření šumu (hluků) v signálu je odstup signálu od šumu SNR (Signal to Noise Ratio), proto je vhodné tuto veličinu v následující podkapitole definovat.

4.1.1 Odstup signálu od šumu

Základní definice odstupu signálu od šumu je dána vztahem (4.1)

$$SNR = 10\log\frac{p_s}{p_n}, \tag{4.1}$$

kde p_S je výkon užitečného signálu a p_n výkon šumu (hluku). Aplikací rovnice (4.1) na navzorkovaný řečový signál, počítá-li se výkon řeči a šumu přes celý signál, se získá globální SNR *GSNR*, tj.

$$GSNR = 10\log\frac{p_s}{p_n} = 10\log\frac{\sum_{n=0}^{l-1} s'^2(n)}{\sum_{n-0}^{l-1} n'^2(n)} = 10\log\frac{E_r}{E_n}, \tag{4.2}$$

kde *s'(n)* je řečový signál a *n'(n)* šum, *l* délka řečového signálu, E_r energie řečového signálu a E_n energie šumu (rozdíl mezi energií a výkonem je záležitostí pouze změny měřítka a jelikož jsou ve vztahu (4.2) energie v poměru, nemá rozdíl na výsledek

GSNR vliv). Globální SNR je však zatíženo chybou, neboť při výpočtu výkonu (energie) řeči jsou zahrnuty také části signálu bez řečové aktivity, které snižují celkový výkon řečového signálu. Korektní výpočet odstupu signálu od šumu, který byl při testování robustnosti detektorů aplikován, je dán (4.3)

$$SNR = 10\log\frac{\sum_{n=0}^{l-1} s'^{2}(n)D_{neg}(n)}{\sum_{n=0}^{l-1} n'^{2}(n)D_{neg}(n)}, \tag{4.3}$$

kde $D_{neg}(n)$ je negovaná identifikační funkce $D(n)$ (3.17) ideálního detektoru.[8]

4.1.2 Zdroje vstupních dat

Vstupní promluva (`OSR.wav`) byla získána z otevřené databáze řečových signálů [9]. Jedná se o půlminutovou promluvu ženského hlasu v anglickém jazyce, která je kódovaná PCM s 16 bitovou délkou slova a vzorkovací frekvencí 8 kHz. K jednotlivým vzorkům promluvy byl přičítán bílý šum a několik hluků. Bílý šum byl generován v MATLABu jako posloupnost náhodných čísel (funkce `randn`) a reálné rušení bylo získáno vlastním nahráním zvuku tekoucí vody (`voda.wav`), hučení klimatizace v servrovne (`servrovna.wav`), zvuku motoru (`motor.wav`), lidského kašle (`kasel.wav`) a směsi zvuků na Burianově náměstí v Brně v rušný den (`burianovo_namesti.wav`). Jednotlivé hluky byly rovněž kódovány PCM s 16 bitovou délkou slova a vzorkovací frekvencí 8 kHz. Délka všech signálů rušení byla upravená na délku shodnou se vstupní promluvou a velikost jednotlivých vzorků pro dílčí experimenty upravená tak, aby splňovala úrovně odstupu signálu od šumu -5, 0, 5, 10, 20 dB.

4.2 Vyhodnocení úspěšnosti detekce

Pro vyhodnocování úspěšnosti detekce byly sestaveny charakteristiky ROC (Receiver Operating Characteristic), které zobrazují závislost pravděpodobnosti planého poplachu na pravděpodobnosti správné detekce. Parametrem pro vykresleni ROC křivek bylo měnící se rušení. Zároveň byly sestaveny křivky závislosti pravděpodobnosti správné detekce pauzy a pravděpodobnosti správné detekce řečového rámce na měnícím se SNR a pro jednotlivé detektory.

4.2.1 ROC charakteristiky

ROC charakteristiky jsou standardním způsobem vyhodnocování úspěšnosti detekce. Jak již bylo řečeno, zobrazují závislost pravděpodobnosti správné detekce na pravděpodobnosti planého poplachu.

Pravděpodobnost správné detekce segmentu signálu řeč neobsahující *HR0* je definována vztahem (4.3)

$$HR0 = \frac{N_{0,0}}{N_0^{ref}}, \tag{4.3}$$

kde $N_{0,0}$ je počet správně detekovaných segmentů a N_0^{ref} je skutečný celkový počet segmentů řeč neobsahujících.

Analogicky je definována pravděpodobnost správné detekce řečového rámce *HR1* (4.4)

$$HR1 = \frac{N_{1,1}}{N_1^{ref}}, \tag{4.4}$$

kde $N_{1,1}$ je počet správně detekovaných segmentů a N_1^{ref} je skutečný celkový počet řečových rámců.

Pravděpodobnost planého poplachu *FAR0* je poté dána vztahem (4.5) []

$$FAR0 = 1 - HR1. \tag{4.5}$$

4.2.2 Výsledky detekcí

Výsledky detekce pro různé typy hluků, SNR, s filtrací preemfázovým filtrem a bez užití preemfázového filtru byly pro přehlednost uspořádány do tabulek 4.1 respektive 4.2, kde současně byly zavedeny zkratky testovaných detektorů:

ED Energetický detektor v časové oblasti

SD Spektrální detektor

1KD Jednokrokový integrální kepstrální detektor

2KD Dvojkrokový integrální kepstrální detektor

DD Diferenciální kepstrální detektor

Tab. 4.1 Výsledky experimentů při použití preemfáze

		ED		SD		1KD		2KD		DD	
hluk	*SNR*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*
bez	-	0,245	0,759	0,854	0,102	0,867	0,059	0,862	0,042	0,850	0,064
bílý šum	-5	0,963	0,724	1,000	0,711	0,980	0,971	1,000	1,000	0,396	0,219
	0	0,968	0,521	0,999	0,451	0,983	0,967	1,000	0,991	0,444	0,150
	5	0,957	0,393	0,995	0,330	0,967	0,895	1,000	0,982	0,512	0,090
	10	0,915	0,340	0,969	0,198	0,980	0,693	1,000	0,793	0,550	0,055
	15	0,753	0,279	0,912	0,128	0,988	0,654	0,998	0,555	0,449	0,045
	20	0,527	0,527	0,873	0,110	0,982	0,414	0,991	0,384	0,618	0,040
kašel	-5	0,415	0,295	0,597	0,216	0,265	0,023	0,223	0,007	0,318	0,054
	0	0,399	0,254	0,462	0,085	0,286	0,017	0,268	0,007	0,323	0,041
	5	0,417	0,246	0,465	0,060	0,311	0,024	0,272	0,005	0,335	0,030
	10	0,417	0,236	0,538	0,070	0,301	0,012	0,294	0,001	0,348	0,027
	15	0,395	0,247	0,579	0,068	0,346	0,021	0,336	0,013	0,370	0,026
	20	0,366	0,254	0,667	0,076	0,474	0,031	0,478	0,024	0,442	0,031
voda	-5	0,538	0,420	0,992	0,928	0,951	0,845	0,998	0,904	0,532	0,321
	0	0,455	0,288	0,991	0,895	0,941	0,724	0,997	0,738	0,532	0,250
	5	0,351	0,296	0,991	0,585	0,944	0,645	0,993	0,667	0,532	0,186
	10	0,323	0,251	0,985	0,364	0,941	0,492	0,984	0,493	0,539	0,103
	15	0,347	0,258	0,964	0,262	0,928	0,327	0,982	0,315	0,564	0,060
	20	0,412	0,261	0,926	0,168	0,911	0,200	0,924	0,124	0,599	0,048
servrovna	-5	0,929	0,678	1,000	0,960	0,986	0,979	1,000	1,000	0,511	0,275
	0	0,839	0,467	1,000	0,654	0,977	0,929	1,000	0,974	0,511	0,194
	5	0,652	0,323	0,997	0,386	0,971	0,787	0,999	0,838	0,520	0,114
	10	0,483	0,278	0,945	0,223	0,983	0,636	1,000	0,622	0,534	0,072
	15	0,448	0,265	0,859	0,137	0,961	0,368	0,994	0,358	0,588	0,550
	20	0,458	0,256	0,822	0,108	0,950	0,205	0,991	0,232	0,706	0,045
motor	-5	0,660	0,590	0,989	0,684	0,994	0,985	1,000	1,000	0,059	0,011
	0	0,581	0,381	0,971	0,391	0,994	0,952	1,000	0,981	0,057	0,010
	5	0,521	0,311	0,906	0,200	0,991	0,887	1,000	0,916	0,056	0,074
	10	0,433	0,292	0,654	0,076	0,988	0,727	0,999	0,777	0,064	0,005
	15	0,432	0,267	0,672	0,057	0,976	0,447	0,996	0,494	0,070	0,005
	20	0,408	0,251	0,745	0,060	0,971	0,294	0,980	0,269	0,092	0,005
náměstí	-5	0,441	0,346	1,000	0,996	0,992	0,814	1,000	0,913	0,988	0,935
	0	0,404	0,273	1,000	0,864	0,991	0,650	1,000	0,722	0,987	0,878
	5	0,360	0,253	0,999	0,561	0,983	0,407	0,999	0,451	0,987	0,739
	10	0,329	0,252	0,993	0,347	0,930	0,225	0,953	0,205	0,976	0,530
	15	0,320	0,255	0,967	0,208	0,949	0,160	0,968	0,148	0,963	0,335
	20	0,347	0,256	0,900	0,146	0,878	0,073	0,916	0,070	0,943	0,212

