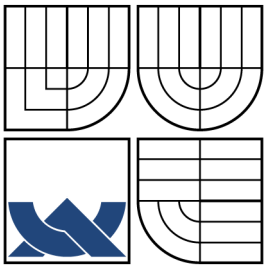
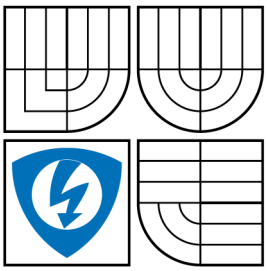


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

OPTIMALIZOVANÁ DETEKCE ŘEČOVÉ AKTIVITY V PROSTŘEDÍ S PROMĚNNÝMI VLASTNOSTMI

ZKRÁCENÁ VERZE DIZERTAČNÍ PRÁCE
BRIEF VERSION OF THE DOCTORAL THESIS

AUTOR PRÁCE
AUTHOR

Ing. IVAN MÍČA

VEDOUCÍ PRÁCE
SUPERVISOR

prof. Ing. ZDENĚK SMÉKAL, CSc.

BRNO 2014

OBSAH

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

CNG Comfort Nois Generator (generátor komfortního šumu)

DTX Discontinuous Transmission (nespojitý přenos)

E short-time Energy (krátkodobá energie)

EM Expectation-Maximization (očekávání-maximalizace)

GMM Gaussian Mixture Model (model směsi Gaussových distribucí)

GSL GNU Scientific Library (knihovna optimalizovaných matematických a statistických funkcí pro jazyky C a Fortran

IID Independent and Identically distributed (nezávislá a identicky rozdělená)

IQR Inter-Quartile Range (mezikvartilové rozpětí)

k-means k Means (k středních hodnot)

ML Maximum Likelihood (maximální věrohodnost)

SNR Signal-to-Noise Ratio (poměr signálu k šumu)

SNR_{seg} Segmentální SNR

SVM Support Vector Machine (metoda podpůrných vektorů)

VAD Voice Activity Detection (detekce řečové aktivity)

ZCR Zero Crossin Rate (míra průchodů nulou)

ÚVOD

První zmínka o praktickém využití automatizované detekce řeči byla v dostupné odborné literatuře nalezena v rámci inženýrského návrhu [BF59] směřujícího k hospodárnějšímu využití podmořských telefonních kabelů přepínáním obvodů tak, že automaticky uvolní telefonní obvod mluvčího, kdykoliv je detekována pauza v řeči a na uvolněný kanál tak může být přepnut jiný mluvčí. Autoři přitom vycházejí z dřívější publikované analýzy charakteristických intervalů v telefonních hovorech, z níž vyplývá, že statisticky je zhruba 60 % každého telefonního hovoru vyplněno pauzami (v článku je citována práce autorů A. C. Norwina a O. J. Murphyho „Characteristic Time Intervals in Telephone Conversation“ z roku 1938).

Jádro úvahy zmíněného inženýrského návrhu spočívá ve statistickém odvození minimálního počtu potřebných kabelů a v návrhu telekomunikačního protokolu, který zajistí jednak optimální vytížení kabelů, tedy primární cíl, a jednak bude momentálně odpojeným koncovým účastníkům pouštět slabý šum, aby neznejistěli, zda nedošlo k přerušení spojení. Autoři pak dochází k závěru, že realizací tohoto návrhu může být zhruba zdvojnásobena přenosová kapacita. K samotným řečovým detektorům se v tomto návrhu přistupuje jako k marginálnímu problému a o jeho technickém řešení zde není zmínka.

Autoři jako jedno z východisek své úvahy předpokládají existenci detektoru schopného detekovat „i nejslabší řeč v pětimilisekundovém nebo kratším intervalu“ a za nejzávažnějším problém, který v této souvislosti řeší návrhem vhodného protokolu, považují riziko ztráty „prvních několika milisekund“ promluvy, než detektor stihne sepnout.

Taková praktická motivace tedy zřejmě byla na počátku několika desetiletí světového výzkumu a tak optimistické byly původní předpoklady ohledně složitosti problematiky detekce řečové aktivity (VAD).

