VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Ing. Martin Plšek

EXTRAKCE ŘEČOVÉHO SIGNÁLU Z HLUKU POZADÍ VE SPEKTRÁLNÍ OBLASTI

SPEECH SIGNAL EXTRACTION FROM NOISY BACKGROUND IN SPECTRAL DOMAIN

ZKRÁCENÁ VERZE PH.D. THESIS

Obor: Teleinformatika

Školitel: Prof. Ing. Zdeněk Smékal, CSc.

Oponenti: Prof. Ing. Andrej Lúč, CSc.

Ing. Robert Vích, DrSc.

Datum obhajoby: 6.5.2005

KLÍČOVÁ SLOVA

řečový signál, potlačování šumu v řeči, spektrogram, mapování spektrogramu

KEYWORDS

speech signal, noise suppression in speech signals, spectrogram, mapping of spectrogram

Disertační práce je uložena na oddělení vědy a výzkumu FEKT VUT v Brně, Údolní 53, Brno, 602 00.

ISBN 80-214-2940-2
ISSN 1213-4198

OBSAH

1 ÚVOD

S masivním nárůstem mobilních komunikačních systémů jsou na kvalitu přenášeného hlasu kladeny čím dále tím větší požadavky. Jedním z hlavních parametrů hodnocení srozumitelnosti lidské řeči po průchodu přenosovým kanálem je odstup signálu od šumu, nebo přesněji od hluku pozadí libovolného charakteru.

Vstupním prvkem do každého hlasového komunikačního systému je mikrofon, který v reálných akustických prostředích snímá i okolní hluk. Hlukem může být například ruch křižovatky, zvuky motoru dopravních prostředků, hlučná místnost, přechodné impulzní hluky a šumy, zvuky zvířat nebo dokonce i současná promluva několika mluvčích nacházejících se v blízkosti elektroakustického snímače. Pokud není vstupní akustický signál ihned převeden do číslicové podoby, je dále degradován tepelnými širokopásmovými šumy pasivních i aktivních prvků přenosového zařízení, vnějším elektromagnetickým polem, přeslechy a impulzním rušením různého původu, které se přes parazitní kapacitní a induktivní vazby velmi snadno váže na užitečný signál. Na výstupu komunikačního řetězce potom může úroveň hluků přerůst úroveň samotné řeči.

Separační techniky nacházejí v převážné míře uplatnění v komunikačních systémech, jakými jsou například mobilní telefony, radiová pojítka dopravních prostředků a centrálního dispečinku, spojovací prostředky ozbrojených sil, kdy je zapotřebí zabezpečit komunikaci i za velmi ztížených podmínek apod. Cílovou skupinou mohou být také hudební studia zabývající se restaurací archivních zvukových nahrávek a v neposlední řadě také kriminalistické ústavy.

Velký odstup signálu od hluku pozadí není vyžadován pouze příjemcem zprávy, kterým byl do nedávné doby výhradně člověk, ale také číslicovými algoritmy nové generace, které v reálném čase analyzují řečový signál přicházející z mikrofonu a na základě dekódovaných povelů provádějí naprogramovanou činnost. Jedná se například o systémy hlasového ovládání telekomunikačních zařízení, pracovních strojů, systémů dopravních prostředků, ale i domácích elektrospotřebičů. Oblast využití daného zařízení je vždy výrazně rozšířena po zařazení bloku extrakce parazitních hluků před vlastní zpracování řečového signálu.

Prudký rozvoj číslicových metod zpracování řečových signálů v reálném čase je možný díky značnému pokroku ve vývoji nových integrovaných digitálních signálových procesorů, kde výpočetní výkon za posledních několik let vzrostl na takovou míru, že již lze s relativně malými ekonomickými náklady realizovat složitá zařízení číslicového zpracování signálů.

2 SOUČASNÝ STAV ŘEŠENÉ PROBLEMATIKY

2.1 OBECNÝ POPIS

Následující část poskytuje obecný náhled na separační techniky používané v současné době ve světě. Metody analýzy separačních technik většinou spočívají v nalezení množiny vhodných parametrů, charakteristických právě pro řečový signál. V některých případech je výhodnější postupovat obráceně, tedy nalézt přesný model parazitního signálu, který je ovšem velmi variabilní pro různé druhy rušení. Parametry obou způsobů zpracování se zpravidla určují z následujících typů analýz:

Klasické techniky analýzy:

- kmitočtová analýza, popř. časově - kmitočtová analýza [1], [8], [9], [11], [13],
- homomorfní analýza, popř. časově - kepstrální analýza [9], [14], [18],
- predikční techniky (LPC modely) [8], [9], [15], [28],
- statistické analýzy (neuronové sítě) [24], [27],

Nové techniky analýzy:

- modelování pomocí simulace sluchových a nervových orgánů [12], [19],
- vlnková (waveletová) transformace a její modifikace [18], [29],
- technika TESPAR založená na přesné matematické reprezentaci signálů [25],
- ostatní spektrální analýzy (např. Pronyho metoda) [30], [35],
- vícekanálové techniky (analýza energetických poměrů a fázových posuvů z množiny snímačů) [12], [20], [21], [23].

Při výběru konkrétní techniky je nutné nejprve posoudit schopnost metody vytvořit takovou transformaci degradovaného řečového signálu, která naznačuje možný výskyt řeči v daném hluku. K aplikaci v inženýrské praxi jsou vhodné jen takové postupy, které jsou schopné nejen detekovat řečovou aktivitu, ale také provést její přesnou lokalizaci v časové i kmitočtové rovině.

2.2 ZÁKLADNÍ DĚLENÍ SEPARAČNÍCH TECHNIK

Techniky pro zvýraznění řeči skryté v šumu jsou v současné době již nedílnou součástí každého komunikačního zařízení. Jedná se většinou o jednoduché algoritmy s malou účinností separace a velkým omezením na charakter parazitních hluků. Každoročně jsou na celém světě do vývoje nových účinnějších systémů vynakládány nemalé finanční prostředky, ale v konkurenčním boji může zvítězit pouze takový algoritmus, který dokáže automaticky, bez vnějšího zásahu uživatele, s minimální výpočetní náročností a minimem proměnných parametrů, odstranit nejběžnější hluky

vyskytující se v okolí člověka. Při detailnějším rozboru však zjistíme, že právě tyto přirozené zvuky jsou na potlačení nejnáročnější. Algoritmus musí být proto schopen patřičně reagovat jak na širokopásmové šumy, impulzní hluky a trvalé rušící periodické signály, tak i na nežádoucí lidskou řeč v pozadí. Zejména pak musí být systém schopen zvýraznit lidskou řeč ze značně zarušeného prostředí, kdy je zcela skryta v hluku a stává se nesrozumitelnou.

Metody pro potlačení parazitního rušení v řečovém signálu lze v základu rozdělit na jednokanálové a vícekanálové [10]. V současné době jsou v inženýrské praxi většinou používány jednodušší jednokanálové techniky založené na potlačení vyšších, popřípadě nižších kmitočtů spektra vstupního signálu v závislosti na odstupu signálu od rušení. Je zřejmé, že nežádoucí hluk v samotné řeči zůstává, pouze je vlivem omezené šířky přenášeného pásma pro lidský sluch méně nápadný. Při výrazném překrytí užitečného signálu nežádoucím hlukem dochází ke snížení srozumitelnosti výsledné řeči, neboť nastává omezení přenášeného kmitočtového pásma samotné řeči.

Podstatně dokonalejšími jednokanálovými metodami jsou algoritmy založené na tzv. spektrálním odečítání parazitních hluků od směsi řeči a hluku [3], [5], [12]. Princip je založen na odhadu výkonového spektra rušení z krátkodobých úseků signálu, ze kterých je následně vypočítána prahová úroveň pro odlišení užitečných částí signálu od hluku pozadí. Účinnost separace je přímo závislá na přesnosti odhadu výkonových parametrů rušení, které jsou určovány pomocí detektorů řečové aktivity v řečových pauzách [22], [26]. Uvedený základní princip odečítacích algoritmů, označovaný zkratkou PSS (Power Spectral Subtraction) [12], vyžaduje stacionární chování hlukových signálů. Nevýhodou je, že výsledná nahrávka často obsahuje výrazný zbytkový šum. V současné době proto existuje mnoho modifikovaných technik založených právě na principu PSS, ovšem s odlišným přístupem k odhadu výkonových vlastností rušení. Jedná se např. o nelineární metodu NSS (Nonliear Spectral Subtraction) [12] nebo např. metodu MMSE STSA (Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator) [12], popř. její vylepšenou verzi MMSE Log-STSA (Minimum Mean-Square Error Short-Time Log-Spectral Amplitude Estimator) [12]. V současné době se odečítací techniky zdokonalují o fyziologické vlastnosti lidského sluchu. Na základě psychoakustického modelu lidského slyšení dochází k dynamické volbě separační prahové úrovně v závislosti na tzv. kmitočtovém maskovacím prahu. Jedná se např. o perspektivní techniku SS-PMHE (Spectral Subtraction with Physiological Model of Human Ear) [12]. Zdokonalené metody spektrálního odečítání jsou většinou ve stadiu vývoje a přes jisté úspěchy nedosahují, vzhledem ke svým výpočetním nárokům, velkých separačních účinností.

Výpočetně méně náročnou a světově rozšířenou jednokanálovou metodou, užívanou firmou MOTOROLA v mobilních telefonech pro jednotky „handsfree“, je technika RASTA [1], [2] (název je odvozen z anglické zkratky RelAtive SpecTrAl), vyvinutá prof. Hermanským a jeho týmem v letech 1994-1995 v USA. Metoda využívá ke své práci adaptivní pásmovou filtraci časově proměnného spektra

vstupního signálu, tzv. časových trajektorií [1]. Potlačení trvalých a rychle se v čase měnících parazitních signálů je dostatečné, ovšem pro většinu běžných hluků není separace příliš účinná. Při malých hodnotách poměru výkonu signálu k výkonu rušení dochází k násilné modulaci spekter vstupního signálu a výstupní řeč se stává spíše nepříjemnou.

Další perspektivní jednokanálové techniky jsou založeny na tzv. vlnkové (waveletové) transformaci [29], [31]. Jedná se o poměrně novou metodu analýzy signálů, založenou na teorii tzv. Hilbertových prostorů. Její základy byly položeny již před desítkami let, ale bouřlivý vývoj dílčích algoritmů nastal až počátkem osmdesátých let 20. století, kdy se teorie vlnkové transformace začala systematizovat a byl vyvinut odpovídající matematický aparát. Vlnková transformace, podobně jako Fourierova transformace, je založena na zobrazení analyzovaného signálu do prostoru bázových funkcí (analogie spektra signálu). Oproti rozvoji do Fourierovy řady se však nepoužívá trigonometrických funkcí, nýbrž tzv. mateřský wavelet, jehož translací (změnou polohy) a dilatací (změnou šířky) vytvoříme bázi celého prostoru. Své uplatnění může vlnková transformace nacházet v oborech jako je seismologie, meteorologie, lékařství, telekomunikace atd., kde se často používá pro analýzu a rekonstrukci hlukem znehodnocených signálů. Aby mohly techniky založené na vlnkové transformaci úspěšně potlačit nežádoucí rušení z řečového signálu, je nutná poměrně dobrá znalost charakteru rušení a tím i znalost nejvýhodnější bázové funkce.

