Detektory řečové aktivity na bázi perceptivní kepstrální analýzy

J. Rajnoha, P. Pollák

České vysoké učení technické v Praze, Fakulta elektrotechnická
Technická 2, 166 27 Praha 6 - Dejvice

Abstrakt

Tento článek se zabývá popisem a implementací detektoru řečové aktivity (VAD) založeného na perceptivní kepstrální analýze řečového signálu. Kepstrální detektory vykazují zvýšenou robustnost vůči šumovému pozadí řeči v porovnání s jednoduššími algoritmy, např. energetickými. Perceptivní analýza řečového signálu realizovaná použitím vhodné banky filtrů s nelineární frekvenční osou pak lépe extrahuje příznaky řečového signálu použitelné pro tuto detekci.

Článek popisuje jednotlivé kroky algoritmu detekce s podrobnějším popisem významných bloků a jejich implementacemi v prostředí MATLAB.

Práce srovnává použitý detektor se standardním algoritmem používaným v hlasovém kodeku G.729. V závěru je diskutována možnost využití detektoru v různých aplikacích s příkladem použití detektoru v úloze robustního rozpoznávání řeči, které přineslo zlepšení úspěšnosti rozpoznání řeči téměř o 50 %.

1 Úvod

Detekce řeči je významnou součástí mnoha aplikací pro zpracování řeči. Nachází využití v systémech pro zvýrazňování řeči k aktualizaci parametrů modelu pozadí řeči, ve vokodéru pro přenos pouze řečového signálu a také v řečových rozpoznávačích pro detekci začátku a konce promluvy a pro odstranění neřečových částí signálu.

První skupinu tvoří algoritmy detekce řeči založené typicky na výkonové analýze signálu, spektrální či kepstrální analýze resp. koherenční analýze. Nejjednodušší formy detekce řeči zkoumají energii signálu nebo počet průchodů nulou [1], [2]. Jejich výhodou je velmi nízká výpočetní náročnost, naopak nevýhodou je vysoká chybovost v případě detekce řeči v šumovém prostředí. Spolehlivější algoritmy pro detekci jsou založeny na spektrálních (kepstrálních) vzdálenostech mezi řečovým signálem a pozadím řeči [3]. V případě zmíněných detektorů se obvykle zjišťuje míra odlišnosti daného bloku signálu od pozadí v dané oblasti (energie, spektrum, entropie spektra [4]). O vlastním výsledku detekce lze pak rozhodnout porovnáním této míry s prahovou hodnotou, kterou lze stanovit globálně jako fixní práh, či ji adaptivně obnovovat [5] podle aktuálních charakteristik pozadí řeči, případně lze využít více sofistikovaných rozhodovacích stromů [1]. Pro telekomunikační systémy se využívají algoritmy, které kombinují několik různých prvků pro zvýšení efektivity detekce [6], [7].

Pro prostředí s velmi vysokou hladinou rušení (např. jedoucí automobil) se dále používají vícekanálové metody [8]. S jejich nasazením ale současně vzrůstá výpočetní náročnost detekčních algoritmů a potřeba vícekanálového nahrávání zvyšuje také hardwarové požadavky systému. Další skupinu algoritmů tvoří detekce na bázi statistického zpracování signálu. Jedná se o přístup využívající GMM modelů řeči a šumu [2], případně umělých neuronových sítí pro nelineární mapování mezi vektorem řečových příznaků a přítomností řeči [9].

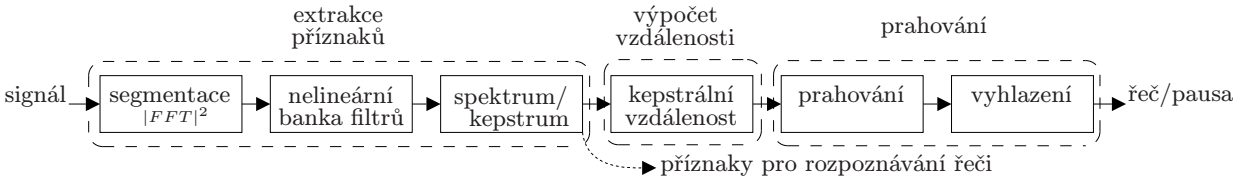
Tento článek se zabývá kepstrálními detektory, jejich implementací a vhodným předzpracováním signálu pro efektivní filtraci v zašuměném prostředí. Prezentované algoritmy jsou postaveny na struktuře standardního VAD systému. Použití jednotlivých algoritmů je inspirováno cílem využití VAD v úloze robustního rozpoznávání řeči.