Současně s rozvojem oboru číslicového zpracování řeči i s rozvojem výpočetní techniky a s nárůstem reálných aplikací rostly také nároky na přesnost detekce a začaly se objevovat potíže při uplatnění automatických detekčních algoritmů vlivem nepříznivých pracovních podmínek, takže přirozeně začala vznikat nová oblast výzkumu.

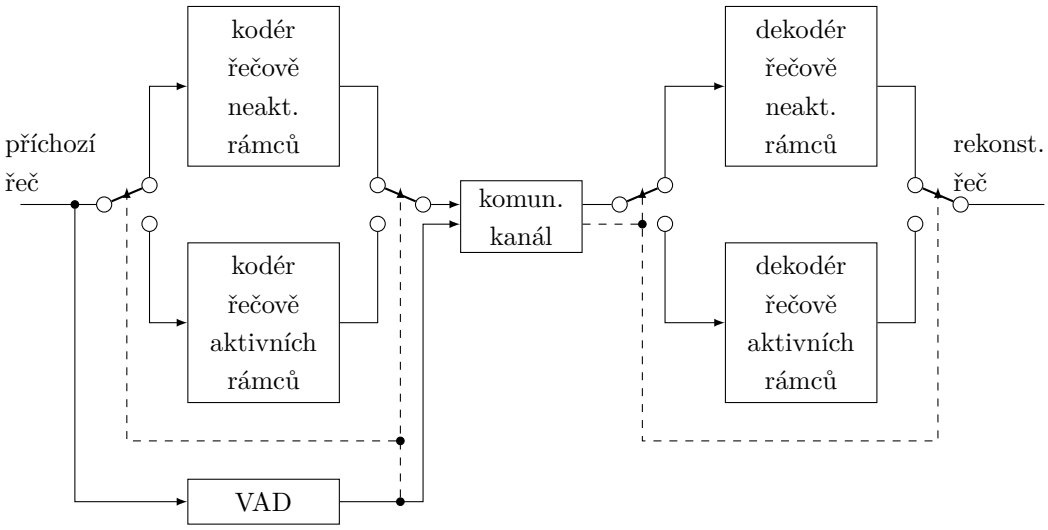
Jeden z prvních ucelených experimentů zcela věnovaných klasifikační úloze typu *znělá řeč–neznělá řeč–ticho* [AR76] je založen na parametrizaci řeči pomocí sady měr periodicity, energie signálu a lineární predikce v samostatných řečových rámcích, tj. bez uvažování korelace navazujících rámců. Uplatňuje parametrický gaussovský model a je typickým příkladem úlohy strojového učení s učitelem. Trénování i testování je prováděno pouze na čistých laboratorních nahrávkách a vliv rušení v reálných podmínkách se zde ještě nediskutuje.

Modernější statistické detekční algoritmy, jejichž rozmach začal [SS98], staví na fisherovské statistice, konkrétně na testu poměrem věrohodností. Tyto metody pracují většinou se spektrálními a od spektra odvozenými příznaky a již také využívají časově korelačních vlastností spektra řeči. Dosahují již velmi dobrých výsledků i v obtížných podmínkách. Spoléhají však na předpoklad čistě šumových několika počátečních rámců signálu, na nichž se trénuje model šumu a jsou tak na přechodu mezi metodami učení s učitelem a učení bez učitele.

Potenciál metod učení bez učitele se rozvíjí nejpozději od [Yin+11], kde se objevily perspektivní návrhy na řešení přetrvávajících problémů s identifikovatelností složek a sekvenční aktualizace. Tato práce staví na percepčních příznacích a rovněž využívá gaussovský model, přičemž ale nestaví na apriorních předpokladech šumového počátku signálu. V současnosti se tento směr zdá být jednou z hlavních oblastí výzkumu v oboru VAD.