Odlišným přístupem k problematice odhlučnění řečových signálů je vývoj vícekanálových metod schopných pracovat v reálném čase. Nejstarší vícekanálová metoda je již dlouhou dobu používána v televizních přenosech mezi moderátorem a přenosovým vozem [20]. Hlas publicisty je v hlučném prostředí snímán směrovým mikrofonem, kterým je převeden na elektrický signál. Hluk okolí se snímá ve vhodné vzdálenosti druhým mikrofonem s všesměrovou charakteristikou. Vzniklé elektrické signály se analogově nebo číslicově odečtou. Aby metoda spolehlivě pracovala, musí zdroj rušení vytvářet v daném prostoru tzv. difúzní (rozptýlené) akustické pole, což není vždy splněno. Tato metoda však není příhodná pro běžné použití.

Moderní číslicové zpracování dovoluje použít algoritmy využívající vzájemných fázových posuvů a amplitudových poměrů elektrických signálů snímaných z množiny vhodně umístěných elektroakustických měničů [12], [20], [21]. Poměrně složitými matematickými operacemi dochází k oddělení hluku pozadí od užitečné řeči. Jedná se např. o perspektivní techniky BSS (Blind Source Separation), jejichž cílem je obnova původních signálů z kompoziční směsi pořízené v reálném akustickém prostředí. Celý problém je značné závislý na konkrétních podmínkách šíření akustické vlny od množiny zdrojových signálů k přijímací soustavě mikrofonů. Obecně lze říci, že vícekanálové techniky jsou díky velkým nárokům na požadovaný výpočetní výkon pouze ve vývoji. K jejím nevýhodám patří nutnost realizovat v místě mluvčího poměrně složitou přijímací soustavu, a proto tyto systémy také zatím nejsou vhodné pro běžnou mobilní komunikaci.

3 CÍLE DISERTACE

Záměrem předkládané práce je návrh a optimalizace nové digitální technologie separace řeči skryté v silném hluku. Hlavní myšlenkou bylo vytvořit soubor algoritmů pro účinné potlačení parazitního rušení z užitečné řeči na základě znalosti stavby hlasového traktu a fyziologických vlastností lidského sluchu. Z principů generování řeči byly hledány identifikační příznaky pro separaci hlásek ze silně znehodnocené směsi řeči a hluku. Inspirací byly v dnešní době nejperspektivnější techniky extrakce hlukových signálů, založené zejména na časově-kmitočtové analýze.

Výstupem řešení je zcela původní digitální technologie separace řeči zlepšující užitné energetické vlastnosti na výstupu algoritmu. Navržená metoda je založena na principiálně zcela novém přístupu k problému extrakce hlukových signálů, který je, jak se ukázalo, mnohem více účinný nežli standardní způsoby řešení realizované pomocí jednorozměrné lineární číslicové filtrace.

Nedílnou součástí vyvinuté technologie je popis a zobecnění základních principů separace, založených na analýze a následném zpracování časově-kmitočtového prostoru (spektrogramu). Pojetí spektrogramu jako dvojrozměrného celku zatím nedovolovalo účinně extrahovat hlukové signály, které v čase vykazují nestacionární charakter. Dalším nedostatkem dnešních separačních algoritmů je obtížná identifikace řeči v rozmanité směsi hluků (od širokopásmových šumů různých rozložení pravděpodobností až k limitnímu případu harmonického rušení). Při vývoji byl také kladen velký důraz na univerzálnost navržené techniky vzhledem k odlišnému charakteru možných hlukových signálů.

Nemalá část disertace je též věnovaná dodatečnému zvýšení srozumitelnosti výstupního řečového signálu zpracováním v časové a energetické rovině. Komplexně navržená metoda byla podrobena rozsáhlému testování na velkém množství hlukových a řečových signálů. Při testování bylo využíváno hlukových a řečových databází vytvořených právě k tomuto účelu. Navržená technika je v předkládané práci hodnocena z hlediska míry potlačení různých druhů rušení a výsledné srozumitelnosti pro lidský sluch.

Vývoj nových technologií je umožněn stále větší dostupností velmi rychlých mikropočítačů a superrychlých DSP, které dovolují realizovat algoritmy separace řeči v reálném čase, tj. bez znatelného časového zpoždění. Část práce se proto zabývá paměťovými nároky a výpočetní náročností celého algoritmu. Zjištěné hodnoty uvedených veličin nám poskytují náhled na možnosti konkrétní hardwarové realizace navrženého a optimalizovaného algoritmu na moderních číslicových signálových procesorech.

4 METODA MAPOVÁNÍ SPEKTROGRAMU

4.1 ÚVOD

Náplní předkládané práce je návrh, vývoj a realizace zcela původního algoritmu extrakce řeči z rušného prostředí a dále jeho optimalizace pro práci v reálném čase za účelem pozdější implementace na vhodný signálový procesor. Základním požadavkem na vyvíjený algoritmus byla vyšší účinnost extrakce řeči, nežli jakou disponují v současné době používané techniky. Již v počátcích vývoje se ukázalo velmi výhodné využít obdobný přístup k problému separace jaký používá poměrně úspěšná technika RASTA [1], [2]. Její princip je založen na statistické analýze časově-kmitočtového prostoru degradovaného řečového signálu. Vlastní číslicové zpracování vypočteného dvojrozměrného prostoru (spektrogramu) je však v našem případě založeno na zcela odlišném přístupu, který se na základě množství provedených experimentů ukázal mnohem více účinný. Odlišným způsobem zpracování vstupního signálu bylo docíleno odstranění většiny nežádoucích vlastností techniky RASTA [1], [2]. Experimenty dále naznačují, že největší předností nově navržené metody je relativně malá citlivost vnitřních proměnných na charakter a úroveň nežádoucího rušení. Odpadá tedy složitá kalibrace algoritmu před vlastní separací.

4.2 ZÁKLADNÍ VLASTNOSTI NAVRŽENÉ TECHNIKY

Vyvinutý algoritmus, pracovně nazvaný „Mapovací metoda“, je jednokanálovou technikou [16], [17], [32], [33], [34], [35]. Jádro mapovacího algoritmu využívá ke své práci krátkodobých spekter hlukem degradovaného řečového signálu. Po výpočetně náročné statistické analýze dvojrozměrného časově-kmitočtového prostoru dochází ke zpracování modulu (popř. reálné a imaginární části) obecně komplexního spektrogramu. Výsledkem je binární mapa (maska), s jejíž pomocí dochází k efektivnímu oddělení užitečné řeči od okolního hluku. Základní myšlenkou celé metody, která zaručuje vysokou účinnost separace, je zachování oblastí řečové aktivity v časově-kmitočtovém prostoru a odstranění okolního rušení. Oblastmi řečové aktivity rozumíme plochu ve spektrogramu, ve které je energie užitečného signálu vyšší než energie pozadí.

Z psychoakustiky [4] je známo, že hluk pozadí je lidským sluchem převážně vnímán mimo oblasti řečové aktivity. V blízkosti dominantních spektrálních čar užitečného signálu je rušení lidským sluchem tzv. kmitočtově maskováno (pokud je energetická úroveň rušení nižší než energetická úroveň samotné řeči). Čím menší bude poměr výkonu signálu k výkonu šumu, tím užší bude kmitočtová oblast maskování a rušení začne být vnímáno výrazněji. Metoda Mapování spektrogramu se snaží nalézt vhodný práh, s jehož pomocí bude vytvořena binarizovaná šablona,

kterou bude následně vynásoben modul spektrogramu zarušené řeči. Součin s vytvořenou maskou způsobí odstranění těch částí spektrogramu, kde je úroveň rušivého pozadí nižší než úroveň užitečné řeči. Výstupem algoritmu je jednorozměrný signál obsahující rušení pouze uvnitř řečové aktivity, kde je ovšem kmitočtově maskováno užitečnou řečí a nepůsobí tudíž rušivě. Lze tedy říci, že metoda, podle velikosti poměru výkonu signálu k výkonu rušení, omezuje (maskuje) časově proměnné modulové spektrum vstupního signálu. Argument (fáze) spekter přitom zůstává nezměněn.

Nový pohled na potlačení parazitních hluků v užitečné řeči spočívá zejména v pojetí časově-kmitočtového prostoru jako dvojrozměrného obrazu, na který lze s úspěchem aplikovat rozsáhlý matematický aparát číslicového zpracování obrazů.

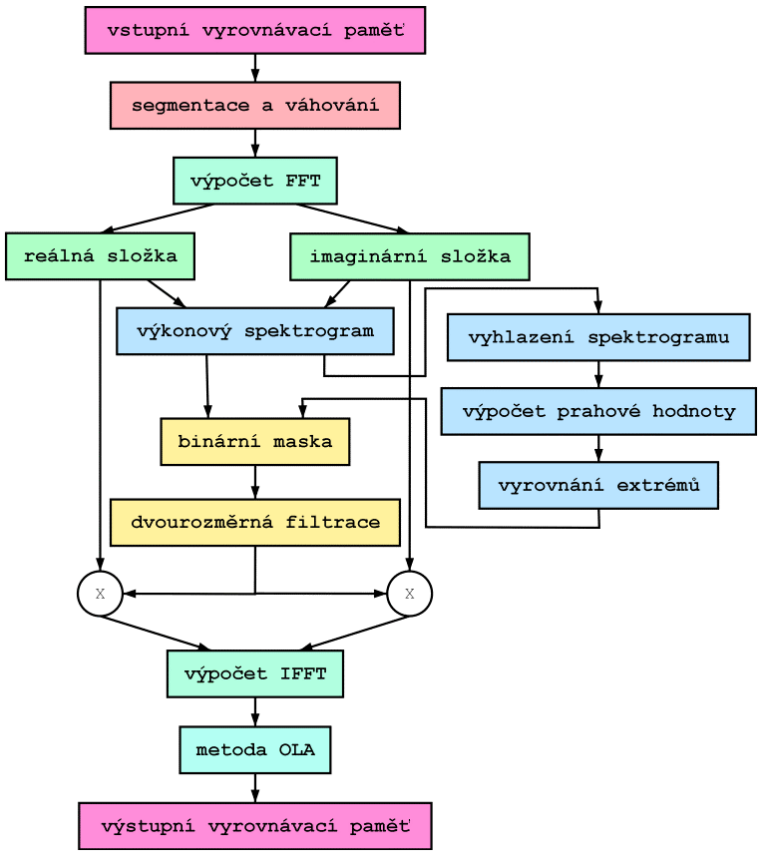
4.3 PŘEHLED ELEMENTÁRNÍCH OPERACÍ

Navržená separační technika neanalyzuje přímo časovou reprezentaci akustického signálu, jako je tomu u běžných technik adaptivní filtrace, ale využívá časově-kmitočtového prostoru určeného vhodnou integrální transformací. Vypočtené hodnoty vzorků jsou následně analyzovány a vhodně pozměněny. Aby bylo možné separovanou řeč zpětně převést do akustické podoby, je nutné dvojrozměrnou reprezentaci signálu inverzně transformovat do jednorozměrného časového průběhu.

Blokové schéma mapovacího algoritmu je znázorněno na vývojovém diagramu obr. 1 a lze jej popsat následujícími operacemi:

1. Načtení N vzorků vstupního akustického signálu $x(i)$, vzorkovaného frekvencí f_{vz} = 8 kHz a kvantovaného na q = 16 bitů, do vstupní vyrovnávací paměti.
2. Umístění posloupnosti vzorků do rámců (segmentů) vhodných délek a vhodných přesahů.
3. Předzpracování segmentů (odstranění stejnosměrné složky, váhování vhodnými okny apod.).
4. Výpočet komplexního spektrogramu z omezeného počtu po sobě následujících segmentů vstupního signálu $x(i)$. Rozdělení komplexního spektrogramu na modul a argument nebo na reálnou a imaginární část.
5. Úprava spektrogramu za účelem zjištění hodnoty prahu potřebného k oddělení řeči od okolního hluku (výpočet výkonového spektrogramu, vyhlazení časových trajektorií).
6. Stanovení základní hodnoty prahu.
7. Zpracování vypočtených hodnot (kompenzace extrémů).
8. Vytvoření binarizované šablony (masky, mapy).
9. Zpracování šablony 2D filtrací.
10. Součin masky a původního modulu spektrogramu, popř. reálné a imaginární části. Matice argumentů zůstává v každém případě nezměněna.