Tab. 4.2 Výsledky experimentů bez použití preemfáze

		ED		SD		1KD		2KD		DD	
hluk	*SNR*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*
bez	-	0,385	0,223	0,852	0,050	0,833	0,049	0,857	0,036	0,849	0,065
bílý šum	-5	0,305	0,348	0,994	0,993	0,990	0,990	1,000	1,000	0,453	0,210
	0	0,367	0,345	0,896	0,858	0,972	0,936	1,000	1,000	0,501	0,222
	5	0,312	0,318	0,995	0,967	0,979	0,879	0,999	0,973	0,309	0,071
	10	0,406	0,309	0,942	0,700	0,987	0,735	1,000	0,823	0,498	0,063
	15	0,404	0,232	0,998	0,484	0,975	0,568	0,999	0,608	0,437	0,047
	20	0,440	0,253	0,989	0,296	0,990	0,552	0,994	0,404	0,490	0,031
kašel	-5	0,476	0,369	0,390	0,080	0,270	0,026	0,234	0,011	0,317	0,054
	0	0,479	0,345	0,336	0,034	0,277	0,014	0,269	0,009	0,323	0,042
	5	0,492	0,316	0,369	0,024	0,314	0,022	0,291	0,013	0,332	0,031
	10	0,480	0,248	0,492	0,037	0,296	0,015	0,304	0,005	0,347	0,027
	15	0,481	0,230	0,550	0,038	0,349	0,022	0,340	0,012	0,369	0,027
	20	0,461	0,228	0,618	0,049	0,493	0,031	0,496	0,024	0,439	0,031
voda	-5	0,406	0,356	0,902	0,856	0,958	0,877	0,982	0,897	0,533	0,321
	0	0,414	0,286	0,902	0,759	0,946	0,733	0,997	0,746	0,531	0,249
	5	0,394	0,240	0,900	0,448	0,949	0,662	0,994	0,665	0,533	0,184
	10	0,357	0,244	0,895	0,242	0,927	0,519	0,968	0,555	0,539	0,100
	15	0,403	0,237	0,861	0,144	0,889	0,329	0,935	0,311	0,564	0,061
	20	0,420	0,229	0,842	0,071	0,904	0,200	0,912	0,118	0,600	0,047
servrovna	-5	0,763	0,628	1,000	0,944	0,979	0,974	1,000	1,000	0,515	0,277
	0	0,762	0,467	1,000	0,654	0,978	0,922	1,000	0,981	0,517	0,197
	5	0,701	0,337	0,997	0,335	0,974	0,755	1,000	0,943	0,526	0,117
	10	0,597	0,284	0,964	0,176	0,985	0,661	0,999	0,754	0,545	0,073
	15	0,552	0,250	0,888	0,084	0,963	0,378	0,996	0,475	0,594	0,056
	20	0,520	0,232	0,854	0,052	0,957	0,206	0,949	0,292	0,703	0,045
motor	-5	0,326	0,331	0,329	0,229	0,995	0,985	1,000	1,000	0,059	0,015
	0	0,331	0,310	0,324	0,136	0,994	0,957	1,000	0,981	0,056	0,009
	5	0,357	0,282	0,306	0,069	0,994	0,921	1,000	0,943	0,056	0,007
	10	0,370	0,246	0,278	0,044	0,989	0,752	0,999	0,754	0,063	0,005
	15	0,416	0,244	0,293	0,026	0,978	0,454	0,996	0,475	0,071	0,004
	20	0,436	0,229	0,464	0,023	0,953	0,356	0,949	0,292	0,092	0,005
náměstí	-5	0,516	0,383	1,000	0,985	0,990	0,808	1,000	0,893	0,988	0,934
	0	0,488	0,303	1,000	0,909	0,986	0,625	1,000	0,689	0,987	0,878
	5	0,415	0,265	0,999	0,581	0,978	0,377	0,999	0,431	0,986	0,740
	10	0,358	0,247	0,996	0,313	0,929	0,215	0,943	0,202	0,975	0,531
	15	0,360	0,232	0,977	0,173	0,946	0,164	0,967	0,158	0,962	0,339
	20	0,403	0,231	0,917	0,084	0,862	0,065	0,912	0,066	0,943	0,218

Pro sestavení ROC křivek, ve kterých byl měnícím se parametrem rušivý signál, byly pravděpodobnosti jednotlivých detektorů pro všechny úrovně vstupního signálu zprůměrňovány a uvedeny do tabulky 4.3 respektive 4.4 v závislosti na použití preemfázového filtru. ROC křivky pro jednotlivé detektory nalezne čtenář v přílohách, které jsou zařazeny na konec diplomové práce.

Tab. 4.3 Pravděpodobnosti detektorů pro různá rušení bez použití preemfázového filtru

	ED		SD		1KD		2KD		DD	
hluk	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*
bílý šum	0,847	0,464	0,958	0,321	0,980	0,766	0,998	0,784	0,495	0,100
kašel	0,402	0,255	0,551	0,096	0,330	0,021	0,312	0,009	0,356	0,035
voda	0,404	0,296	0,975	0,534	0,936	0,539	0,980	0,540	0,550	0,161
serverovna	0,635	0,378	0,937	0,412	0,971	0,650	0,997	0,671	0,562	0,208
motor	0,506	0,349	0,823	0,245	0,986	0,715	0,996	0,739	0,066	0,018
náměstí	0,367	0,273	0,977	0,520	0,954	0,388	0,973	0,418	0,974	0,605
bez	0,245	0,759	0,854	0,102	0,867	0,059	0,862	0,042	0,850	0,064