Pro účely kepstrální analýzy řečového signálu se obvykle používají standardizované parametrizace MFCC – Melovské kepstrální koeficienty či PLP – percepční predikční koeficienty. Ty jsou založeny na principech tvorby řečového signálu v hlasovém traktu a vnímání řeči lidským

sluchem. Základem těchto metod je aplikace nelineárního zkreslení frekvenční osy ve spektrální oblasti a následný přechod do kepstrální oblasti. Článek ukazuje, jak je možné s využitím maticového zpracování dat v prostředí MATLAB efektivně realizovat tyto parametrizace.

2 Detekce řečové aktivity

Proces detekce řečového signálu lze rozdělit do tří základních kroků podle obr. 1. V prvním kroku je nalezena kepstrální reprezentace signálu. Ta umožňuje nejen snížit výpočetní náročnost algoritmu snížením objemu dat, ale také vystihnout nejdůležitější příznaky v řeči, vhodné pro další zpracování. Tyto kepstrální parametry bývají standardně využívány v procesu rozpoznávání řeči. Detektor založený na těchto příznacích tak nevyžaduje nadbytečný výpočet vhodné reprezentace signálu, pokud je dále využíván v ASR. Dalším krokem je výpočet vzdálenosti pro popis odlišnosti daného segmentu signálu od odhadu charakteristik pozadí řeči. Pro určení této odlišnosti se vychází přímo z kepstrálních koeficientů, nebo jsou použity diferenční kepstrální koeficienty, které jsou rovněž jednou z charakteristik signálu používaných pro rozpoznávání řeči. V posledním kroku je získaná vzdálenost porovnána s prahovou hodnotou pro konečné rozhodnutí o přítomnosti řeči. Tento výsledek může být mírně vyhlazen pro potlačení krátkých falešných skoků.



Obrázek 1: Blokové schema použitého detektoru řeči

V následujícím textu jsou podrobněji popsány jednotlivé kroky detekce řečové aktivity na bázi analýzy diferenčního kepstra signálu a rozebrána jejich realizace v prostředí MATLAB. Dále je uveden algoritmus integrální detekce, kde se vzdálenost počítá od charakteristik pozadí řeči.

2.1 Parametrizace signálu

Uvedený algoritmus detekce řečové aktivity je založen na kepstrální reprezentaci signálu ve formě PLP či MFCC koeficientů. Obě metody využívají obdobného výpočetního postupu, naznačeného v blokovém schematu na obr. 1 v části "extrakce příznaků". Postup lze rozdělit do následujících částí:

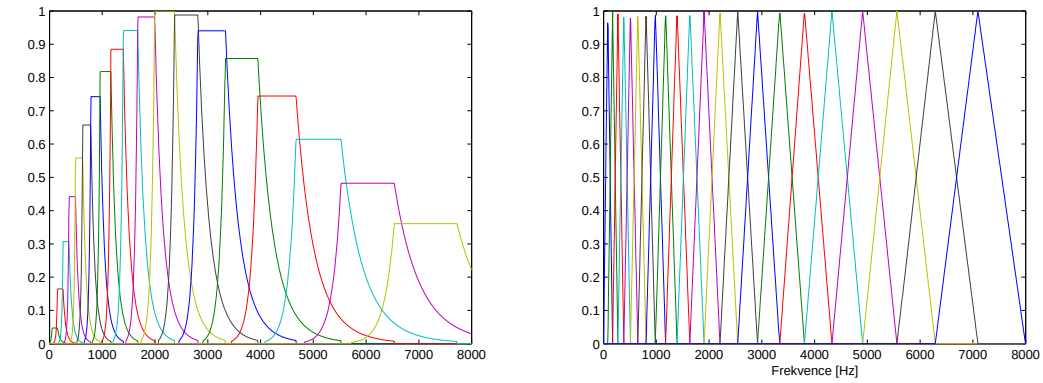
Krátkodobá Fourierova transformace signálu

Krátkodobá Fourierova transformace signálu je generována funkcí `spectrogram(...)`, která umožňuje uživatelsky definovat další parametry krátkodobé analýzy. Signál je rozdělen na segmenty o délce 25ms s krokem 10ms a váhován Hammingovým oknem. Z výstupní matice lze umocněním získat požadované výkonové spektrum.