11. Zpětná Fourierova transformace komplexního spektrogramu složeného z nového modulu spektrogramu a původního argumentu, popř. z totožně upravené reálné a imaginární části spektrogramu.
12. Zpětný převod zpracované matice rámců do jednorozměrného akustického signálu pomocí metody OLA (OverLap Add) a uložení vzorků do výstupní vyrovnávací paměti.



Obr. 1: Blokové schéma algoritmu metody Mapování spektrogramu.

Uvedený výčet obsahuje řadu standardních kroků zpracování řečových signálů číslicovými metodami. Pro softwarovou realizaci uvedených bodů lze tedy použít běžných číslicových signálových procesorů určených pro zpracování akustických signálů v reálném čase. Detailní rozbor jednotlivých operací bude popsán v následujících kapitolách.

4.4 VÝPOČET ČASOVĚ-KMITOČTOVÉHO PROSTORU

4.4.1 Teoretický úvod

Předpokládejme diskrétní vstupní signál $x(i)$, který je součtem užitečné řeči $s(i)$ a hluku pozadí $n(i)$

$$x(i) = s(i) + n(i), \tag{1}$$

kde i je pořadí jednotlivých vzorků, tzv. časový index, který souvisí s reálným časem příchozího signálu dle vztahu

$$t = i \cdot T_{vz} = \frac{i}{f_{vz}}, \tag{2}$$

kde T_{vz} je vzorkovací perioda a f_{vz} vzorkovací kmitočet. Jak vyplývá ze vztahu (1) uvažujeme tzv. aditivní rušení [2], [8]. Dále předpokládejme, že $s(i)$ a $n(i)$ jsou na sobě statisticky nezávislé, musí tedy platit

$$\mathrm{E}\{x^2(i)\} = \mathrm{E}\{s^2(i) + n^2(i)\} = \mathrm{E}\{s^2(i)\} + \mathrm{E}\{n^2(i)\}, \tag{3}$$

kde operátor E značí střední hodnotu diskrétního signálu.

4.4.2 Předzpracování signálu v časové oblasti

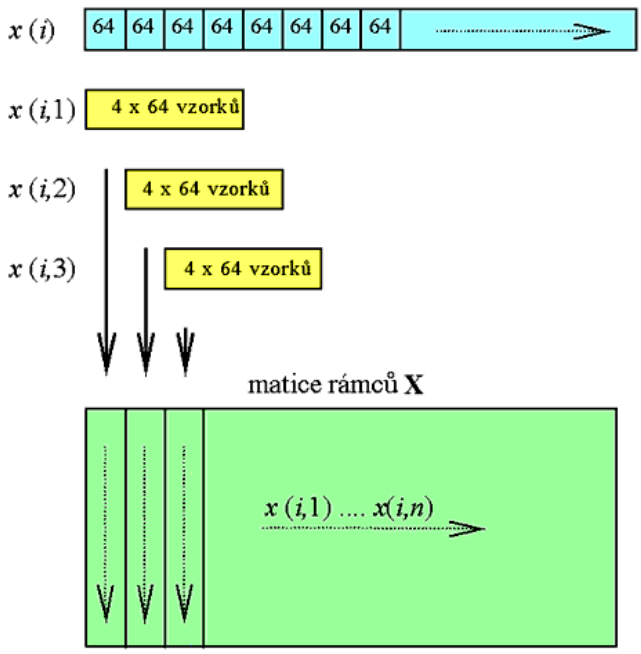
Nejprve předpokládejme šumem znehodnocený řečový signál $x(i)$ délky N navzorkovaný kmitočtem f_{vz} = 8 kHz a kvantovaný na q = 16 bitů. Standardní číslicové předzpracování diskrétního signálu $x(i)$ se skládá ze tří základních kroků:

- rozdělení signálu na krátké úseky (rámce, segmenty),
- odstranění stejnosměrné složky signálu,
- váhování segmentů vhodnými okny.

Rozdělení signálu na rámce (segmenty)

Souvislý číslicový signál je velmi účelné zpracovávat metodami tzv. krátkodobé analýzy, založenými na rozdělení příchozího signálu na krátké navzájem se překrývající úseky dle zvolených pravidel (viz obr. 2). Každý segment je vhodné rozdělit do čtyř stejně dlouhých části. Doba trvání každého segmentu byla zvolena na hodnotu t_l = 32 ms, překrytí na t_p = 24 ms a posun na t_s = 8 ms. Uvedeným časovým intervalům odpovídají počty vzorků dle trojice analogických vztahů

$$L = t_l \cdot f_{vz} = 0{,}032 \cdot 8000 = 256 \text{ vzorků}, \tag{4}$$
$$P = t_p \cdot f_{vz} = 0{,}024 \cdot 8000 = 192 \text{ vzorků}, \tag{5}$$
$$S = t_s \cdot f_{vz} = 0{,}008 \cdot 8000 = 64 \text{ vzorků}. \tag{6}$$



Obr. 2: Rozdělení vstupního signálu na rámce a jejich uspořádání do matice.

Hodnoty vznikly z požadavku dostatečné rozlišovací schopnosti v kmitočtové oblasti (toho dosáhneme velkou délkou rámce), ale zároveň také snahou o velké rozlišení v čase (toho naopak dosáhneme malou délkou rámce nebo také vzorkováním s vyšším vzorkovacím kmitočtem f_{vz}). Vzhledem ke skutečnosti, že řeč je během časového intervalu t_l = 32 ms téměř stacionární, lze volit větší délku okna a časové rozlišení zvýšit dodatečným překrytím sousedních rámců. Jednotlivé segmenty je vhodné ukládat do sloupců matice **X** (obr. 2), kde budou připraveny k dalšímu zpracování.

Odstranění stejnosměrné složky signálu

Parazitní vychýlení vstupního signálu způsobené analogovou části komunikačního řetězce lze potlačit následující číslicovou korekcí. Naprogramovaný algoritmus odečte od každého vstupního vzorku odhad střední hodnoty příslušného segmentu délky N

$$x'(i,n) = x(i,n) - \mu_x(n), \tag{7}$$

kde střední hodnotu $\mu_x(n)$ odhadneme dle vztahu

$$\mu_x(n) = \frac{1}{L}\sum_{i=0}^{N-1} x(i,n), \tag{8}$$

kde L značí délku segmentu ve vzorcích. Uvedená korekce bude spolehlivě pracovat pouze za předpokladu konstantní střední hodnoty $\mu_x(n)$ na dlouhém časovém intervalu (řádově sekundy). Odstranění stejnosměrné složky není nutné provádět, pokud vstupní signál $x(i)$ vykazuje zanedbatelně malou střední hodnotu μ_x.

Váhování segmentů vhodnými okny

Před dalším zpracování (výpočtem krátkodobých spekter) je výhodné jednotlivé segmenty vynásobit vhodnou váhovou posloupností, oknem $w(i)$, které pozmění hodnoty příslušných vzorků vstupního signálu takovým způsobem, aby následně vypočtená krátkodobá spektra co nejlépe odpovídala skutečnému spektru analyzovaného signálu. I vlastní segmentaci číslicového signálu, popsanou v předchozí části, je možné matematicky vyjádřit pomocí součinu vstupního signálu s různě časově posunutou obdélníkovou funkcí konečné délky. Avšak omezením signálu v časové rovině způsobíme nežádoucí zkreslení spektra segmentovaného signálu - rozmazání a rozptyl spektra.

Matematicky lze proces segmentace časově neomezeného signálu $x(i)$ s využitím vhodně časově posunuté váhové posloupnosti $w(i)$ vyjádřit vztahem

$$x(i,n) = x(i) \cdot w(i - nS), \tag{9}$$

kde i vyjadřuje běžící čas ve vzorcích, n pevný pozorovací čas v segmentech a konstanta S posun okna mezi sousedními segmenty. Jevy vzniklé ve spektru signálu $x(i,n)$ lze odvodit z principu konvoluce spektra $X(k)$ časově neomezeného signálu $x(i)$ se spektrální funkcí $W(k)$ okna $w(i)$

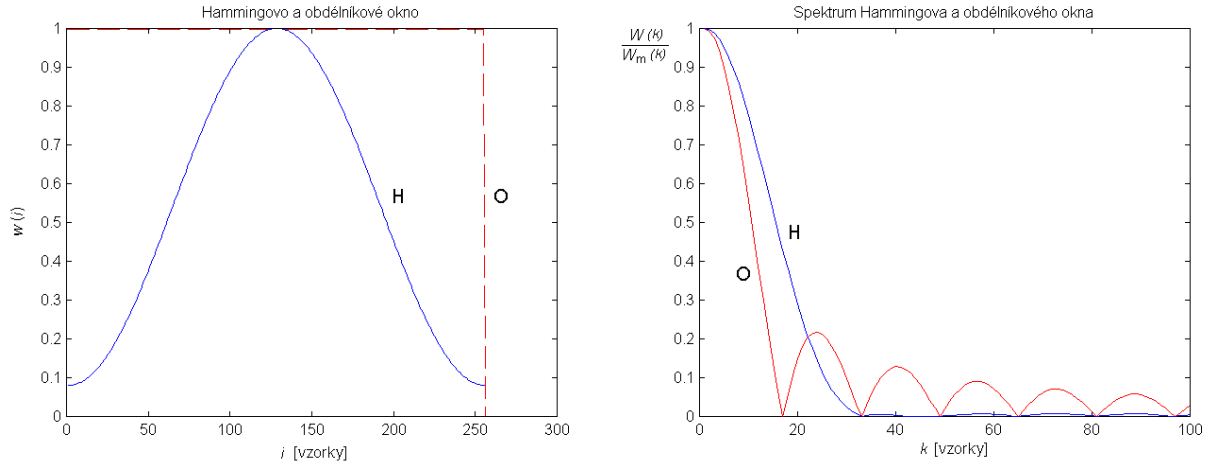
$$X(k,n) = X(k) * W(k), \tag{10}$$

kde k značí pořadí vzorků ve spektru.

Nežádoucí vlastnosti pravoúhlé segmentace lze částečně eliminovat nahrazením obecného obdélníkového okna jinými váhovými posloupnostmi, které vykazují odlišné vlastnosti nejen v čase, ale zejména ve spektru. V technické praxi číslicového zpracování řečových signálů se obecné obdélníkové okno nejčastěji nahrazuje Hammingovým oknem, které v časové oblasti částečně utlumí signál na okrajích rámce, což v kmitočtové rovině přinese požadované snížení úrovně postranních laloků modulového spektra (obr. 3). Současně dochází ke snížení strmosti přechodu spektrální funkce z propustného do nepropustného pásma, to naopak vede k mírnému rozostření vypočteného modulového spektra. Na obr. 3 je vykresleno detailní porovnání Hammingova okna (modrá barva) a běžného obdélníkového okna (červená barva) v časové i kmitočtové rovině. Hammingova posloupnost je definována vztahem [8]

$$w(i) = 0{,}54 - 0{,}46 \cdot \cos(\frac{2\pi}{L} \cdot i)\,,\ \ i=0,\ 1,\ldots,\ L\text{-}1. \qquad (11)$$

Nyní je vstupní signál dostatečně předzpracován a může být přistoupeno k výpočtu časově-kmitočtového prostoru, který je jádrem navržené separační techniky.



Obr. 3: Porovnání časových průběhů a amplitudových spekter Hammingova okna (modrá barva) a obdélníkového okna (červená barva) .