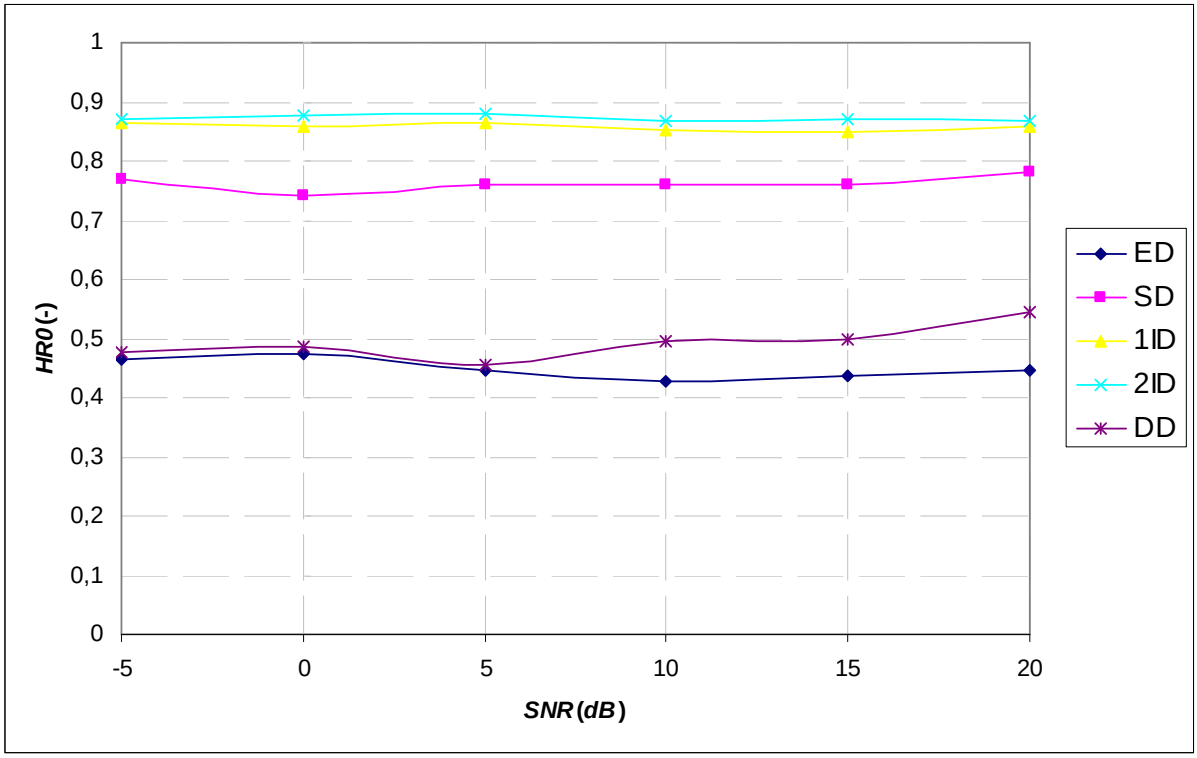
Tab. 4.4 Pravděpodobnosti detektorů pro různá rušení s použitím preemfázového filtru

	ED		SD		1KD		2KD		DD	
hluk	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*	*HR0*	*FAR0*
bílý šum	0,372	0,301	0,969	0,716	0,982	0,777	0,999	0,801	0,448	0,107
kašel	0,478	0,289	0,459	0,044	0,333	0,022	0,322	0,012	0,355	0,036
voda	0,399	0,265	0,884	0,420	0,929	0,554	0,965	0,549	0,550	0,160
serverovna	0,649	0,366	0,951	0,374	0,973	0,649	0,991	0,741	0,567	0,127
motor	0,373	0,274	0,332	0,088	0,984	0,737	0,991	0,741	0,066	0,008
náměstí	0,423	0,277	0,982	0,508	0,949	0,376	0,970	0,407	0,973	0,607
bez	0,385	0,223	0,852	0,050	0,833	0,049	0,857	0,036	0,849	0,065

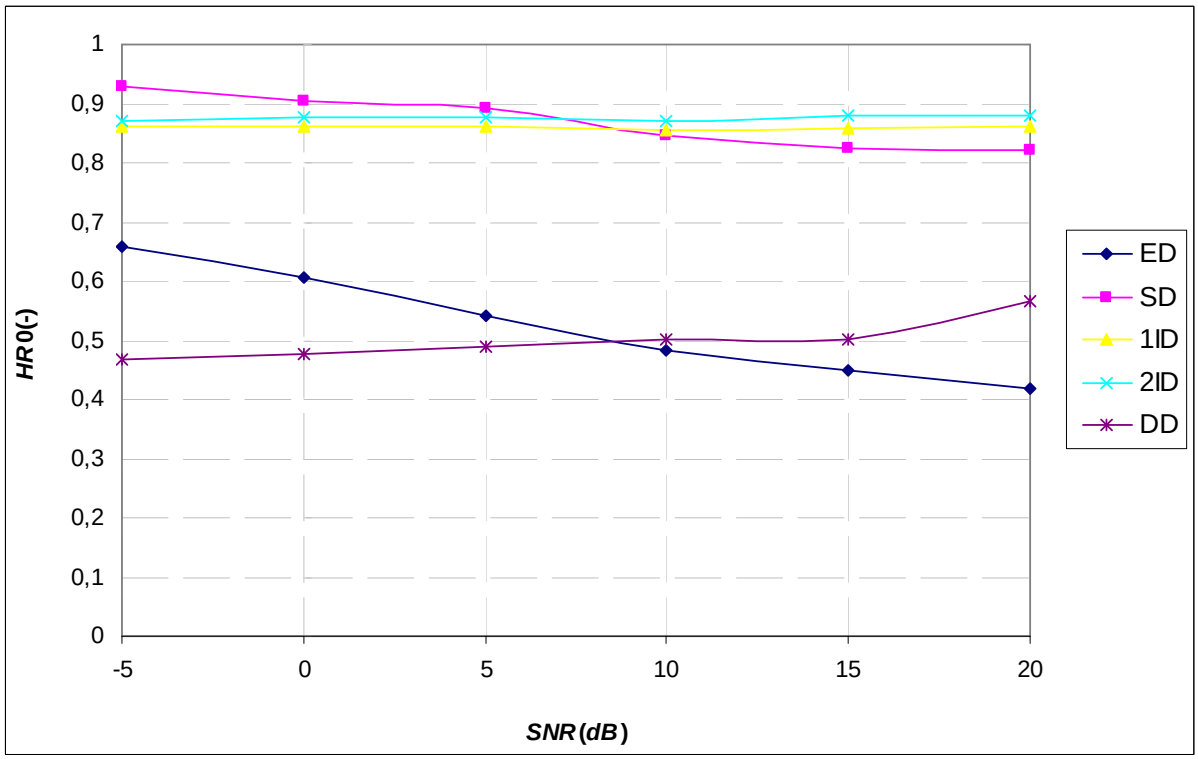
Současně s ROC křivkami byly pro jednotlivé detektory sestaveny křivky závislosti správné detekce segmentu signálu řeč neobsahující *HR0* a pravděpodobnosti správné detekce řečového rámce *HR1*. Obě pravděpodobnosti byly sestaveny jako průměr hodnot *HR0* a *HR1* pro různá rušení při zachování konstantní hodnoty odstupu signálu od šumu. Tyto závislosti při použití preemfáze a bez preemfáze jsou uvedeny na obr. 4.1, 4.2, 4.3, 4.4.