1 PROBLEMATIKA DETEKCE ŘEČI

1.1 Aplikační význam detekce řeči



Obr. 1.1: Význam bloku VAD pro kodek G.729B [IT12]

Typickou telekomunikační aplikací automatické detekce řeči je kódování řeči pro úsporný přenos. Na obrázku 1.1 je zařazení detekčního bloku (VAD) v rámci standardního kodeku G.729. Annex B tohoto standardu obsahuje schéma komprese řečově neaktivních úseků hovoru. Toto schéma sestává ze tří algoritmů využívaných k redukci přenášeného datového toku.

Význam a funkce bloků je stručně popsána následovně:

- *Detektor řečové aktivity (VAD)* přepíná vokodér mezi kódováním/dekódováním řečových nebo tichých pasáží. Rozhoduje se každých 10 ms ve shodě s použitou délkou vysílaných rámců. Je-li rámec označen za aktivní, použije se ke kódování a k dekódování řečový kodek. Je-li označen za neaktivní, pak se uplatní další dva algoritmy (DTX, CNG).
- *Nespojitý přenos (DTX).* Dostává informace o řečové aktivitě aktuálně zpracovávaného rámce od modulu VAD, sleduje změny v řečově neaktivních rámcích a podle potřeby aktualizuje parametry neaktivních rámců. Je-li vyžadována aktualizace, vyšle kodér informace o energetické hladině a o spektrální obálce, aby na straně přijímače mohl být generován signál podobný původnímu řečově neaktivnímu signálu.
- *Generátor komfortního šumu (CNG).* Na straně přijímače generuje pseudo bílý budicí signál s parametry přijatými od DTX modulu na základě rozhodnutí VAD modulu. Šum je dále zpracován interpolovanými LPC filtry stejně, jako kdyby se jednalo o řečově aktivní rámec. Tím je na přijímací straně přibližně reprodukován šum pozadí strany vysílací.

Pomineme-li detaily, tak oproti průkopnické práci zmíněné v úvodu, kdy se úspory zamýšlelo dosáhnout čistě přepínáním účastnických okruhů, je v moderním pojetí navíc pouze blok pro kódování řečově aktivních rámců. Základní idea, včetně „reprodukce“ šumu, zůstala tedy i po padesáti letech nezměněna. Avšak s odstupem půl století praktických aplikací a s využitím některých znalostí z oboru zpracování řeči je dnes možno realističtěji formulovat samotnou detekční úlohu i požadavky kladené na výkonnost algoritmu za předpokládaných provozních podmínek.

1.2 Intuitivní detektor

Na obrázku 1.2 nahoře je časový průběh krátkého úseku bezšumové promluvy z upravené laboratorní databáze [Pel11] s vyznačením řečové aktivity a se schematickým přepisem lingvistického obsahu. Pod řečovým signálem jsou průběhy čtyř často používaných segmentálních příznaků, a to logaritmu krátkodobé energie $E_{\log}$, počtu průchodů signálu nulovou úrovní ZCR, prvního koeficientu lineární predikce LPC_1 a plochosti spektra X_{flt}. Pro každý z příznaků je rovněž uveden průběh jeho první diference Δ, čili *rychlostního, delta* příznaku. Signál v tomto případě je převzorkován na 16 kHz, a výpočet příznaků je prováděn v 20 ms rámcích vážených Hammingovým oknem s posuvem po 10 ms.

Průběhy všech příznaků vykazují jistou korelaci s vyznačením řeči. Zjevně nejlépe však vystihuje řečovou aktivitu samotný příznak $E_{\log}$, jehož hodnota v pauzách pravidelně klesá hluboko pod průměrnou hladinu řeči, což se též projeví na odvozeném rychlostním příznaku, jehož vysoká absolutní hodnota vyznačuje hranice řečové aktivity. Na základě těchto pozorování je možno intuitivně sestavit jednoduchý algoritmus detekce řeči prahováním hodnoty příznaku $E_{\log}$, který je popsán jako

$$va = \begin{cases} 1 & \text{pro } E_{\log} \geq \vartheta_{\text{E}}, \\ 0 & \text{jinak}, \end{cases} \tag{1.1}$$

kde $va = 1$ značí řečově aktivní, $va = 0$ řečově neaktivní segment a ϑ_{E} je pevně daná hodnota diskriminačního prahu. Jediným volným parametrem tohoto detektoru je diskriminační práh sledovaného příznaku.