Výpočet spektrogramu

Spektrogram je soubor spekter krátkých úseků signálu. Bývá často názorně vykreslován jako dvojrozměrný obraz, v němž svislá souřadnice reprezentuje kmitočet, vodorovná čas a barevná stupnice modul nebo argument komplexní spektrální funkce. Uspořádání spekter do matice je velmi výhodné i z toho důvodu, že na vzniklý spektrogram je možné aplikovat rozsáhlý matematický aparát číslicového zpracování obrazů [6], [7], [8], a právě tento úhel pohledu s úspěchem využívá navržená metoda Mapování spektrogramu.

Základem pro výpočet komplexních spekter jednotlivých segmentů je diskrétní Fourierova transformace DFT (Discrete Fourier Transform) [8], [13]. Realizace algoritmu DFT na krátkém časovém intervalu (např. během jednoho segmentu) bývá nazývána krátkodobou diskrétní Fourierovou transformací STFT (Short Time Fourier Transform), kterou lze definovat vztahem

$$\boldsymbol{X}(n,k) = \sum_{i=0}^{L-1} x(n,i) \cdot e^{-\mathrm{j}\frac{2\pi}{L}ki}\,, \qquad (12)$$

kde i je index vzorků vstupního signálu v časové rovině, k je pořadí vzorků v kmitočtové oblasti a n vyjadřuje segmenty délky L jdoucí v čase za sebou. Vlastní

výpočet se v praxi výhradně realizuje rychlými algoritmy FFT (Fast Fourier Transform), popis této techniky lze nalézt např. v [18].

Kompromisem mezi separační schopností navrženého algoritmu a jeho výpočetní náročností je použití N_{FFT}-bodové Fourierovy transformace, kde $N_{\mathrm{FFT}} = 256$ vzorků. Šířku pásma mezi sousedními vzorky spektra lze popsat a následně vyčíslit úměrou

$$\Delta f_{\mathrm{FFT}} = \frac{f_{\mathrm{vz}}}{N_{\mathrm{FFT}}} = \frac{8000}{256} = 31{,}25\ \mathrm{Hz/vzorek}, \tag{13}$$

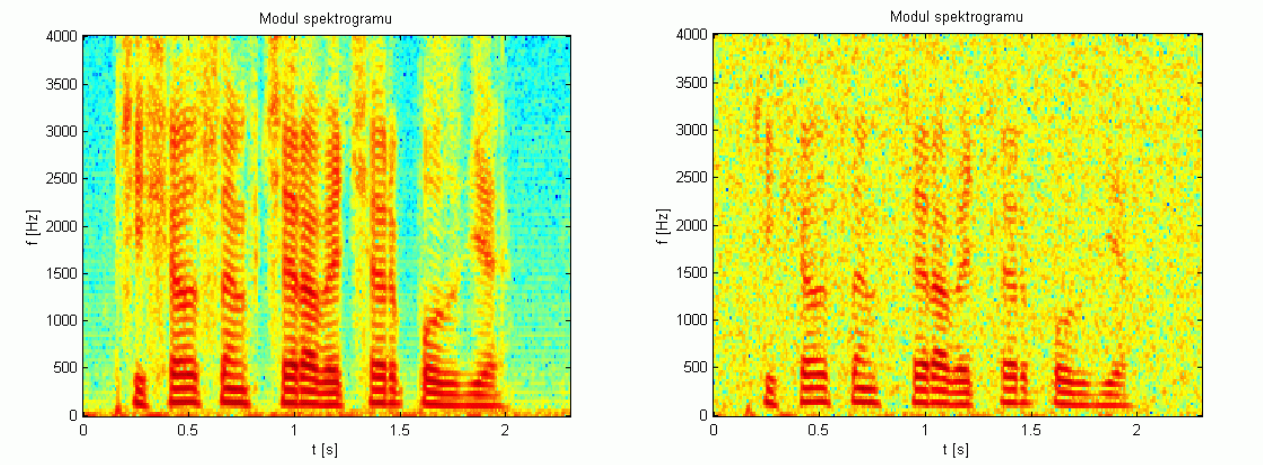
kde vzorkovací kmitočet $f_{vz} = 8$ kHz a počet bodů transformace $N_{\mathrm{FFT}} = L = 256$.

Z vypočteného souboru komplexních spekter dle vztahu (12) snadno určíme potřebnou modulovou a argumentovou část, kterou je možné dále vyjádřit pomocí reálné a imaginární složky jako

$$X(n,k) = \left|\boldsymbol{X}(n,k)\right| = \sqrt{\mathrm{Re}\{\boldsymbol{X}(n,k)\}^2 + \mathrm{Im}\{\boldsymbol{X}(n,k)\}^2}, \tag{14}$$

$$\varphi(n,k) = \arg[\boldsymbol{X}(n,k)] = \mathrm{arctg}\,\frac{\mathrm{Im}\{\boldsymbol{X}(n,k)\}}{\mathrm{Re}\{\boldsymbol{X}(n,k)\}}. \tag{15}$$

Matice sloupcových vektorů vyjádřená vztahem (14) bývá nazývána modulem spektrogramu a matice sloupcových vektorů podle vztahu (15) argumentem či fází spektrogramu.



Obr. 4: Modul spektrogramu věty *„Přišel jsem včera večer pozdě“* bez rušení (vlevo) a s aditivním Gaussovým šumem o směrodatné odchylce $\sigma = 0{,}05$ (vpravo).

V levé části obr. 4 je vykreslen modul spektrogramu nezkresleného řečového signálu *„Přišel jsem včera večer pozdě“* namluveného mužem. Pravá část obr. 4 znázorňuje modul spektrogramu stejné promluvy, ovšem znehodnocené aditivním

rušením v podobě Gaussova šumu o směrodatné odchylce $\sigma = 0{,}05$ (platí pro vzorky vstupního signálu normované do intervalu <-1;+1>). Hodnota poměru výkonu signálu k výkonu šumu je rovna *SNR* = 7,6 dB. Svislá osa u obou spektrogramů popisuje kmitočet, vodorovná osa čas a barevná stupnice odpovídá velikosti jednotlivých koeficientů modulu a fáze spektrogramu dle schématu  (nejnižší hodnota -> nejvyšší hodnota).

4.5 STANOVENÍ ROZHODOVACÍHO PRAHU METODOU SPEKTRÁLNÍHO ODEČÍTÁNÍ

Klíčovým bodem mapovací techniky je dostatečně přesný výpočet prahové úrovně, s jejíž pomocí dochází k vytvoření binarizované dvojrozměrné šablony odlišující oblasti řečové aktivity od okolního rušení. Z důvodu silné závislosti prahu $T_{\text{práh}}$ na okamžitém poměru výkonu signálu k výkonu rušení *SNR* bylo zapotřebí nalézt vhodný algoritmus, který by byl schopný příchozí řečový signál v daném okamžiku analyzovat a parametr *SNR* v jednotlivých časových trajektoriích vyčíslit. Za nejvhodnější postup byl vybrán algoritmus spektrálního odečítání podle Rainera Martina [5]. Mezi jeho největší přednosti patří skutečnost, že ke své činnosti nepotřebuje detektor řečové aktivity.

4.5.1 Implementace odečítacího algoritmu

Princip metody spektrálního odečítání spočívá v odhadu výkonového spektra stacionárního a ergodického rušení [8] z obecných $N_{\text{práh}}$ výkonových spekter vstupní nahrávky. K odhadu se využívá minimálních hodnot souboru výkonových spekter na časově omezeném okně. Hlavní výhodou uvedeného postupu je možnost zjištění odhadu výkonu rušení přímo z degradovaného řečového signálu. Odpadá tedy nutnost detekce řečové aktivity a s ní související zjišťování parametrů nežádoucích signálů pouze v řečových pauzách.

Základem algoritmu je výpočet výkonového spektrogramu

$$S_x(n,k) = \frac{1}{N_{\text{FFT}}} \left| \boldsymbol{X}(n,k) \right|^2 = \frac{1}{N_{\text{FFT}}} \left(\text{Re}\{X(n,k)\}^2 + \text{Im}\{X(n,k)\}^2 \right), \quad (16)$$

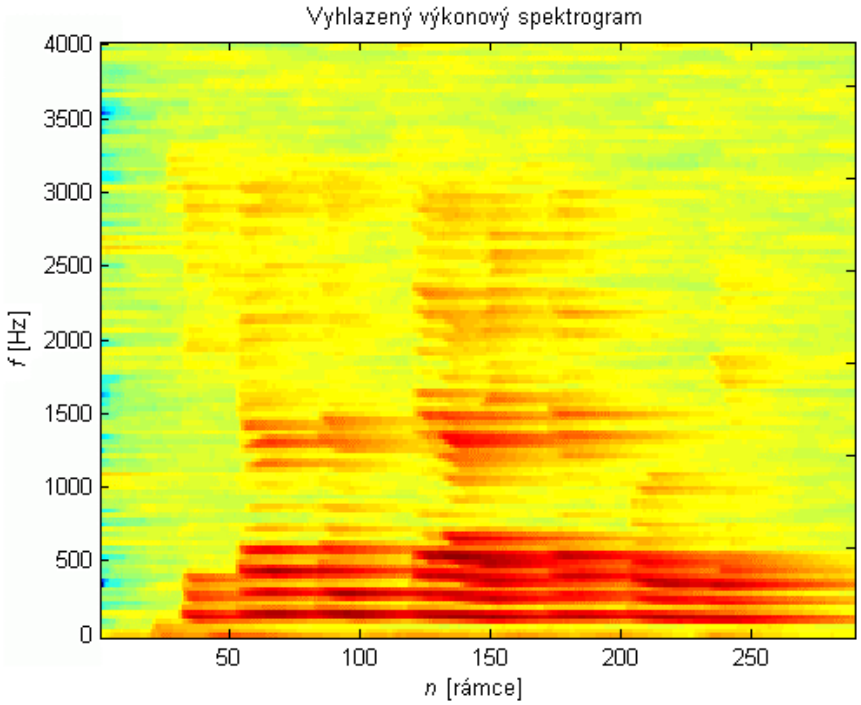
kde k je pořadí vzorků ve spektru, n je index zpracovávaných segmentů a N_{FFT} odpovídá počtu bodů spektrální funkce.

Vzhledem ke značnému rozptylu hodnot sousedních vzorků v jednotlivých časových trajektoriích, způsobených hlukem pozadí, je nutné provést vyhlazení průběhů přes $N_{\text{práh}}$ segmentů. Použijeme jednoduchý číslicový IIR filtr 1. řádu typu dolní propusti popsaný diferenční rovnicí

$$P_x(n,k) = \alpha \cdot P_x(n-1,k) + (1-\alpha) \cdot S_x(n,k), \qquad (17)$$

kde $\alpha \in$ <0,1> je tzv. vyhlazovací konstanta [5]. Její hodnota závisí jak na nestacionárním charakteru řečového signálu, tak i na časových změnách energie parazitního hluku a obvykle se pohybuje v intervalu <0,90 ; 0,95>. Experimentálně bylo zjištěno, že pro separaci řečového signálu českého mluvčího znehodnoceného stacionárním širokopásmovým šumem zcela vyhoví hodnota $\alpha = 0{,}95$.

Obr.5 znázorňuje vyhlazené časové trajektorie výkonového spektrogramu z obr. 4 (slovní spojení *„Přišel jsem včera večer pozdě“* s aditivním Gaussovým šumem o směrodatné odchylce $\sigma = 0{,}05$). Zde je zapotřebí zdůraznit, že vyhlazený výkonový spektrogram slouží pouze pro výpočet rozhodovacího prahu $T_{\text{práh}}$ a není tedy určen pro další zpracování řečového signálu, proto si můžeme dovolit jeho degradaci.