Z obr. 4.1 a 4.2 je patrné, že vliv preemfáze na pravděpodobnost správné detekce segmentu signálu řeč neobsahující *HR0* (která byla pro různá rušení průměrkována) při měnícím se odstupu signálu od šumu se výrazně projevil u energetického detektoru a spektrálního detektoru. Toto zjištění bylo předpokládáno, poněvadž preemfáze mění energetické vlastnosti signálu a oba zmíněné detektory identifikují části promluvy bez řečového signálu na základě výpočtu energie. Vliv

preemfáze se již neprojevil u kepstrálních detektorů. Čtenář by mohl namítnout, že první kepstrální koeficient přece nese informaci o energii signálu, parametrem pro vyhodnocení kepstrálních detektorů je však kepstrální vzdálenost vypočtená jako

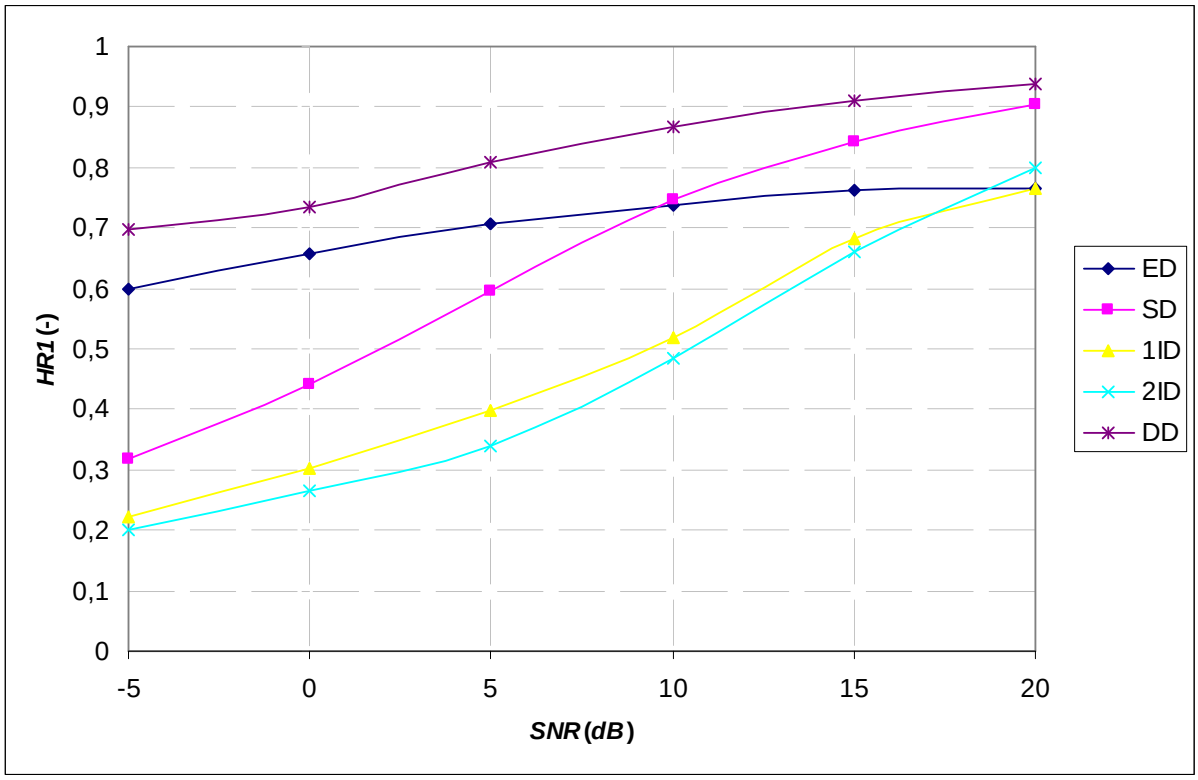


Obr. 4.1 Závislost *HR0* na *SNR* pro jednotlivé detektory při použití preemfáze

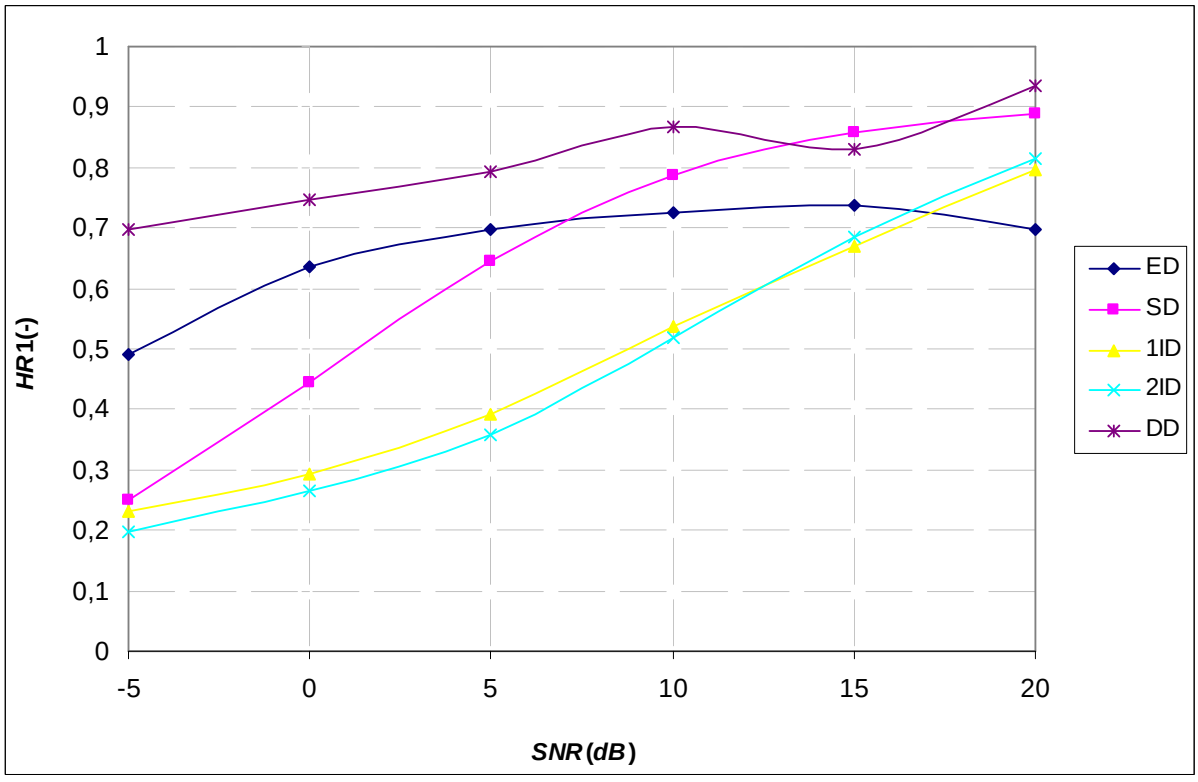


Obr. 4.2 Závislost *HR0* na *SNR* pro jednotlivé detektory bez použití preemfáze

rozdíl mezi aktuálního kepstra a kepstra pozadí. Tímto je tedy vliv energie na pravděpodobnost správné detekce eliminován.



Obr. 4.3 Závislost *HR1* na *SNR* jednotlivé detektory při použití preemfáze



Obr. 4.4 Závislost *HR1* na *SNR* pro jednotlivé detektory bez použití preemfáze

Ze závislostí obr. 4.3 a 4.4 je vidět, že vliv preemfáze na pravděpodobnost správné detekce řečového segmentu *HR0* je již zanedbatelný. Z výše popsaných zjištění vyplývá, že použití preemfáze má spíše smysl pro rozpoznávání řeči, než pro identifikaci řečového signálu či segmentů řeč neobsahujících.

4.3 Popis skriptovacích souborů a funkcí realizovaných v MATLABu

Jak již bylo několikrát řečeno, realizované detektory byly odsimulovány v programovém prostředí MATLAB. Veškeré skriptovací soubory a funkce, které byly pro testování v tomto prostředí napsány, jsou přiloženy v elektronické verzi na CD nosiči. V této podkapitole bude uveden pouze jejich stručný popis, který byl pro přehlednost seřazen do tabulky 4.5.