Výsledek predikce přítomnosti řeči pomocí tohoto algoritmu a označení chyb v detekci oproti referenčnímu značení při zvoleném prahu $\vartheta_{\text{E}_{\log}} = -3{,}7$ je vidět v grafu na obrázku 1.3.

Pro souhrnné vyhodnocení výkonnosti detektoru a pro účely srovnání účinnosti detekce v různých podmínkách je třeba vhodně definovat míry.

1.3 Vyhodnocení výkonnosti

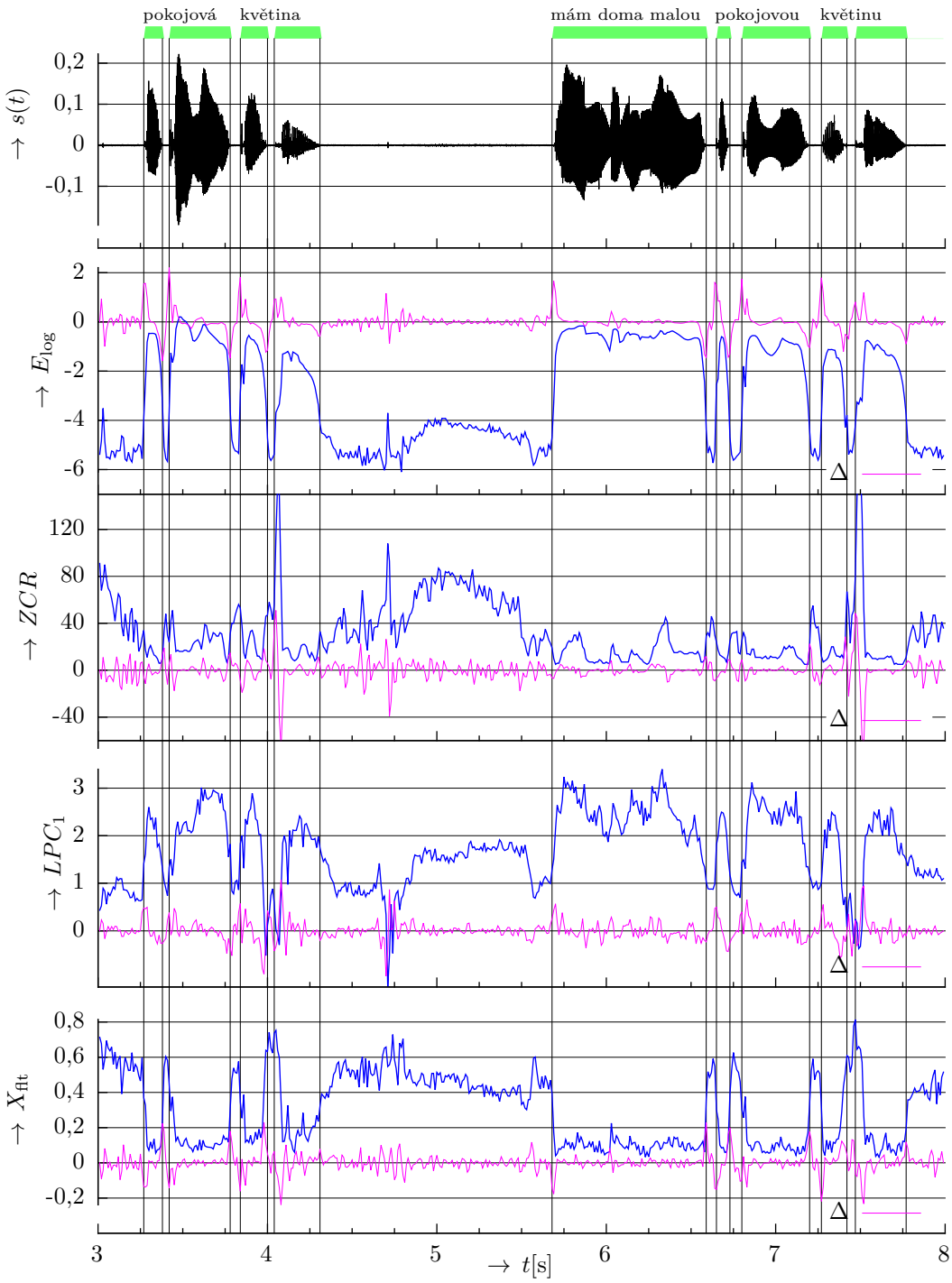
Výkonnost binárního klasifikátoru je možno popsat kontingenční tabulkou 2×2, která je v tomto kontextu nazývána též *matice záměn* a definujeme ji