Obr. 5: Vyhlazené časové trajektorie výkonového spektrogramu věty *„Přišel jsem včera večer pozdě“* s aditivním Gaussovým šumem o směrodatné odchylce $\sigma = 0{,}05$, konstanta $\alpha = 0{,}95$.

Odhad výkonového spektra rušení $P_n(k)$ lze podle [5] vyjádřit jako vážené minimum vyhlazených krátkodobých výkonových spekter $P_x(n,k)$ dle vztahu

$$P_n(k) = o_{\min} \cdot P_{\min}(k), \qquad (18)$$

kde velikost součinitele $o_{\min}$ závisí na charakteru hlukového signálu a tedy i na statistických parametrech vektoru $P_{\min}(k)$. Vektor $P_{\min}(k)$ je možné vyčíslit z rovnice

$$P_{\min}(k) = \min_{N_{práh}} P_x(n,k), \tag{19}$$

kde jednotlivé prvky sloupcového vektoru $P_{\min}(k)$ jsou minimální hodnoty časových trajektorií z $N_{\text{práh}}$ vyhlazených výkonových spekter $P_x(n,k)$.

4.5.2 Parametrický popis parazitního signálu

Způsob výpočtu statistických parametrů nežádoucího rušení také částečně vychází z algoritmu spektrálního odečítání Rainera Martina [5]. Využijeme odhad nevychýleného výkonového spektra šumu $P_{\min}(k)$. Pro účely separační techniky Mapování spektrogramu je výhodnější přejít od výkonového popisu šumu k jeho modulu, tedy přesněji k odhadu modulového spektra jedné realizace náhodného procesu. Inverzně ke vztahu (16) bude platit

$$\left|\boldsymbol{X}_n(k)\right| = \sqrt{N_{\text{FFT}} \cdot P_{\min}(k)}\,, \tag{20}$$

kde $\left|\boldsymbol{X}_n(k)\right|$, nebo jen $X_n(k)$, značí výše zmíněný odhad modulové spektrální funkce jedné realizace náhodného procesu.

Nyní je zapotřebí nalézt vhodný postup, na základě kterého bude ze získaného odhadu modulového spektra šumu $X_n(k)$ určena prahová hodnota $T_{\text{práh}}$, s jejíž pomocí nastane oddělení řečové aktivity od hlučného pozadí. Ukázalo se velmi výhodné využít základních statistických veličin, kterými jsou střední hodnota $\mathrm{E}\{.\}$ a směrodatná odchylka $\sigma\{.\}$, resp. jejich odhady

$$\mathrm{E}\{X_n\} \doteq \mu = \frac{1}{K} \cdot \sum_{k=0}^{K} X_n(k) \tag{21}$$

$$\sigma\{X_n\} \doteq \sigma = \sqrt{\frac{\sum_{k=0}^{K} (X_n(k) - \mu)^2}{K}}\,, \tag{22}$$

kde $K = \frac{N_{\text{FFT}}}{2}$ (spektrum diskrétního signálu je periodické a symetrické kolem poloviny vzorkovacího kmitočtu).

Pokud má rušící signál charakter širokopásmového šumu, například s Gaussovým rozložením pravděpodobnosti, bude určení prahu velmi jednoduché a matematicky snadno zdůvodnitelné. Využijeme poznatku o téměř 100% pravděpodobnosti

skutečnosti, že okamžitá hodnota šumu $n(i)$ nebude nabývat vyšších hodnot než známé pravidlo tří-sigma

$$|n(i) - \mu| \leq 3\sigma . \tag{23}$$

Praxe podložená množstvím testovacích nahrávek ukázala, že změna hodnoty číselné konstanty před směrodatnou odchylkou o jednotku libovolným směrem má jen velmi malý vliv na výslednou filtraci. Pro naprostou většinu širokopásmových šumů, včetně šumu bílého, je vhodné volit konstantu v těsném okolí čísla 2 (pro úzkopásmové rušení je rozptyl prahů poněkud vyšší). Vyšší hodnoty číselné konstanty mají za následek až příliš dokonalé ohraničení jednotlivých harmonických složek řečového signálu, což v důsledku malého rozlišení spektra (v našem případě Δf_{FFT} = 31,25 Hz/vzorek) často vede k degradaci samotného řečového signálu. Hodnotu prahu lze tedy obecně vyjádřit vztahem

$$T_{práh} = \mu + \tau\sigma , \tag{24}$$

kde konstanta τ vyjadřuje výše uvedenou závislost na charakteru rušícího signálu.

4.6 VYTVOŘENÍ BINÁRNÍ MASKY Z MODULU SPEKTROGRAMU A JEJÍ ZPRACOVÁNÍ

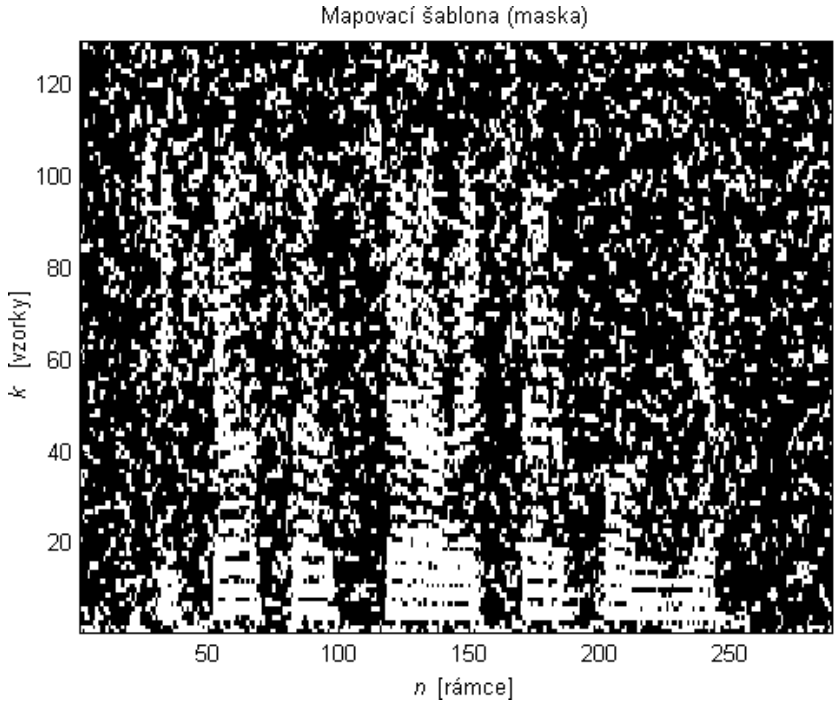
4.6.1 Binarizace

Nyní se dostáváme ke klíčovému bodu celého separačního algoritmu, a to k vytvoření binární masky (mapy, šablony). Binarizací neboli prahováním rozumíme transformaci jasových úrovní dvojrozměrného signálu (obrazu) do dvojúrovňové stupnice černá – bílá. Hlavní myšlenkou mapovací techniky je možnost pohlížet na spektrogram jako na dvojrozměrný obraz s využitím rozsáhlého matematického aparátu číslicového zpracování obrazů. Nyní již tedy zbývá pomocí hodnoty prahu $T_{\mathrm{práh}}$, vypočteného v předchozí kapitole z $N_{\mathrm{práh}}$ segmentů, provést binarizaci modulu spektrogramu $X(n,k)$. Vznikne matice $\mathbf{F}_{\mathrm{mask}}$, jejíž prvky nabývají diskrétních hodnot 0 nebo 1 podle následujícího pravidla

$$F_{\mathrm{mask}}(n,k) = \begin{cases} 1 & \text{pro } X(n,k) > T_{\mathrm{práh}} , \\ 0 & \text{pro } X(n,k) \leq T_{\mathrm{práh}} . \end{cases} \tag{25}$$

Na obr. 6 je znázorněna binární maska Gaussovým šumem znehodnoceného testovacího signálu *„Přišel jsem včera večer pozdě“* (σ = 0,05; *SNR* = 7,6 dB). Jsou zde zřejmé četné osamocené shluky saturovaných bodů o úrovni jedna (bílá), které

jsou způsobeny nenulovou pravděpodobností výskytu okamžitých hodnot šumu nad prahem určeným dle vztahu (24) při $\tau = 3$. V teorii zpracování obrazů se zbytkový šum tohoto typu příznačně nazývá „pepř a sůl“ (v našem případě hlavně „sůl“).



Obr. 6: Binární maska vypočtená prahováním modulu spektrogramu.

4.6.2 Filtrace dvojrozměrného prostoru

Kdybychom v tomto okamžiku násobili maskou $\mathbf{F}_{\text{mask}}$ z obr. 6 původní modul spektrogramu $X(n,k)$ z obr. 4, bude výsledná nahrávka obsahovat výrazný „hrubozrnný šum“. Zbytkový syntetický hluk má v časové i kmitočtové rovině charakter ostře ohraničených impulzů. Série experimentů ukázala, že zbytkové rušení uvedených vlastností je lidským sluchem vnímáno velmi nepříjemně. Je tedy nutné všechny ostré přechody v binární masce odstranit následným zpracováním pomocí poměrně výpočetně náročných operací číslicového zpracování obrazů.

Mediánová dvojrozměrná filtrace

K účinnému potlačení impulzního rušení „pepř a sůl“ slouží mediánový filtr [6], [8], [18]. Jeho princip je založen na výběru prostřední hodnoty setříděné posloupnosti libovolné množiny bodů. Množina nebo též okno je definována tvarem a velikostí. Vzhled okna se většinou volí experimentálně s ohledem na charakter impulzního rušení. Pro náš účel postačí okno čtvercové o rozměru 5x5 bodů. Pokud

je počet prvků v množině lichý, je střed setříděné posloupnosti jednoznačný. Pokud je počet prvků sudý, je výstupní vzorek získán aritmetickým průměrem prostředních vzorků v setříděné posloupnosti.

Mediánový filtr je nelineárním a většinou také nerekurzivním systémem. Jeho vlastností je schopnost bezezbytku odstranit impulzní rušení a přitom nedeformovat strmosti hran vstupního signálu, jako je tomu u lineárních (průměrujících) filtrů. Matematický zápis mediánového filtru je následující

$$y_{i,j} = \mathrm{med}\{x_{i-1,j-1}; x_{i,j}; x_{i+1,j}; x_{i,j-1} \ldots .\}, \tag{26}$$

kde proměnné i a j představují souřadnice středu okna ve zpracovávaném obrazu x a výstupním obrazu y. Mediánový filtr je výpočetně poměrně náročným algoritmem pro velké množství srovnávacích operací. Z hlediska rychlosti algoritmu je účelné volit okno co možná nejmenší.

Proces dvojrozměrné mediánové filtrace aplikované na binární matici $\mathbf{F}_{\mathrm{mask}}$ lze napsat jako

$$\mathbf{F}_{\mathrm{2Dmed}} = \underset{5x5}{\mathrm{med}}\{\mathbf{F}_{\mathrm{mask}}\}. \tag{27}$$

Na obr. 8a je pomocí dvojrozměrného mediánového filtru s oknem o rozměru 5x5 bodů dle vztahu (27) zpracovaná binární matice $\mathbf{F}_{\mathrm{mask}}$ z obr. 6. Je zřejmé, že naprostá většina rušivých impulzů a jejich shluků byla odstraněna. Zbylé rozsáhlejší oblasti saturovaných bodů již filtr s oknem 5x5 nedokázal odstranit. Nabízí se tedy možnost zvětšit rozměr okna, ale poté by již s velkou pravděpodobností docházelo k odstranění důležitých pasáží řeči, převážně na vyšších kmitočtech.