Tab. 4.5 Popis skriptovacích souborů a funkcí realizovaných v MATLABU

Název souboru/funkce	**Popis**
Mainfinal.m	Hlavní skriptovací soubor, z něhož se volají jednotlivé funkce a definují vstupní parametry – velikost úrovně šumu, vstupní signál, délka segmentu, jeho překrytí. Zároveň je možnost provedení preemfáze V tomto souboru se také realizuje výpočet HR0, HR1 a FAR0.
Segmentace.m	Funkce, která pro provede nesegmentování vstupního signálu.
Energie.m	Realizace energetického detektoru popsaného v podkapitole 3.3.2
Spectral_energy.m	Realizace spektrálního detektoru popsaného v podkapitole 3.3.3
One_step.m	Realizace jednokrokového integrálního kepstrálního detektoru popsaného v podkapitole 3.3.4
Two_step.m	Realizace dvojkrokového integrálního kepstrálního detektoru popsaného v podkapitole 3.3.5
Diferential_cepstral.m	Realizace diferenciálního kepstrálního detektoru popsaného v podkapitole 3.3.6
Spektrogram.m	Skriptovací soubor pro vykreslení spektrogramu (obr. 3.4)
Idealni_detektor	Skriptovací soubor pro vykreslení ideální detekce pauz (obr. 3.4)
Harmonicky.m	Skriptovací soubor pro vykreslení vzorkování (obr. 3.3)

Současně jsou k těmto souborům v elektronické verzi přiloženy také *.wav soubory testovaných záznamů zvuku a hluků.

5. ZÁVĚR

Cílem této diplomové práce bylo navrhnout účinnou metodu pro označení pauz v řeči z jednokanálového záznamu. V prostředí MATLAB bylo navrženo pět detektorů segmentů bez řečové aktivity v různých oblastech zobrazení řečového signálu. V časové oblasti to byl energetický detektor, ve skektrální oblasti dvojkrokový detektor, který v prvním kroku pro detekci využíval energetických vlastností signálu v různých dílčích pásmech spektra, ve druhém kroku statistických výpočtů. V kepstru byly realizovány dva detektory založené na integrálním algoritmu a třetí pro identifikaci pauz využíval diferenciální algoritmus.

Všechny detektory byly testovány pro právě jednu promluvu ženského hlasu získanou z otevřené databáze řečových signálů, a to bez i s přítomností nežádoucího rušení. Jako nežádoucí rušení byl použit bílý šum i vlastní nahrané reálné zvuky okolí (tekoucí voda, zvuk motoru, odpolední ruch na náměstí, kašel a hučení klimatizace v serverovně).

Pro vyhodnocení robustnosti jednotlivých detektorů byly sestaveny ROC křivky – závislosti pravděpodobnosti správné detekce na pravděpodobnosti planého poplachu. Měnícím se parametrem bylo právě nežádoucí rušení.

Byl zároveň zkoumán vliv preemfáze na změnu závislosti správné detekce segmentu signálu řeč neobsahující a pravděpodobnosti správné detekce řečového rámce při měnícím se odstupu signálu od šumu. Z výše popsaných zjištění vyplynulo, že vliv je zanedbatelný a použití preemfáze má spíše smysl pro rozpoznávání řeči, než pro identifikaci řečového signálu či segmentů řeč neobsahujících.

POUŽITÁ LITERATURA

[1] PSUTKA, J., MÜLLER, L., MATOUŠEK, J., RADOVÁ, V. *Mluvíme s počítačem česky.* Academia, Praha 2006. ISBN 80 – 200 – 1309 – 1.

[2] SIGMUND, M. *Analýza řečových signálů*. FEKT VUT v Brně, Brno 2000. ISBN 80 – 214 – 1783 – 8.

[3] ARIM, E., COSTA, F., FREITAS, T. *An empirical account of the relation between discourse structure and pauses in Portuguese.* ILTEC, Lisabon

[4] SMÉKAL, Z., *Číslicové zpracování řeči*. Skriptum FEKT VUT, Brno 2008.

[5] SMÉKAL, Z., Šebesta, V. *Signály a soustavy*. Skriptum FEKT VUT, Brno 2003.

[6] VONDRÁŠEK, M. *Odhad SNR řečového signálu snímaného v hlučném prostředí*. Diplomová práce ČVUT FEL, Praha 2004.

[7] POLLÁK P., SOVKA P., UHLÍŘ J. *Cepstral Speech/Pause Detectors*. Praha 1995. http://noel.feld.cvut.cz/speechlab/

[8] POLLÁK P. *Metody odhadu odstupu signálu od šumu v řečovém signálu.* Praha 2001. http://noel.feld.cvut.cz/speechlab/

[9] *Open speech repository*. http://www.voiptroubleshooter.com/open_speech/index.html

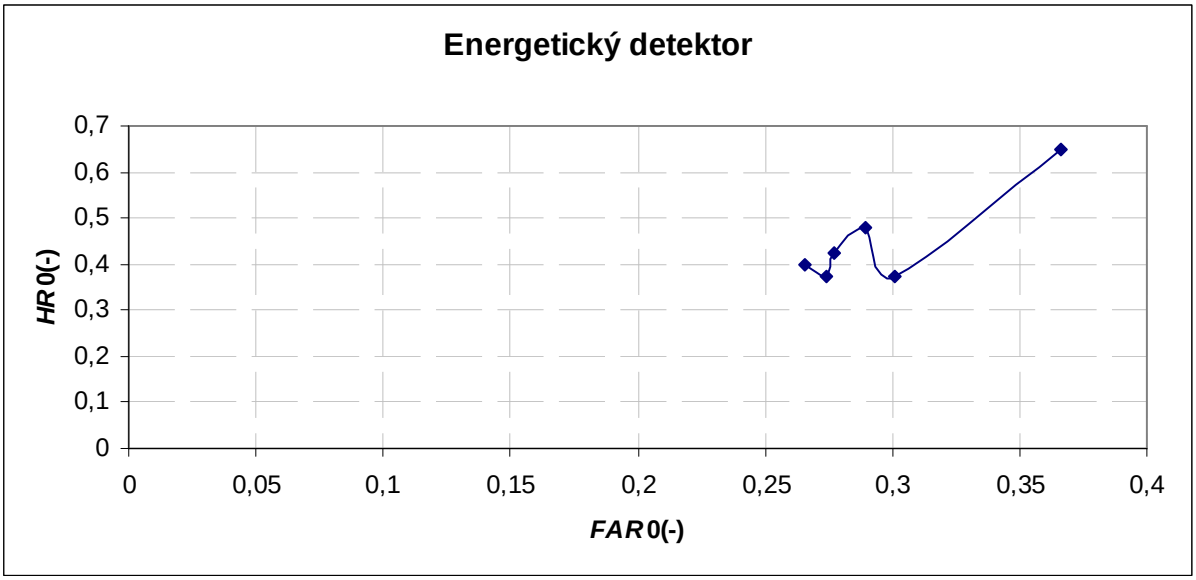
[10] RAMINEZ, J., SEGURA, J. C., BENITEZ, C., TORRE, A., RUBIO, A. *Efficient voice activity detection algoritms using long-term speech information*. Granada 2003. http://www.sciencedirect.com/