$$\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}, \tag{1.2}$$

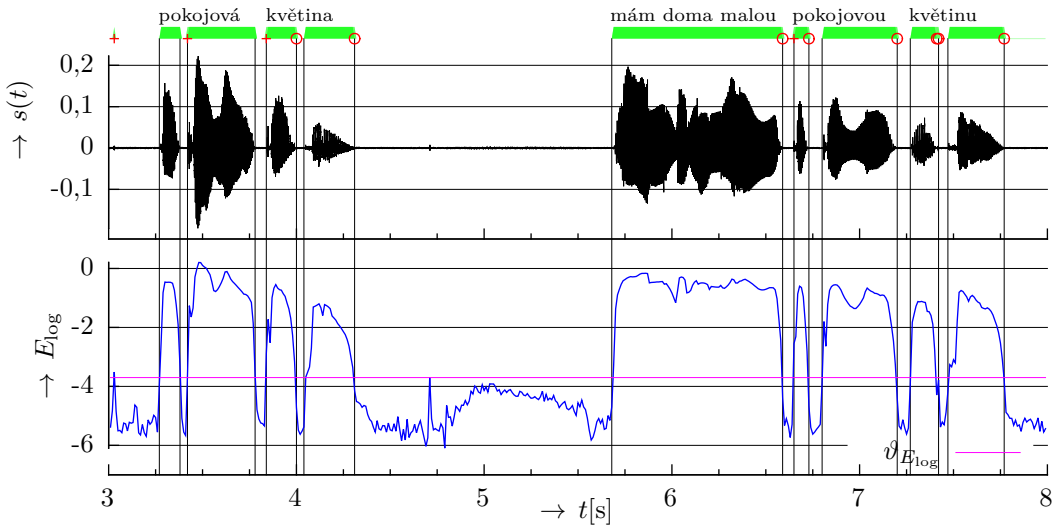
kde TP, TN značí počty správně detekovaných řečových (P), resp. neřečových (N) rámců, FP, FN značí počty chybných detekcí.

V matici záměn je obsažena veškerá informace o výkonnosti binárního klasifikátoru, jak skutečný počet rámců v jednotlivých třídách, tak jejich zatřídění detektorem. Výhodou uvedení absolutních počtů pro porovnání algoritmů je možnost srovnání jak délek testovacích nahrávek a tedy i relativních vah výsledků, tak poměrného zastoupení řeči a pauz v celé nahrávce. Nevýhodou je však nižší přehlednost takového srovnání. Proto se z matice záměn odvozují další míry [Pre08], zaměřující se na aplikační preference. Pro vyhodnocení detekčních algoritmů se obvykle používají:

- *vybavení* (angl. recall), též *podíl skutečně pozitivních* (angl. true positive rate – TPR) je poměrová míra vyjadřující, jak velký díl všech řečových rámců byl detektorem odhalen, tj. $recall = \frac{TP}{TP+FN}$,
- *přesnost* (angl. precision), též *pozitivní predikční hodnota* (angl. positive predictive value) se definuje jako $prec = \frac{TP}{TP+FP}$, čili postihuje snahu, aby mezi všemi pozitivně detekovanými byly pouze skutečně pozitivní,



Obr. 1.2: Označkovaný řečový signál se segmentálními a Δ příznaky.