Nepříjemnou vlastností zpracované matice $\mathbf{F}_{\mathrm{2Dmed}}$ je přítomnost ostře ohraničených oblastí v časové i kmitočtové rovině. Kdybychom nyní násobili maticí $\mathbf{F}_{\mathrm{2Dmed}}$ modul spektrogramu $X(n,k)$ z obr. 4, bude výsledná nahrávka i nyní znít synteticky a navíc bude obsahovat v místech menších osamocených oblastí umělý úzkopásmový hluk, který lze přirovnat ke zvuku „znějících zvonů“. Syntetické nežádoucí hluky odstraníme dvojrozměrnou lineární filtrací.

Dvojrozměrná lineární filtrace

K dostatečně velkému potlačení uvedených nežádoucích vlastností je vhodné použít lineární dvojrozměrný FIR (Finite Impulse Response) filtr [6], [8] s konečnou impulzní odezvou aplikovaný na matici $\mathbf{F}_{\mathrm{2Dmed}}$. Výběrem dvojrozměrného filtru vhodných vlastností docílíme potlačení ostrých přechodů jak v kmitočtu, tak i v čase. Energie rušivých shluků saturovaných bodů bude v matici $\mathbf{F}_{\mathrm{2Dmed}}$ rozostřena do větší plochy a nebude již působit rušivě.

Dvojrozměrnou filtraci lze v předmětové rovině vyjádřit diskrétní dvojrozměrnou konvolucí vstupního obrazu **x** o rozměru [x_x ; x_y] s dvojrozměrnou impulzní odezvou **h** o rozměru [h_x ; h_y] jako

$$y(i,j) = \sum_{k=0}^{k=h_x-1} \sum_{l=0}^{l=h_y-1} h(k,l) \cdot x(i-k, j-l)\,, \tag{28}$$

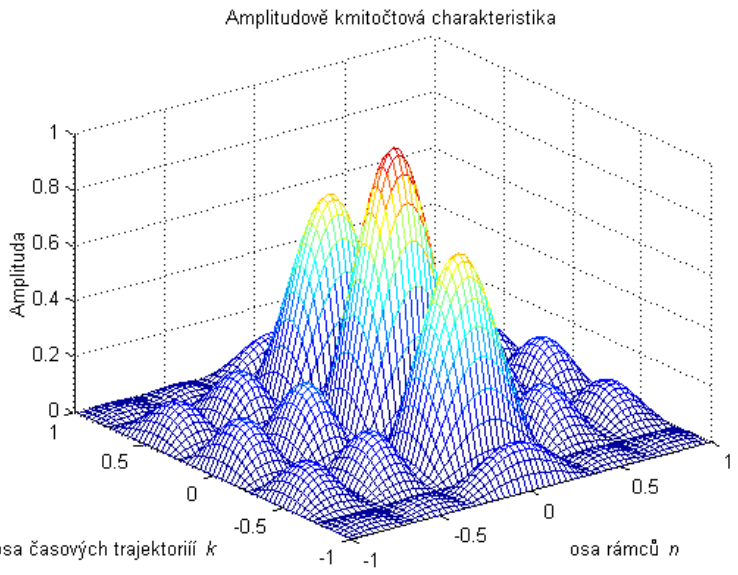
kde proměnné k a l jsou indexy impulzní odezvy **h** a proměnné i a j indexy vstupního a výstupního obrazu **x** a **y**. Impulzní odezvu **h** je v teorii zpracování obrazů zvykem nazývat bodovou rozptylovou funkcí PSF (Point Spread Function), dvojrozměrným operátorem či korelačním jádrem.

Z předchozího vztahu je zřejmé, že pokud bude vstupní obraz **x** konečných rozměrů x_x x x_y a velikost rozptylové funkce **h** bude h_x x h_y, potom bude rozměr výstupního obrazu y_x x y_y a bude platit

$$y_x = x_x + h_x - 1 \quad \text{a} \quad y_y = x_y + h_y - 1\,. \tag{29}$$

Pro realizaci dvojrozměrné FIR filtrace aplikované na matici $\mathbf{F}_{2Dmed}$ je žádoucí navrhnout takový filtr, který by respektoval charakter řečového signálu. Použijeme filtr typu dolní propusti s nižším prostorovým horním mezním kmitočtem v rovině segmentů nežli v rovině kmitočtů. Výsledkem bude vyšší rozostření matice $\mathbf{F}_{2Dmed}$ podél jednotlivých časových trajektorií, čímž docílíme respektování základního tónu řeči a jeho násobků. Experimentálně byl za nejvhodnější filtr vybrán systém typu dolní propusti charakterizovaný impulzní odezvou z obr. 7. Amplitudově kmitočtová charakteristika navrženého 2D filtru je také vykreslena na obr. 7.

$$\mathbf{h}_1 = \frac{1}{24}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$



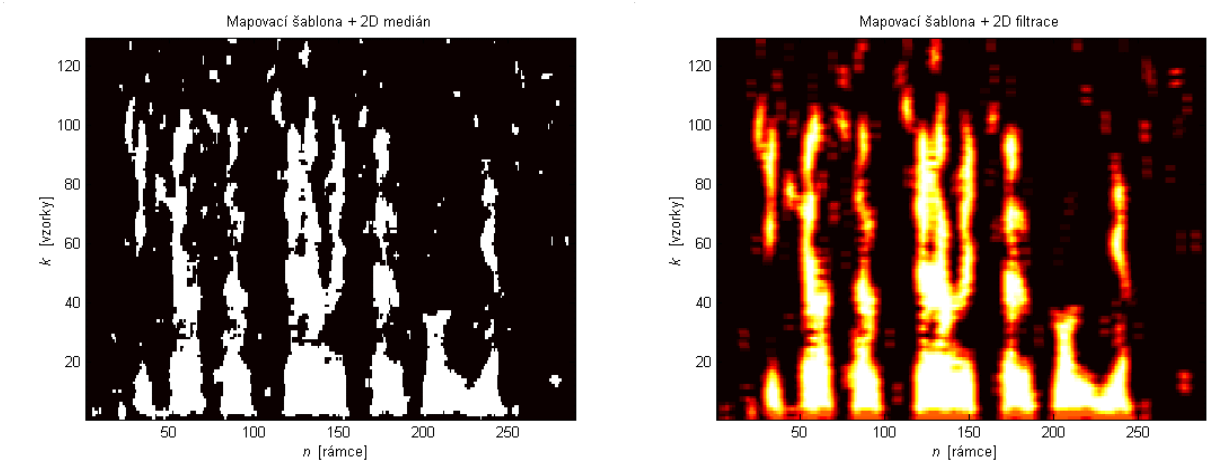
Obr. 7: Impulzní odezva a amplitudově kmitočtová charakteristika dvojrozměrné dolní propusti.

Z obr. 7 je skutečně zřejmé respektování časového průběhu základního tónu řeči a jeho násobků. Prostorové kmitočty budou v obraze $\mathbf{F}_{2\mathrm{Dmed}}$ potlačeny výrazněji v ose segmentů (v ose času) než v ose časových trajektorií (ose kmitočtů).

Proces dvojrozměrné FIR filtrace aplikované na binární matici $\mathbf{F}_{2\mathrm{Dmed}}$ lze psát ve výsledku jako

$$\mathbf{F}_{2\mathrm{Dfilt}} = \mathbf{F}_{2\mathrm{Dmed}} * \mathbf{h} \,. \tag{30}$$

Na obr. 8b je vykreslen prototyp víceúrovňové masky $\mathbf{F}_{2\mathrm{Dfilt}}$ zbavené šumu „pepř a sůl“ a následně zpracované lineárním 2D FIR filtrem typu dolní propust s impulzní odezvou $\mathbf{h}_1$ z obr. 7. Praktické zkoušky ukázaly výrazné potlačení ostrých přechodů v binární matici a rozptýlení rušivých oblastí do větší plochy. Snížení syntetického hluku nastalo při rušení širokopásmovým Gaussovým šumem asi 4x, ovšem subjektivní hodnocení lidským sluchem bylo několikanásobně vyšší.



Obr. 8: Zpracování binární masky 2D filtrací, (a) binární maska zpracovaná 2D mediánovým filtrem se čtvercovým oknem o rozměru 5x5 bodů, (b) víceúrovňová maska upravená dvojrozměrnou lineární FIR filtrací.

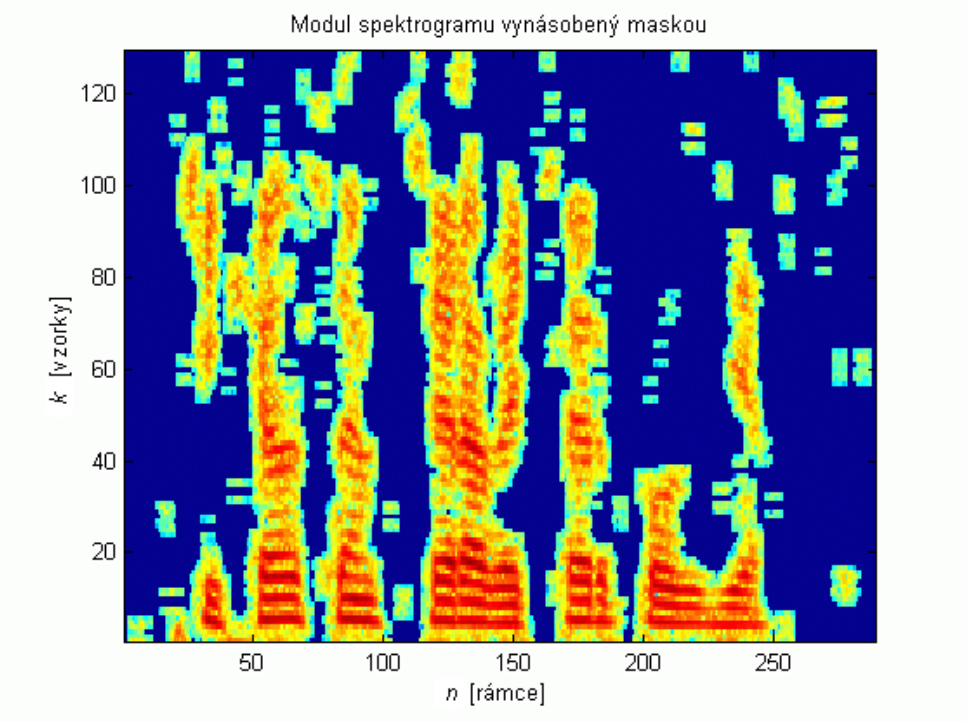
4.7 PROCES SEPARACE PARAZITNÍCH HLUKŮ

Víceúrovňová mapovací šablona, vypočtená podle předchozích kapitol, obsahuje přesnou informaci o výskytu užitečné řeči v časově-kmitočtovém prostoru. Vlastní proces separace řečové aktivity je výhodné realizovat skalárním násobením (bod po bodu) vypočtené a zpracované matice $\mathbf{F}_{2\mathrm{Dfilt}}$ s původním modulem spektrogramu $X(n,k)$. Součin jednotlivých prvků uvedených matic lze vyjádřit následujícím vztahem

$$Y(n,k) = F_{2\mathrm{Dfilt}}(n,k) \cdot X(n,k) \,. \tag{31}$$

Výsledek operace součinu je znázorněn na obr. 9. Zpracovaný spektrogram nyní obsahuje rušení pouze uvnitř oblastí řečové aktivity, okolní hluk pozadí je zcela odstraněn. Je zřejmé, že metoda Mapování spektrogramu je schopná, na rozdíl od ostatních metod, při správném nastavení parametrů jednotlivých operací, parazitní hluk zcela odstranit, nikoliv jen potlačit.