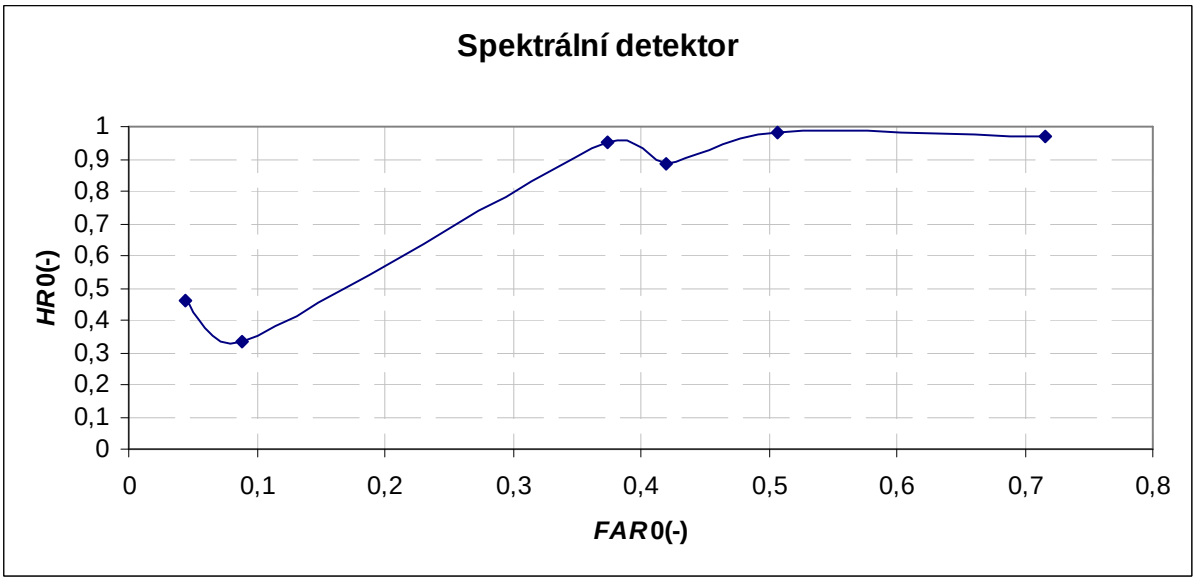
[11] ZAPLATÍLEK, K., DOŇAR B., Matlab začínáme se signály. BEN – technická literatura, Praha 2006. ISBN 80 – 7300 – 200 – 0.

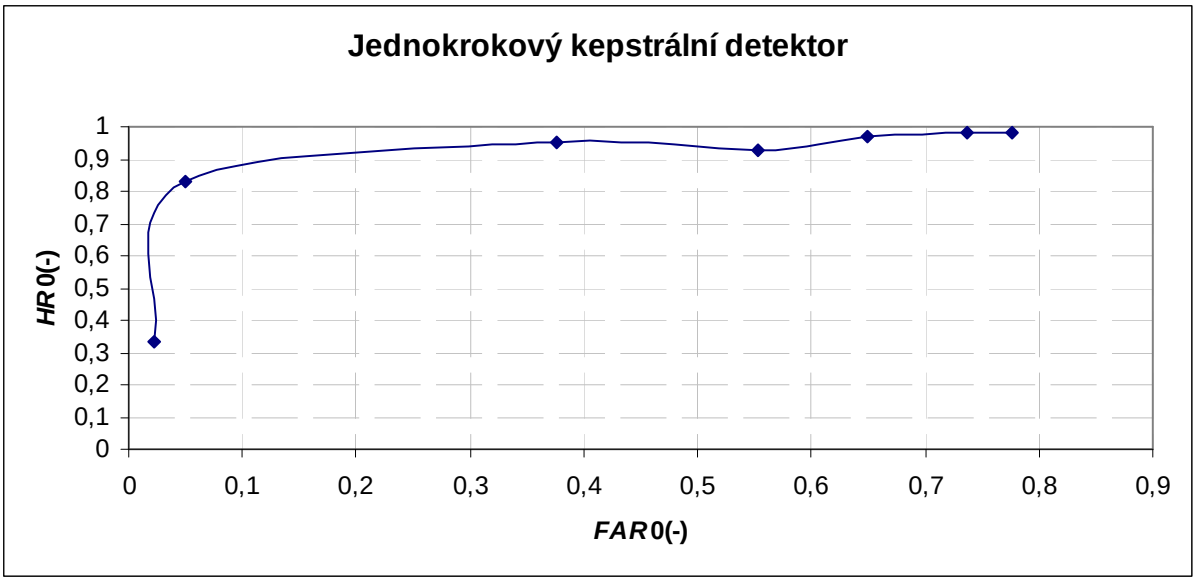
[12] Prasad, R. V., Sawan, A., Jamadagni, H., Chiranth, M. C., Sah, R., Gaurav, V. *Comparasion of Voice Activity Detection for VoIP.* IEEE 2002.