Nelineární zkreslení frekvenční osy

Je známo, že vnímání tónů lidským uchem není lineárně závislé na frekvenci poslouchaného tónu. S měnící se frekvencí se mění vnímání zvuku, mimo jiné například subjektivní výška tónu nebo schopnost rozlišovat blízké tóny. Tuto nelinearitu je možné zahrnout do procesu výpočtu parametrů signálu rozdělením signálu do jednotlivých pásem, která odpovídají tzv. kritickým pásmům. Ta jsou definována na základě experimentálních měření. V dalším zpracování signálu se poté pracuje pouze s energií v jednotlivých pásmech. Vzhledem k rozdílné šířce jednotlivých pásem je často tato energie normována k šířce pásma, aby nedocházelo k zvýšení vlivu širších pásem na vyšších frekvencích oproti pásmům užším.

Při zpracování je nelineární filtrace realizována pomocí tzv. melovské (MFCC) resp. barkovy (PLP) banky filtrů. Tyto banky se liší zejména ve tvaru použitých filtrů a v šířce jednotlivých pásem především na vyšších frekvencích (viz obr. 2).



Obrázek 2: Banky filtrů pro analýzu řeči - barkova BF (vlevo), melovská BF (vpravo)

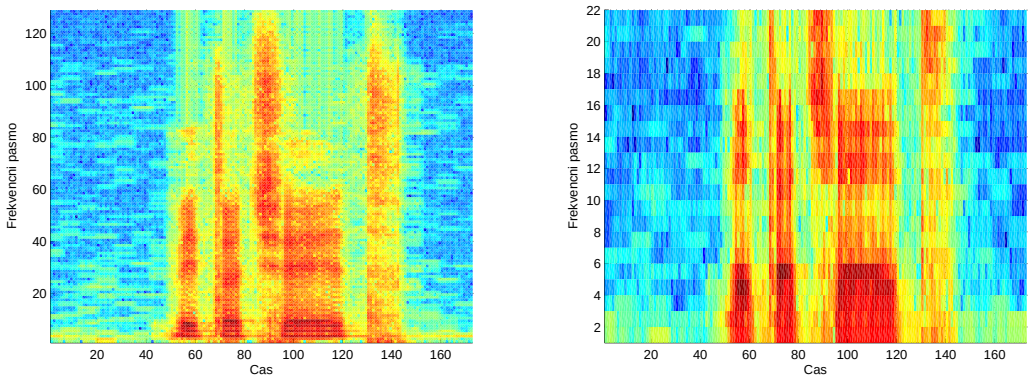
Funkce `spectrogram(...)` generuje matici, která obsahuje spektrální reprezentaci pro jednotlivé segmenty signálu. Máme-li definovánu banku filtrů jako matici M jednotlivých filtrů o délce, která odpovídá počtu vzorků spektra segmentu, lze získat spektrum frekvenčně zkresleného signálu jednoduchým součinem těchto matic. V MATLABu pak zápis nelineární analýzy signálu může vypadat takto:

```
% Vykonove spektrum puvodniho signalu
spg = abs(spectrogram(signal,N,OVER,NFFT,fs)).^2;

% BF - matice jednotlivych filtru melovske banky filtru,
%    - viz obr. 2

% Vykonove spektrum filtrovaneho signalu
spg_nelin = abs(sqrt(spg')*BF).^2;
% Pozn.: vystup spektrogramu dava matici, kterou je
% nutne pretocit pro spravne nasobeni matic
```

Výsledek zkreslení spektra ukazuje spektrogram na obr. 3.



Obrázek 3: Spektrogram řečového signálu před nelineární filtrací (vlevo) a po filtraci melovskou BF s 22 pásmy (vpravo)

V průběhu výpočtu parametrů je na signál možné aplikovat tzv. pre-emfáze (MFCC i PLP) a transformace intenzity zvuku na hlasitost (PLP). Tyto procesy způsobí potlačení dynamiky řeči a umožní tak lépe popsat signál výslednými koeficienty.