Nejvýraznějším nežádoucím jevem, který snižuje kladné separační vlastnosti předkládané metody, je možné zhoršení srozumitelnosti výstupní řeči způsobené nepřesným odhadem výkonového spektra rušení. Chybný výpočet způsobí nedokonalé ohraničení užitečné řeči ve spektrogramu. Nepřesná identifikace charakteru hlukového signálu je nejvýraznější při silně nestacionárním rušení přítomném ve všech časových trajektoriích.



Obr. 9: Modul spektrogramu vstupního signálu vynásobený víceúrovňovou maskou F_{2Dfilt}.

4.8 ZPĚTNÁ TRANSFORMACE DVOJROZMĚRNÉHO ČASOVĚ-KMITOČTOVÉHO PROSTORU

Po zpracování modulu spektrogramu hlukem degradovaného řečového signálu je nutné provést návrat z časově-kmitočtové reprezentace jednotlivých vzorků zpět do jednorozměrného časového průběhu. Nejprve sestavíme výstupní komplexní spektrogram $\boldsymbol{Y}(n,k)$ ze zpracovaného modulu spektrogramu $Y(n,k)$ a původního argumentového spektrogramu $\varphi(n,k)$ jako

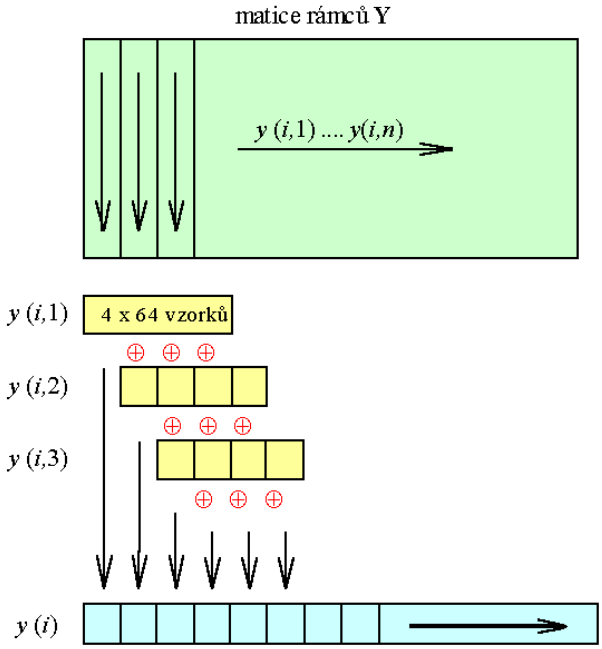
$$\boldsymbol{Y}(n,k) = Y(n,k) \cdot e^{j \cdot \varphi(n,k)}, \tag{32}$$

kde n je opět pořadí segmentů v čase a k je pořadí vzorků ve spektru.

Základem zpětného převodu spektrální funkce do časové oblasti je zpětná diskrétní Fourierova transformace, často označovaná zkratkou IDFT (Inverse Discrete Fourier Transformation) [8], [13]. Algoritmus IDFT aplikovaný na časově omezené úseky signálu (segmenty) bývá nazýván zpětnou krátkodobou diskrétní Fourierovou transformací ISTFT (Inverse Short Time Fourier Transform) a lze jej definovat inverzně ke vztahu (12) jako

$$y(n,i) = \frac{1}{L}\sum_{k=0}^{L-1} \boldsymbol{Y}(n,k)\cdot \mathrm{e}^{\mathrm{j}\frac{2\pi}{L}ki}, \tag{33}$$

kde i značí index vzorků výstupního signálu v časové rovině, k pořadí vzorků v kmitočtové oblasti a n jednotlivé segmenty délky L jdoucí v čase za sebou.

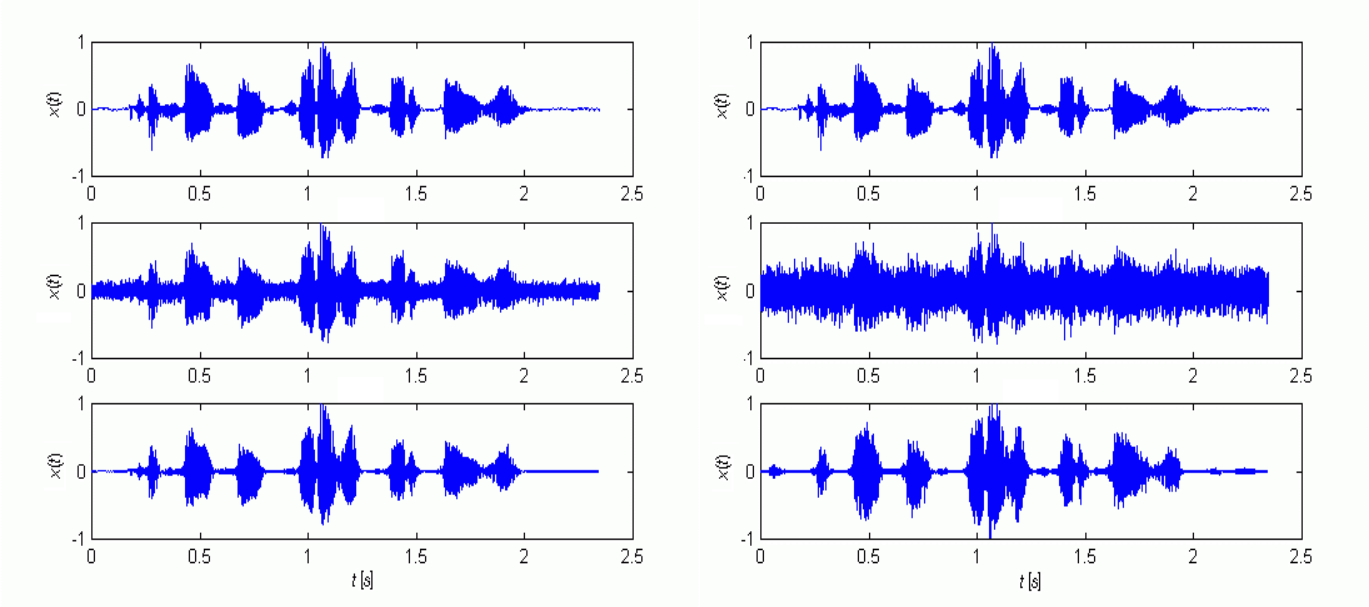


Obr. 10: Zpětná rekonstrukce jednorozměrného řečového signálu pomocí metody OLA.

Zkreslení způsobené segmentací signálu v časové rovině je účelné eliminovat poslední operací metody Mapování spektrogramu, kterou je zpětné rozložení matice segmentů $y(n,k)$ na jednorozměrný signál $y(i)$. Využívá se metoda OLA (OverLap Add – přičtení přesahu) [1], [2] založená na součtu přesahů sousedních segmentů. Výstupní segmenty jsou uspořádány vodorovně pod sebe s respektováním vzájemných časových posunů. Výstupní signál $y(i)$ je pak získán součtem vzorků ve svislém směru. Princip techniky OLA je schématicky znázorněn na obr. 10.

4.9 ZHODNOCENÍ KVALITY VÝSTUPNÍHO SIGNÁLU

Uvažujme dva vstupní řečové signály obsahující slovní spojení *„Přišel jsem včera večer pozdě“* namluvené mužem. Nahrávky byly pořízené v akusticky odstíněné komoře a neobsahují tedy parazitní odrazy od okolních překážek. Užitečné signály byly uměle znehodnoceny Gaussovým širokopásmovým šumem o směrodatných odchylkách $\sigma = 0{,}05$ a $\sigma = 0{,}15$. Před vlastní superpozicí řeči a hluku byly nezkreslené nahrávky shodně normovány tak, aby hodnoty jednotlivých vzorků výsledného součtu plně využily dynamický rozsah <-1;+1>. V první nahrávce dosahuje poměr výkonu signálu k výkonu šumu vypočtený přes všechny časové trajektorie hodnoty $SNR_{IN1} = 7{,}6$ dB. Ve druhé nahrávce je hodnota poměru výkonu signálu k výkonu šumu pouze $SNR_{IN2} = -2{,}0$ dB.



Obr. 11: **(a) Původní řečový signál *„Přišel jsem včera večer pozdě“*, tentýž signál doplněný aditivním Gaussovým šumem ($\sigma = 0{,}05$) a výstupní signál zpracovaný metodou Mapování spektrogramu, (b) původní řečový signál *„Přišel jsem včera večer pozdě“*, tentýž signál doplněný aditivním Gaussovým šumem ($\sigma = 0{,}15$) a výstupní signál zpracovaný metodou Mapování spektrogramu.**

Na obr. 11a a obr. 11b jsou pod sebou postupně vykresleny časové průběhy nezkresleného řečového signálu *„Přišel jsem včera večer pozdě“*, téhož řečového signálu doplněného širokopásmovým Gaussovým šumem a výstupního signálu vyfiltrovaného navrženou metodou Mapování spektrogramu. Na spodních obrázcích je v obou případech zřejmé výrazné potlačení hlučného pozadí mezi jednotlivými úseky řeči.

Po zpracování obou testovacích signálů bylo možné na výstupu metody zjistit tyto hodnoty $SNR_{OUT1} = 12{,}3$ dB a $SNR_{OUT2} = 4{,}6$ dB. Je zřejmé, že v porovnání se vstupními hodnotami SNR_{IN1} a SNR_{IN2} nastalo velké zlepšení šumových poměrů v obou testovacích nahrávkách. V prvním případě byl šum potlačen téměř třikrát a ve druhém případě dokonce více než čtyřikrát. Z obr. 11a a obr. 11b je dále zřejmé úplné odstranění hluku pozadí mezi oblastmi řečové aktivity. Uvnitř užitečné řeči

(v čase i kmitočtu) byl hluk ponechán, ovšem zde je do jisté míry kmitočtově maskován vlastnostmi lidského sluchu.

Na základě subjektivního zhodnocení separovaných průběhů je možné konstatovat, že v případě prvního testovacího signálu nastalo výrazné zvýšení srozumitelnosti oproti vstupnímu degradovanému signálu. V jistých pasážích nahrávky byl separovaný průběh dokonce srovnatelný s originálem. Druhý testovací signál již vykazoval vyšší úroveň zbytkových úzkopásmových impulzních hluků, ovšem i zde byla výstupní nahrávka zcela zbavena energetického širokopásmového šumu, který byl zachován pouze v úzkých oblastech řečové aktivity.

Účinnost separace řečového signálu metodou Mapování spektrogramu výše uvedeným postupem je pro širokopásmové rušení, v podobě např. Gaussova šumu, zcela mimořádná. Technika tak svými separačními schopnostmi při širokopásmovém rušení zcela předčila všechny běžně ve světě užívané techniky shrnuté v kap. 2.

5 ZÁVĚR

V rámci disertační práce byl navržen zcela původní algoritmus extrakce řečových signálů z hlučného prostředí. V prvotní fázi byl kladen důraz na potlačení širokopásmových šumů, které byly obecně považovány za nejnáročnější druh rušení pro svoji konstantní spektrální hustotu výkonu. Princip separace parazitních hluků byl založen na analýze a následném zpracování časově-kmitočtového prostoru. Byl vyvinut zcela netradiční způsob zpracování akustického signálu pomocí standardizovaných operací číslicového zpracování obrazů. Vývoj probíhal s použitím vývojového prostředí MATLAB a celosvětově rozšířeného strukturovaného jazyka C.