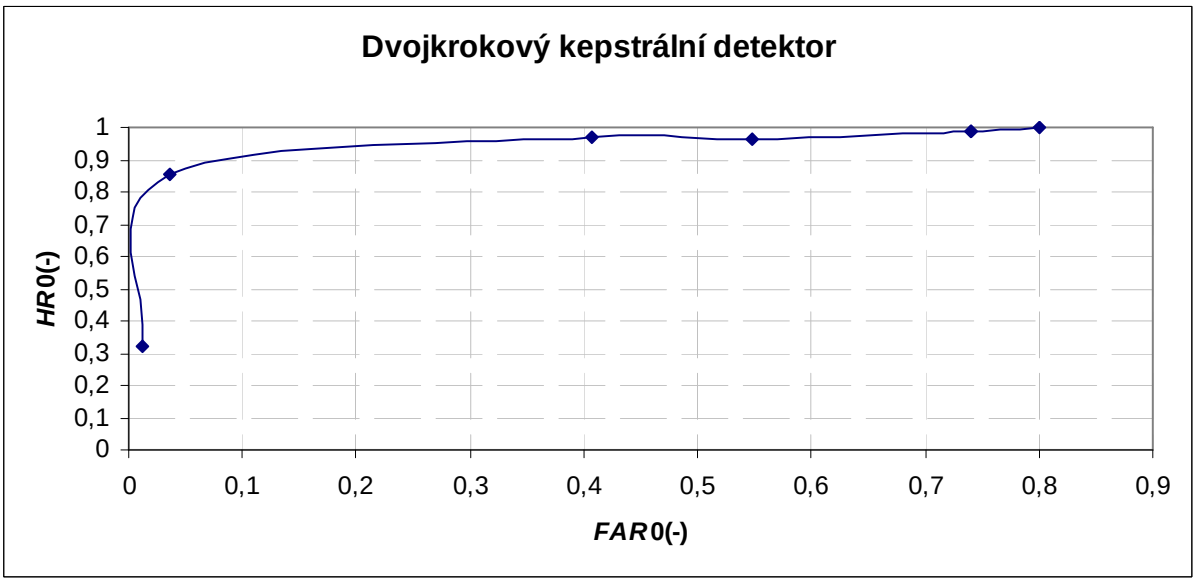
SEZNAM PŘÍLOH

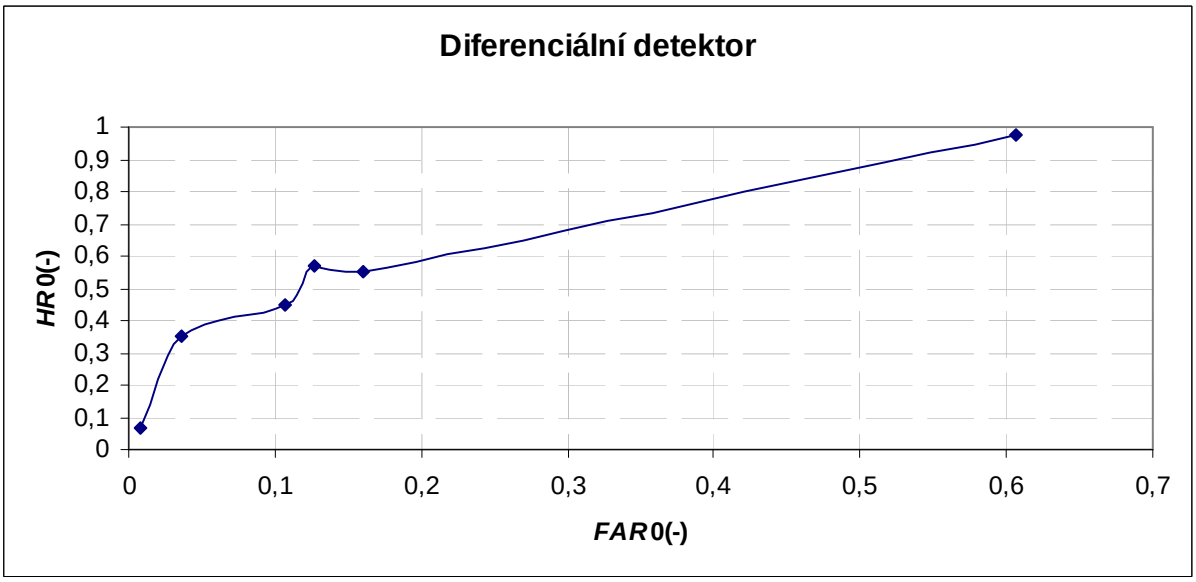
Příloha č. 1 ROC křivky pro jednotlivé detektory při použití preemfáze



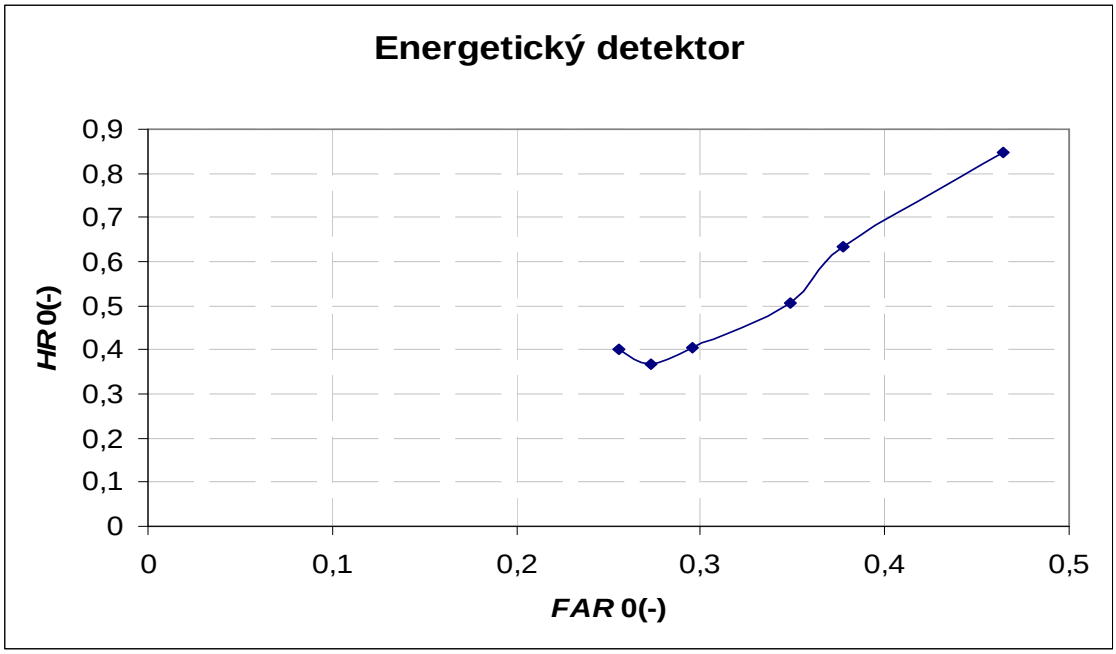


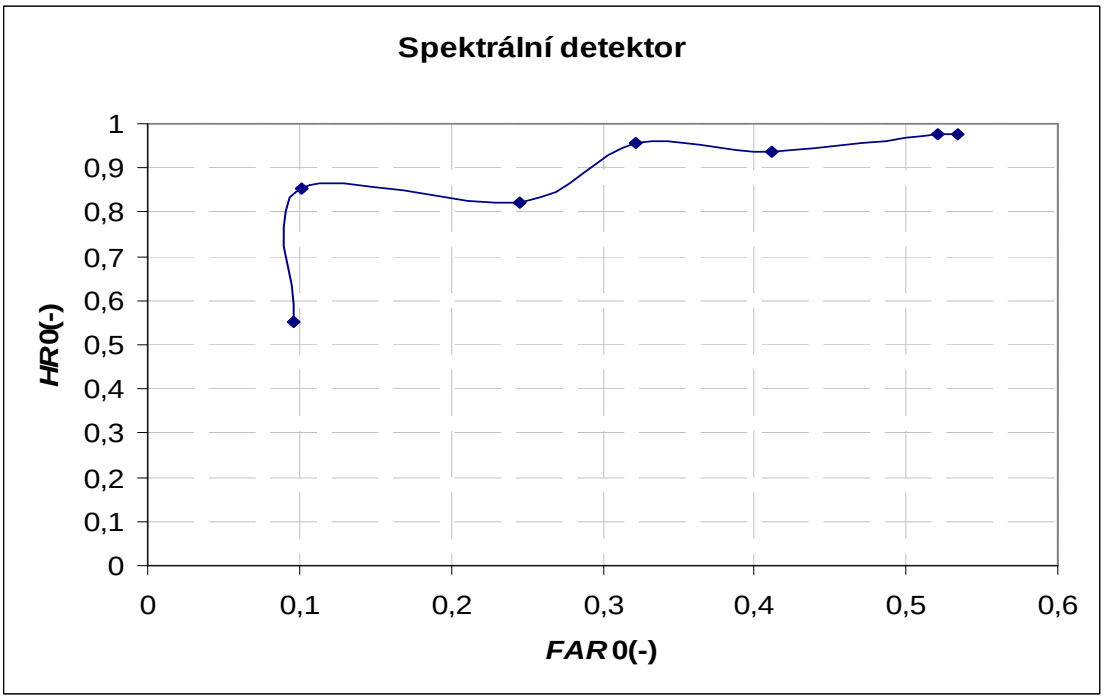

Jednokrokový kepstrální detektor
HR0(-)
FAR0(-)
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

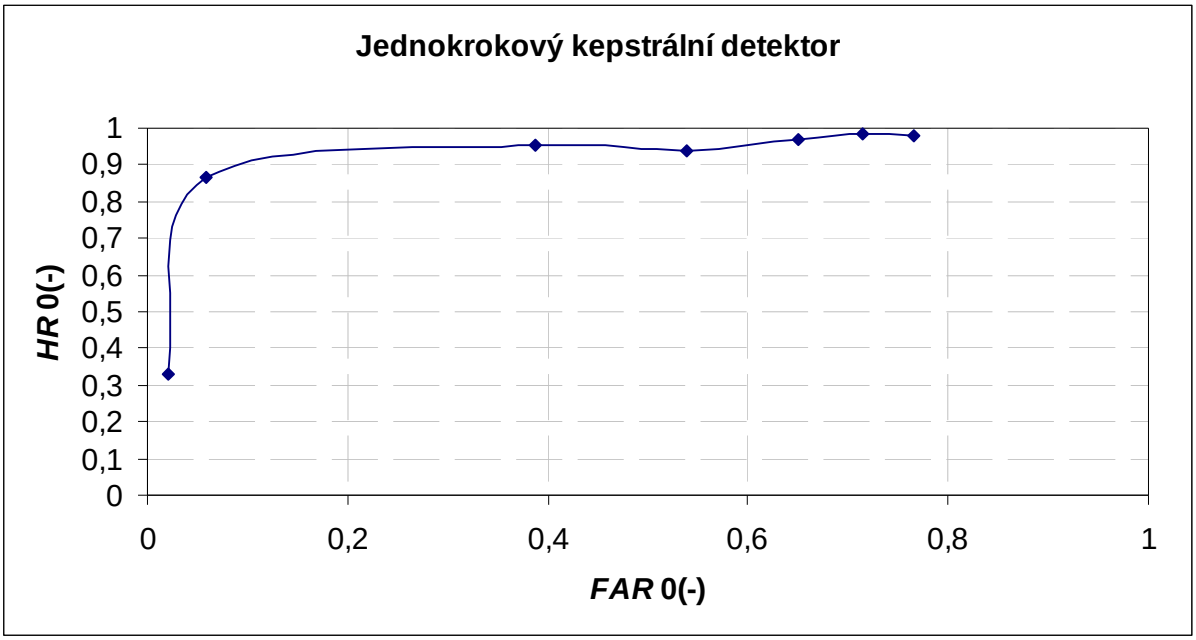

Dvojkrokový kepstrální detektor
HR0(-)
FAR0(-)
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
1