Kepstrální analýza

Pro přechod do kepstrální oblasti využívají zmíněné parametrizace rozdílný způsob. Melovské kepstrální koeficienty jsou získány zpětnou Fourierovou transformací logaritmu spektra. Vzhledem k reálnému a symetrickému výkonovému spektru je použito jednodušší DCT, jejíž aplikaci je možné opět provést maticově. Pro výpočet PLP kepstrálních koeficientů je použito LP analýzy signálu řádu Q. Tu lze získat po přechodu zpětnou Fourierovou transformací spektra do autokorelační oblasti. Následnou rekurzí jsou pak vypočteny hledané kepstrální koeficienty.

2.2 Kepstrální vzdálenost

Pro účely detekce řeči kepstrálním detektorem je obvykle využívána standardní kepstrální vzdálenost (CD) mezi aktuálním segmentem signálu a šumem pozadí. Potřebný odhad kepstra šumu pozadí je v tomto případě možno získat z průměrného kepstra signálu v pauzách řeči. Kepstrální vzdálenost pro i-tý segment lze vyjádřit jako

$$CD[i] = \sum_{k=1}^{p} (c_k[i] - \bar{c}_{o,k}[i])^2 \tag{1}$$

nebo v jednodušší podobě s nižší dynamikou jako

$$CD'[i] = \sum_{k=1}^{p} |c_k[i] - \bar{c}_{o,k}[i]| \tag{2}$$

pro p kepstrálních koeficientů. Detektory založené na této kepstrální vzdálenosti využívají výsledku detekce k získání průměrované kepstrální vzdálenosti $\bar{c}_{o,k}[i]$, což zavádí do algoritmu zpětnou vazbu. Tuto nevýhodu eliminuje použití kumulované kepstrální vzdálenosti (CDC) s využitím diferenční kepstrální analýzy.

Diferenční kepstrum je typicky aproximováno vztahem

$$\delta_k^{(M)}[i] = \left[\sum_{j=1}^{M} j(c_k[i+j] - c_k[i-j])\right] / \left[2\sum_{j=1}^{M} j^2\right] \tag{3}$$

kde M vyjadřuje řád odhadu diferenčního kepstra. Takto získaný odhad lze již použít pro stanovení vzdálenosti pomocí kumulativních součtů (integrace) jednotlivých diferenčních kepstrálních koeficientů jako

$$CDC_k[i] = \sum_{j=0}^{i} \delta_k^{(M)}[j]. \tag{4}$$

Vzhledem k nekorelovanosti jednotlivých kepstrálních koeficientů lze tyto koeficienty počítat jednotlivě a poté získat celkovou kepstrální vzdálenost jako součet jednotlivých vzdáleností

$$CDC[i] = \sum_{k=1}^{p} |CDC_k[i]| = \sum_{k=1}^{p} \left|\sum_{j=0}^{i} \delta_k^{(M)}[j]\right|. \tag{5}$$

Takto vyjádřená kepstrální vzdálenost pak odpovídá vztahu 2. V dalším textu bude pro jednoduchost CDC již značena jako CD.

Výše uvedená metoda pro určení kepstrální vzdálenosti přináší několik výhod. Použití diferenčních koeficientů zavádí vyhlazení výsledků, které omezí vliv náhodných výchylek. Metoda dále odstraňuje konstantní složku ze získané kepstrální vzdálenosti a účinně detekuje významné změny v signálu.

2.3 Prahování

Na základě získané kepstrální vzdálenosti lze již provést rozhodnutí o výsledku detekce. K tomu jsou typicky využívány dva základní přístupy k prahování:

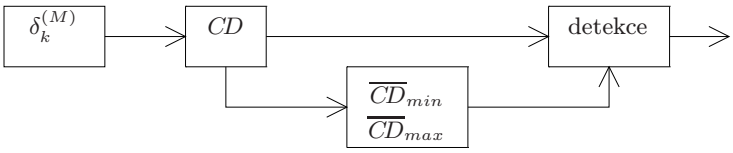
Dynamicky nastavený pevný práh (fixed)

Hodnota prahu THR_f

$$THR_f = \overline{CD}_{min} + \frac{p}{100}\,(\overline{CD}_{max} - \overline{CD}_{min}) \tag{6}$$

je určena na základě analýzy celé promluvy a toto počáteční nastavení prahu není v průběhu detekce řeči v dané promluvě modifikováno. Hodnota p je volena experimentálně na hodnotu typicky okolo 20 %.