Základní separační technika, pracovně nazvaná Mapování spektrogramu, byla detailně matematicky rozebrána a odladěna v prostředí MATLABu a jazyka C. Následovalo rozšíření základní techniky o dílčí algoritmy schopné vhodně reagovat na kmitočtově úzkopásmové a časově nestacionární rušení. Tím byla cílová oblast použití nové techniky mnohonásobně rozšířena. Již během vývoje jednotlivých algoritmů byl kladen důraz na co nejmenší množství vnitřních parametrů, které by před implementací programu do konkrétních zařízení vyžadovaly složité nastavovaní pro optimální provoz v daném akustickém prostředí.

Dále byla provedena rozsáhlá analýza všech použitých operací a provedena náročná optimalizace vyvinutých algoritmů pro práci v reálném čase se zachováním užitných vlastností původní metody. Byla vyvinuta technika dvojité segmentace, která dovoluje rozdělit příchozí akustický signál do vzájemně se překrývajících bloků segmentů a tím snížit zpoždění celé separační techniky způsobené dobou zpracování teoreticky nekonečného signálu. V závěru práce byla objektivně zhodnocena výpočetní a paměťová náročnost jednotlivých operací mapovacího algoritmu a byly stanoveny minimální požadavky na digitální signálové procesory.

LITERATURA

[1] Avendano, C., Hermansky, H.: Temporal Processing of Speech in a Time-Feature Space. Oregon. April 1997.

[2] Hermansky, H., Wan, E. A., Avendano, C.: Speech enhancement based on temporal processing. In Proceedings of the IEEE International Conference on Acoustics - Speech and Signal Processing. May 1995, p. 405-408.

[3] Sovka,P., Pollak,P., Kybic,J.: Extended Spectral Subtraction, In Proceedings of European Conference on Signal processing and Communication. Trieste, September, 1996.

[4] Psutka, J.: Komunikace s počítačem mluvenou řečí. Praha: Academia, 1995. ISBN 80-200-0203-0.

[5] MARTIN, R.: Spectral Subtraction Based on Minimum Statistics. In Proceedings of the EUSIPCO-94. Edinburgh, Scotland, 13.-16. September 1994, p. 1182-1185.

[6] Gonzales, R. C., Woods, R. E.: Digital image processing. Addison-Wesley Publishing Company, Inc. 1992. ISBN 0-201-50803-6.

[7] Ježek, B.: Počítačová grafika I a II. Elektronické studijní materiály - http://lide.uhk.cz/home/fim/ucitel/fujezeb1/www/. Univerzita Hradec Králové, Fakulta informatiky a managementu.

[8] JAN, J.: Číslicová filtrace, analýza a restaurace signálů. Brno: VUTIUM, 2003. ISBN 80-214-1558-4.

[9] SIGMUND, M.: Analýza řečových signálů. Brno: VUT Brno, 2000. ISBN 80-214-1783-8.

[10] DAVIS, G. M.: Noise Reduction in Speech Applications. CRC Press. Florida, 2000. ISBN 0-8493-0949-2.

[11] Quatieri, T. F.: Discrete-Time Speech Signal Processing. New Jersey: Prentice Hall, 2002. ISBN 0-13-242942-X.

[12] Poruba, J.: Separace řečového signálu ze šumového prostředí. Ph.D. Thesis. Brno University of Technology, Faculty of Electrical Engineering and Communication Technology, Department of Telecommunications. Brno 2003.

[13] Oppenheim, A. V., Schafer, R. V.: Discrete-Time Signal Processing. New Jersey: Prentice Hall, 1999. ISBN 0-13-754920-2.

[14] Kahrs, M., Brandenburg, K.: Applications of Digital Signal Processing to Audio and Acoustics. Kluwer Academic Publishers, 1998. ISBN 0-7923-8130-0.

[15] Varho, S., Alku, P.: Separated Linear Prediction - A new all-pole modelling technique for speech analysis. Speech Communication. May 1998, vol. 24, p. 111-121.

[16] Képesi, M., Nagy, Z.: Speech Identification in Noisy Recordings. In Proceedings of the TSP ’99. Brno 1999, p. 110-113. ISBN 80-214-1154-6.

[17] Képesi, M., Nagy, Z.: Potlačení šumu pozadí z řečového signálu mapováním spektrogramu. Elektrorevue – www.elektrorevue.cz. VUT Brno, 2000.

[18] Proakis, J. G., Manolakis, D. G.: Digital Signal Processing, Principles, Algorithms, and Applications. New Jersey: Prentice Hall, 1996. ISBN 0-13-373762-4.

[19] O’Shaughnessy, D.: Speech Communications, Human and Machine. IEEE Press, 1999. ISBN 0-7803-3449-3.

[20] Fischer, S., Simmer, K. U.: Beamforming Microphone Arrays for Speech Acquisition in Noisy Environments. Speech Communication. December 1996, p. 215-229.

[21] Elko, G. W.: Microphone Array Systems for Hands-Free Telecommunication. Speech Communication. December 1996, p.229-241.

[22] Jongseo, S., Wonyong S.: A Voice Activity Detector Employing Soft Decision Based Noise Spectrum Adaptation. In Proceeding of ICASSP. 1998, p. 365.

[23] Te-Won Lee: Independent Component Analysis. Boston: Kluwer Academic Publisher, 1998.

[24] Wu H., Principe J.: Minimum Entropy Algorithms for Source Separation. In Proceeding of the 41st Midwest Symposium on Circuits and Systems Society. Notre Dame, Indiana, 1998.

[25] King, R. A.: TESPAR/FANN An effective new capability for voice verification in the defence environment. Royal Aeronautical Society "The Role of Intelligent Systems in Defence". London, March 1995.

[26] Pollak, P., Sovka, P.: Cepstral Speech/Pause Detectors. In Proceeding of IEEE Workshop on Nonlinear Signal and Image Processing. Neos Marmaras, Halkidiki, Greece, June 20-22, 1995, p. 388-391.

[27] Kambhatla, N.: Local Models and Gausian Mixture Models for Statistical Data Processing. Ph.D. Thesis. Oregon Graduate Institute of Science & Technology. Oregon, January 1996.

[28] Markel, J. D., Gray, A. H.: Linear Prediction of Speech. Springer Verlang, Berlin, 1976.

[29] Rajmic, P.: Exact Risk Analysis of Wavelet Spectrum Thresholding Rules. In Proceedings of 10th IEEE International Conference on Electronics, Circuits and Systems. December 2003, University of Sharjah, United Arab Emirates.

[30] Breit, G., Intaglietta, M.: A modeling cross-spectral analysis technique based on the Prony Spectral Line Estimator (PSLE). Department of AMES-Bioengineering, University of California, San Diego, 1994.

[31] VIDAKOVIC, B.: Statistical Modeling by Wavelets (Wiley Series in Probability and Statistics). New York: John Wiley&Sons, 1999.

[32] KÉPESI, M., PLŠEK, M.: One-Channel Speech Separation Using Spectrogram Modifications In Speech Processing. In Proceedings of 11th Czech-German Workshop - Speech Processing. Prague: BCS, Ltd., September 2001, p. 75 – 75. ISBN 80-86269-07-8.

[33] PLŠEK, M., KÉPESI, M.: One-Channel Speech Separation by The Spectrogram Mapping Method. In Proceedings of Research in Telecommunication Technology RTT 2001. International Scientific Conference. Czech Republic, Lednice, September 2001, p. 275 – 280. ISBN 80-214-1938-5.

[34] PLŠEK, M., KÉPESI, M.: Jednokanálová separační technika Mapování spektrogramu. Elektrorevue - www.elektrorevue.cz, vol. 2001, no. 10, p. 0 - 19. ISSN 1213-1539.

[35] PLŠEK, M.: The Basic Spectrogram Mapping Method of Noisy Speech Signal In Research. In Proceedings of Research in Telecommunication Technology RTT 2003. International Scientific Conference. Slovak Republic, Bratislava, September 2003, p. 1 – 4.

[36] PLŠEK, M.: Speech spectrum smoothing methods by Cepstrum and Pseudo-Cepstrum Weighting. In Proceedings of Telecommunications and Signal Processing. International Scientific Conference TSP 2003. Czech Republic, Brno, 2003, p. 1 – 4.

CURRICULUM VITAE

Osobní údaje

Jméno: **Ing. Martin Plšek**
Datum a místo narození: 4.12.1976 v Brně
Kontakt: plsek@feec.vutbr.cz, xplsek01@seznam.cz, martin.plsek@siemens.com

Vzdělání

1991– 1995	Střední průmyslová škola elektrotechnická v Brně. Studijní obor Elektronické počítačové systémy. Vzdělání ukončeno maturitou s vyznamenáním.
1995 – 2001	Vysoké učení technické v Brně, denní studium na Fakultě elektrotechniky a informatiky. Studijní obor Elektronika a sdělovací technika. Studium ukončeno v červnu 2001 Státní závěrečnou zkouškou.
2001 – 2005	Studium presenční formy postgraduálního doktorského studijního programu na Fakultě elektrotechniky a komunikačních technologií Vysokého učení technického v Brně. Studijní obor Teleinformatika.

Praxe

2002	Technický pracovník Ústavu telekomunikací FEKT VUT v Brně (vývoj moderních metod extrakce řečového signálu z hlučného prostředí).
2002 - 2005	Odborný pracovník Ústavu radioelektroniky Akademie věd České republiky v Praze (odborný pracovník vědy a výzkumu pro syntézu řeči Text-to-Speech a vývoj moderních algoritmů potlačení hluků v řečových signálech).

Účast na řešení projektů

FRVŠ 106/2001/G1	Digitální technologie odstranění hluku z pozadí řečových signálů
FRVŠ 2162/2003/G1	Modelování průběhu základního tónu v systémech TTS
GAČR 102/00/1084	RTD technologie hláskové separace zamaskované v šumu
GAČR 102/04/1097	Zvýrazňování řečového signálu zamaskovaného v šumu
COST OC277	Non-linear methods of speech enhancement
MPO ČR FD-K/040	Příprava služby "Digitální operátor": Nové metody potlač. rušivých signálů
MPO ČR FD-K/125	Aplikace digitální separace řeči v komunikačních technologiích.
MŠMT 262200011	Výzkum elektronických komunikačních systémů a technologií
MŠMT LI002008	Interaktivní oborová knihovna
MŠMT LP01060	Vytvoření encyklopedie komunikačních technologií a její zpřístupnění pomocí internetu

Pedagogické aktivity

Po dobu doktorského studia jsem se podílel na výuce předmětů Číslicové filtry, Konstrukce elektronických zařízení a Datová komunikace. Byl jsem vedoucím 9 ročníkových projektů a 4 diplomových prací.

Další aktivity

V roce 2003 jsem absolvoval tříměsíční odbornou zahraniční stáž na KHBO Oostende v Belgii v rámci programu Socrates – Erasmus.

ABSTRACT

Speech signal carrying message is often degraded by environmental agents during its transmission from acoustic source to listener. Such factors can be a termic wideband noise coming from active and passive devices, external electromagnetic fields, crosstalks and other impulse-sounds. Mostly common speech background noise is caused by sound sources in the speaker's neighbourhood. Microphone receives also parasitic signals, which can be for example crossroads noise, sound of engines, noisy room alternatively loudly music, animal sound and speech signals of other speakers. This work describes a new two-dimensional speech suppression method in time-frequency signal representation. The main idea is not to filter, but to map the signal's spectrogram. First a new algorithm was designed to be able suppress wideband Gaussian and white noise in speech recordings. The algorithm was consecutively expanded about natural shortband separation and nonstationary background noise separation as well. Result programs were simultaneously written in MATLAB and C language. The implementation of the new method highly increases the signal-to-noise ratio and word intelligibility in transmitting speech. The algorithm was tested on a big amount of recordings containing different speakers and background noise. Output speech quality is always very excellent. In the last part of this work, all algorithms were transformed into real-time processing with the later implementation on arbitrary digital signal processors.