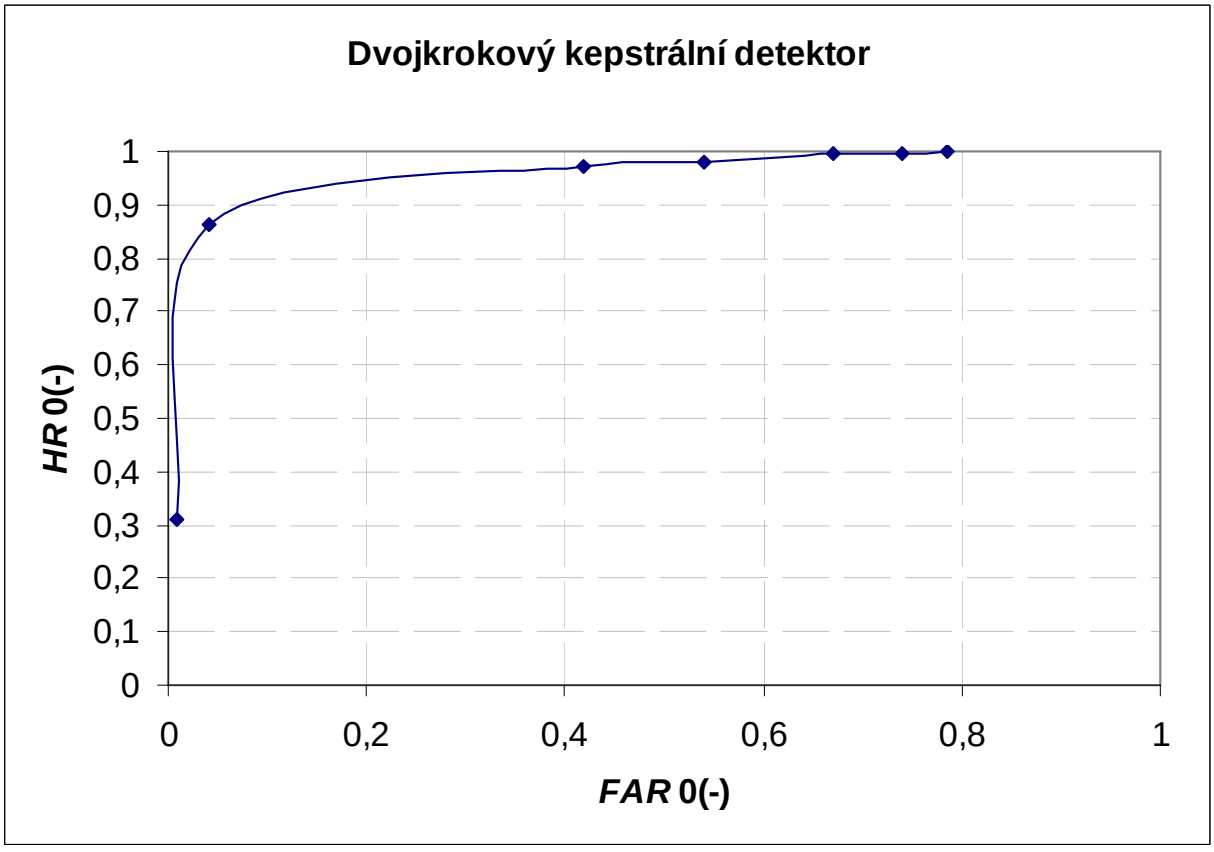

Diferenciální detektor
HR0(-)
FAR0(-)
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
1

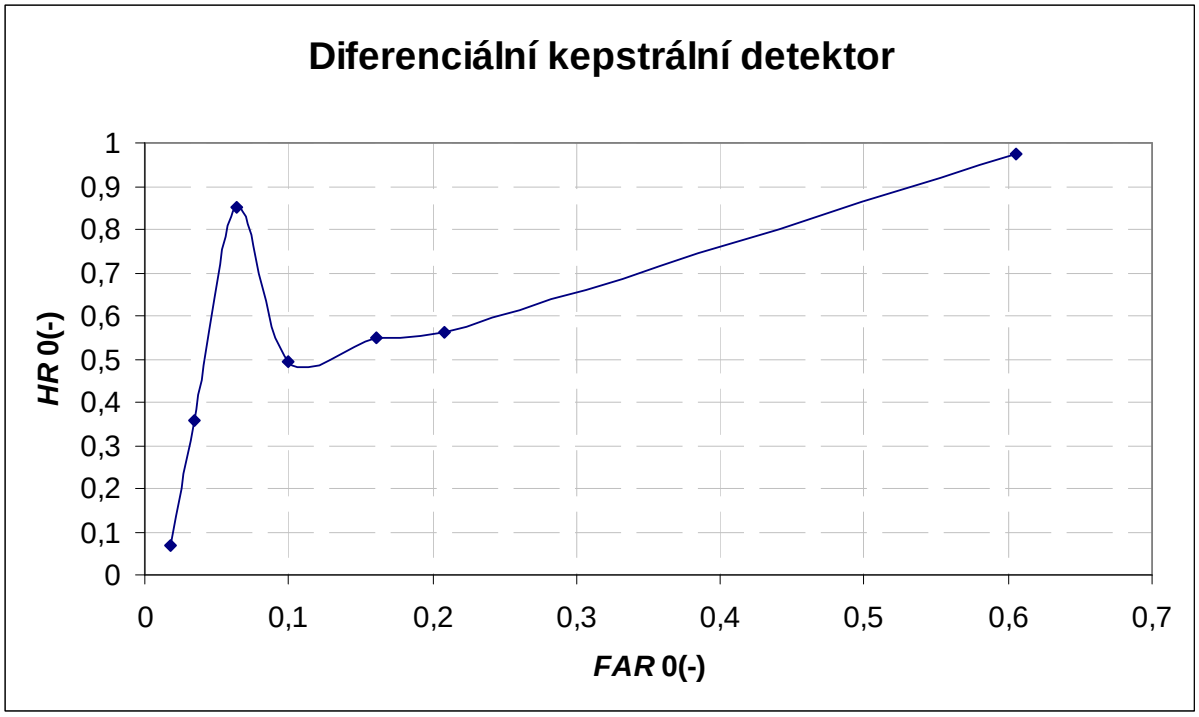
Příloha č. 2 ROC křivky pro jednotlivé detektory bez použití preemfáze






Jednokrokový kepstrální detektor
HR 0(-)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,2
0,4
0,6
0,8
1
FAR 0(-)


Dvojkrokový kepstrální detektor
HR 0(-)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,2
0,4
0,6
0,8
1
FAR 0(-)


Diferenciální kepstrální detektor
HR 0(-)
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
FAR 0(-)