Uvedené průměry $\overline{CD}_{min}$ a $\overline{CD}_{max}$ udávají průměr z 5 % nejnižších resp. nejvyšších hodnot získané kepstrální vzdálenosti CD. Toto průměrování umožní omezit vliv ojedinělých extrémních výchylek CD.



Obrázek 4: Blokové schéma pro algoritmus fixního prahování

Adaptivní práh řízený podle pozadí řeči (adapt)

Adaptivní práh THR_a

$$THR_a[i] = \bar{\mu}_{CD_n}[i] + 2\,\bar{\sigma}_{CD_n}[i] \tag{7}$$

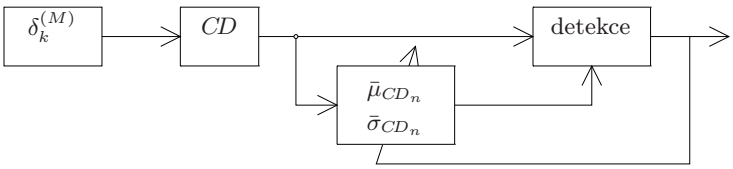
je nastaven podle střední hodnoty kepstrální vzdálenosti v pauzách řeči CD_n zvýšené o dvojnásobek standardní odchylky. To umožňuje postihnout variabilitu charakteristik prostředí. K obnově hodnoty THR_a dochází pro každý vyhodnocený segment pausy.

Adaptivní nastavení prahu detekce již zavádí do algoritmu detekce zpětnou vazbu, neboť k obnově hodnot prahu dochází v pauzách řeči, vyhodnocených daným detektorem. Hodnoty $\bar{\mu}_{CD_n}[i]$ a $\bar{\sigma}_{CD_n}[i]$ se obnovují pomocí exponenciálního průměrování s relativně dlouhou časovou konstantou pro zamezení vlivu náhodných výchylek v charakteristice pozadí řeči

$$\bar{\mu}[i+1] = q\,\bar{\mu}[i] + (1-q)\,CD[i] \tag{8}$$

pro hodnotu q typicky 0.95-0.98.

Adaptivní algoritmus vyžaduje na začátku promluvy krátký úsek pauzy, ve kterém se inicializují hodnoty prahu. V této fázi jsou první segmenty signálu využity pro nastavení hodnot $\bar{\mu}_{CD_n}[i]$ a $\bar{\sigma}_{CD_n}[i]$, které vystihují charakteristiky pozadí řeči.



Obrázek 5: Blokové schéma pro algoritmus adaptivního prahování

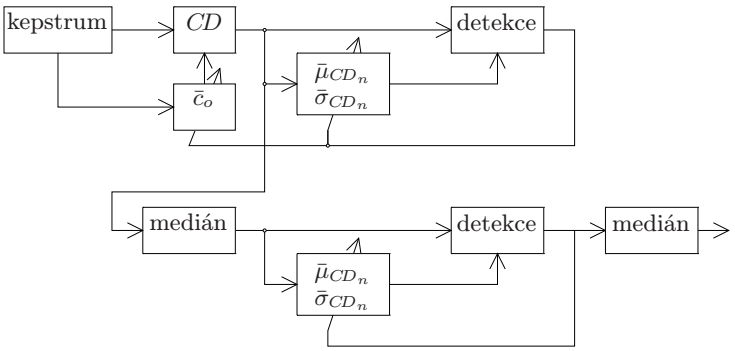
2.4 Vyhlazení výsledků detekce

Výstup detektoru často obsahuje krátké úseky falešné detekce řeči nebo pausy, které lze odstranit vyhlazením výsledků. To je prováděno typicky mediánovým filtrem. V našem algoritmu je použit filtr řádu 3, přičemž je možné použít i řád 5. Vyšší řád již může vést ke ztrátě informace a není proto vhodný. K určitému vyhlazení dochází navíc také již při výpočtu diferenčních kepstrálních koeficientů.

2.5 Integrální algoritmus s adaptivním prahováním

Výše uvedený adaptivní algoritmus zavádí zpětnou vazbu a podle výsledku detekce vyhodnocuje charakteristiky pauzy řeči. Tohoto postupu lze využít k adaptaci hodnoty $\bar{c}_{o,k}$ ze vztahu 1. V každém kroku detekce je tedy obnovena nejen hodnota prahu, ale také průměrného kepstra pozadí řeči, s nímž je srovnáván vyhodnocovaný signál.

V první části algoritmu dochází k nalezení průměrného kepstra pro určení kepstrální vzdálenosti *CD* a charakteristik pozadí $\bar{\mu}_{CD_n}$ a $\bar{\sigma}_{CD_n}$ pro druhou fázi algoritmu, v níž jsou vyhlazené hodnoty kepstrální vzdálenosti zpracovány adaptivním algoritmem prahování.



Obrázek 6: Blokové schéma integrálního detektoru

3 Možnosti nastavení algoritmů s ohledem na aplikaci

Jak je ukázáno v předchozím textu, celý detekční algoritmus má mnoho možných variant a nastavení, jimiž lze ovlivnit výsledné vlastnosti detektoru, které jsou ve velké míře určeny konečným využitím VAD. Například pro použití v rozpoznávačích řeči je důležité, aby na základě chybné detekce nebyly odstraňovány řečové segmenty, které jsou dále potřebné pro vlastní rozpoznávač. Naopak při detekci pauz v řeči, které jsou využity pro následnou úpravu charakteristik šumového pozadí pro metody zvýrazňování řeči, je důležité korektní označení segmentů pauzy.

Volbou vhodné parametrizace lze docílit již zmíněného vyhlazení, které je potřebné pro odstranění nejvýraznějších odchylek v signálu a tím vzniklých případných falešných detekcí. To provádí i uvedené parametrizace MFCC a PLP použitou nelineární filtrací. V případě PLP je navíc vyhlazení podpořeno provedenou LP-analýzou. Následný výpočet kepstrální vzdálenosti založený na diferenčním kepstru ovlivňuje míru vyhlazení řádem M aproximace diferenčního kepstra $\delta_k^{(M)}$. Zvýšením řádu se docílí vyššího vyhlazení, které ale může způsobit ztrátu informace v podobě nedetekovaných krátkých segmentů. Podobně může ovlivnit výsledné vlastnosti systému volba řádu mediánové filtrace výsledné detekce.

Pro různé aplikace detektoru je možné volit také algoritmus prahování. Pro on-line zpracování, typické při detekci začátku a konce promluvy pro ASR, je nutné použít adaptivní algoritmus. Zde není známa informace o celém signálu a tedy jeho dynamice, podle níž se určuje hodnota prahu pro detekci. Tyto adaptivní algoritmy jsou však velmi citlivé na nastavení parametrů adaptace a pro obecné šumové podmínky není možné zajistit jejich optimální volbu. Je tedy důležité tyto parametry volit velmi pečlivě.

Na druhou stranu fixní prahování poskytuje výhodné vlastnosti pro off-line zpracování signálu, například při trénování rozpoznávače řeči nebo při blokovém zpracování řečových dat po detekci začátku a konce promluvy.

4 Experimentální část

Výše uvedené algoritmy byly porovnány na úrovni chyby detekce řečových segmentů a segmentů pauzy podle referenčních dat. Pro obecné srovnání byl použit i algoritmus detekce používaný v hlasovém kodeku G.729. Jednotlivé detektory jsou také srovnány v experimentech, které využívají VAD při parametrizaci řečového signálu pro účely automatického rozpoznávání řeči.

4.1 Přesnost detekce řečové aktivity

Pro testování přesnosti detekce řeči byly použity dvě rozdílné sady dat na bázi české databáze SPEECON:

sada 1 - 150 vět od různých mluvčích v rozdílných šumových podmínkách s ručně označenými hranicemi hlásek

sada 2 - cca. 300 promluv obsahujících izolované číslovky s výraznější pauzou mezi slovy. Referenční detekce řečové aktivity je generována automatickým zarovnáním hlásek pomocí rozpoznávače řeči na bázi HMM.

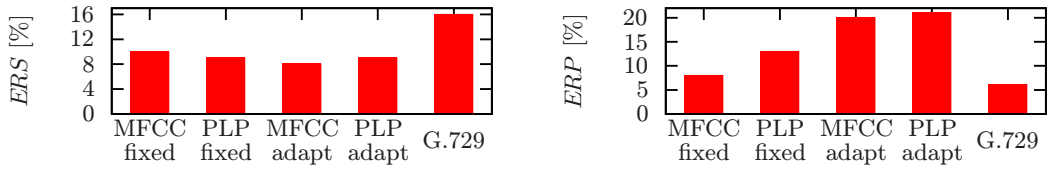
Testy byly provedeny pro obě parametrizace PLP i MFCC s použitím fixního i adaptivního nastavení prahu detekce (fixed, adapt). Tyto algoritmy byly srovnány s algoritmem ITU-T G.729.

Pro posouzení přesnosti detekce byla zvolena následující kritéria: chyba detekce řeči (ERS) a chyba detekce pausy (ERP).

$$ERS = \frac{1}{L_S} \sum_{i=0}^{L-1} (\mathrm{vad}_{ref}[i] - \mathrm{vad}[i])\mathrm{vad}_{ref}[i] \tag{9}$$

$$ERP = \frac{1}{L_P} \sum_{i=0}^{L-1} (\mathrm{vad}[i] - \mathrm{vad}_{ref}[i])(1 - \mathrm{vad}_{ref}[i]) \tag{10}$$

kde L je celkový počet segmentů, L_S a L_P počet segmentů řeči a pauzy.



Obrázek 7: Průměrné ERS (vlevo) a ERP (vpravo) pro obě testovací sady dat

Srovnání prezentovaných detektorů s algoritmem založeným na standardu G.729 ukazuje lepší výsledky detekce řečových segmentů pro obě volby parametrizace i prahování. Vzhledem ke schopnosti ASR systému modelovat pauzu v řeči by v případě použití detektoru pro ASR systém byla vyšší chyba detekce pauzy méně významná.

4.2 Detekce řeči pro ASR

V dalších experimentech byly srovnány jednotlivé algoritmy z pohledu přínosu k úloze rozpoznávání řeči. K tomu bylo použito rozpoznávače sekvence číslovek založeného na skrytých Markovovských modelech kontextově nezávislých fonémů.

Pro posouzení výsledku rozpoznávání bylo použito standardního kritéria

$$WER = (S + D + I)/N \times 100\,\% \tag{11}$$

kde N, D,S a I odpovídají celkovému počtu rozpoznávaných slov, počtu smazaných, zaměněných a chybně vložených slov.

V prvním kroku byly srovnány algoritmy s fixním a adaptivním prahem. Pro tento účel byl rozpoznávač natrénován na části databáze SPEECON z méně hlučného prostředí (kancelářské prostory). Detekce řeči je prováděna pouze v trénovací fázi.

	MFCC		PLP	
bez detekce	5.33		3.33	
fixní práh	5.33	(0 %)	2.49	(25.2 %)
adaptivní práh	2.97	(44.3 %)	2.13	(36.0 %)

Tabulka 1: *WER* a jeho relativní zlepšení vztažené k zpracování bez VAD pro různá nastavení algoritmu prahování v čistých šumových podmínkách

Tabulka chybovosti rozpoznávače 1 ukazuje, že použití detektoru přináší významné zvýšení přesnosti rozpoznávání. Volba adaptivního prahu vedla ke snížení chybovosti až o 44 % pro parametrizaci MFCC oproti případu bez použití detekce v rámci parametrizace signálu. Nižší zlepšení pro fixní nastavení prahu detekce může být způsobeno variabilitou charakteristik prostředí v rámci zkoumané nahrávky, kterou fixní práh nemůže postihnout.

Pro druhý experiment byl použit algoritmus s fixním prahem a integrální detektor. Rozpoznávač byl trénován i testován na celé databázi SPEECON, zároveň byly použity dva různé přenosové kanály - kvalitnější řečový signál z head-set mikrofonu a signál obsahující vyšší hladinu šumu pozadí z hands-free mikrofonu. VAD byl použit v trénovací fázi i při vlastním rozpoznávání.

	MFCC		PLP	
	head-set	hands-free	head-set	hands-free
bez detekce	9.05	10.69	8.82	14.21
fixní práh	6.72 (25.74 %)	9.10 (14.87 %)	4.75 (46.15 %)	7.31 (48.56 %)
integrální detektor	7.36 (18.67 %)	8.50 (20.49 %)	7.31 (17.12 %)	8.59 (39.55 %)

Tabulka 2: *WER* a jeho relativní zlepšení vztažené k zpracování bez VAD pro jednotlivé algoritmy detekce v obecném šumovém prostředí

Z výsledků rozpoznávání v tabulce 2 je zřejmé, že oba typy detekce přináší zlepšení. Vyšší míra zlepšení výsledků rozpoznávače pro fixní práh může být způsobena méně snadnou možností optimalizace adaptačních konstant integrálního algoritmu.

5 Shrnutí

Článek prezentuje implementaci a použití detektoru řečové aktivity založeného na analýze příznaků používaných při rozpoznávání řeči. To umožňuje jeho snadnou aplikaci v ASR systému, jeho využití je ale možné i v dalších oblastech, např. kódování řeči nebo v algoritmech pro zvýrazňování řeči. V následujících bodech jsou shrnuty nejdůležitější závěry:

- Kepstrální analýza založená na MFCC nebo PLP umožňuje lépe rozlišit řečovou aktivitu a šum pozadí, než čistá kepstrální nebo LPC analýza.
- Efektivní aplikace výpočetních algoritmů v maticové podobě (nelineární banky filtrů, DCT) zvyšuje efektivitu výpočtu parametrizačních koeficientů.

- Použití kumulovaných součtů pro výpočet kepstrální vzdálenosti bez nutnosti výpočtu průměrného kepstra pozadí odstraňuje zpětnou vazbu z výpočetního algoritmu.
- Popsané detekční algoritmy vedou v porovnání s detektorem G.729B na nižší chyby detekce řeči. Algoritmus s fixním prahem, vhodný pro off-line zpracování signálu (např. pro trénování ASR) pak dosahuje celkově nejlepších výsledků ze zkoumaných detektorů.
- Použití detekčních algoritmů v parametrizaci signálu pro úlohu rozpoznávání řeči přináší výrazné zlepšení přesnosti rozpoznávání. V prostředí bez šumového pozadí došlo ke snížení chyby až o 44 %, pro obecné šumové podmínky bylo zlepšení téměř o 49 %.

6 Poděkování

Tento výzkum byl podporován granty GAČR 102/08/0707 "Rozpoznávání mluvené řeči v reálných podmínkách", GAČR 102/08/H008 "Analýza a modelování biologických a řečových signálů", výzkumným záměrem MSM 6840770014 "Výzkum perspektivních informačních a komunikačních technologií".

Reference

[1] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. SAP-10, no. 2, pp. 109–118, FEB 2002, ISSN 1063-6676.

[2] Y. Kida and T. Kawahara, "Evaluation of voice activity detection by combining multiple features with weight adaptation," in *Proc. of Interspeech 2006, 9-th International conference on Spoken Language Processing*, Pitssburgh, Sep 2006.

[3] J. A. Haigh and J. S. Mason, "A voice activity detector based on cepstral analysis.," in *Eurospeech'93 - Proceedings of the 3rd European Conference on Speech, Communication, and Technology*, Berlin, Sept. 1993, pp. 1103–1106.

[4] Z. Tüske, Péter Mihajlik, Z. Tobler, and T. Fegyó, "Robust voice activity detection based on the entropy of noise-suppressed spectrum," in *Proc. of Interspeech 2005, 9-th European Conference on Speech Communication and Technology*, Lisbon, Sep 2005.

[5] P. Sovka and P. Pollák, "The study of speech/pause detectors for speech enhancements methods," in *EUROSPEECH'95 - Proceedings of the 4th European Conference on Speech Communication and Technology*, Madrid, Spain, September 1995, pp. 1575–1578.

[6] ITU, "International Telecommunication Union Recommendation G.729, annex b - A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70," 1996.

[7] ETSI, "European Standard EN 300 965 - digital cellular telecommunications system (Phase 2+); Full rate speech; Voice Activity Detector (VAD) for full rate speech traffic channels," 2000.

[8] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, and V. Gilg, "Multichannel voice detection in adverse environments," in *Proc. of EUSIPCO 2002*, Toulouse, France, Sep 2002.

[9] R. Gemello, F. Mana, and R. De Mori, "Non-linear estimation of voice activity to improve automatic recognition of noisy speech," in *Proc. of Interspeech 2005, 9-th European Conference on Speech Communication and Technology*, Lisbon, Sep 2005.

Josef Rajnoha
rajnoj1@fel.cvut.cz

Petr Pollák
pollak@fel.cvut.cz