Obr. 1.3: Výsledek detekce řeči pomocí intuitivního algoritmu prahováním krátkodobé energie s pevným prahem. Chyby klasifikace v porovnání s referenčními značkami jsou vyznačeny pro každý chybný segment. Přesnost je 98,5%, vybavení 97,8%, celková správnost 97,6%.

- F_1 hodnota je definována jako harmonický průměr přesnosti a vybavení, tedy $F_1 = \frac{2 \cdot prec \cdot recall}{prec + recall}$, nabývá hodnot od 0 do 1, takže je přehlednou mírou vyjadřující výkonnost vhodnou pro porovnání detekčních algoritmů. Vybavení, přesnost a F_1 se obvykle používají v aplikacích, kde jsou tolerovány spíše falešně pozitivní než falešně negativní detekce – typicky v řečových kodecích.
- *podíl falešně pozitivních* (angl. false positive rate – FPR) vyjadřuje, jak velký díl všech neřečových rámců byl algoritmem chybně označen za řečové a jeho využití je častější v aplikacích, kdy jsou falešně pozitivní detekce považovány za horší chybu než falešně negativní, např. při některých automatických analýzách, je definován $FPR = \frac{FP}{FP+TN}$,
- *správnost* (angl. accuracy) je aplikačně neutrální měrou definovanou jako $acc = \frac{TP+TN}{TP+TN+FP+FN}$

Výkonnost popsaného energetického detektoru, jak je vidět na obrázku 1.3, je popsána maticí záměn $\left(\begin{smallmatrix} 266 & 4 \\ 8 & 221 \end{smallmatrix}\right)$.

Jednotlivé chybně detekované segmenty na hranicích řeč/neřeč jsou prakticky nevyhnutelné, když uvážíme, že nejmenší časovou jednotkou, se kterou algoritmus pracuje, je délka posuvu okna – obvykle 10 ms. Při ručním vyznačování referenčních značek žádné takové omezení neplatí, takže hranice referenčních značek nejsou synchronizované na pevnou periodu, a tak v místech, kde jsou přesahy delší než polovina posuvu okna, jsou detekovány chyby. Tyto chyby jsou však percepčně nepostřehnutelné.

Procentuální správnost detekce byla ve zmíněném případě 97,6%, přičemž většina chyb jde na vrub hraničním přesahujícím segmentů, a tak je možno tento výsledek prohlásit za téměř ideální.

Tak jednoduchý přístup je však účinný pouze ve velmi příznivých laboratorních podmínkách, jeho uplatnění ve variabilních reálných podmínkách by bylo problematické. Jednak proto, že je zde fixní práh nastaven jednorázově experimentátorem pro konkrétního mluvčího a prakticky neměnné nahrávací podmínky. Není ani schopen se přizpůsobit změně podmínek, např. začne-li vypravěč mluvit hlasitěji, ani nebude optimální pro jiné mluvčí či jinak nastavené zesílení mikrofonu. Především ale detekce založená na tomto jediném příznaku není *robustní* vůči rušivým vlivům *šumů*, které jsou v reálných podmínkách všudypřítomné.

1.4 Vliv šumu

Poměr signálu k šumu v řeči SNR je dle [RS78] definováno v kontextu kvantizačního šumu jako

$$SNR = \frac{\sigma_{\mathrm{s}}^2}{\sigma_{\mathrm{i}}^2} = \frac{\sum_i s^2(i)}{\sum_i n^2(i)}, \tag{1.3}$$

kde $s(i)$ je čistý řečový signál a $n(i)$ je čistý šumový signál, i indexuje vzorky celého řečového signálu, σ^2 značí rozptyl – signály se předpokládají centrované, takže je rozptyl ekvivalentní energii signálu. Dále pomocí (1.3) definujeme *segmentální* SNR_{seg} jako

$$SNR_{\mathrm{seg}} = \frac{\sum_i (w(i)s(i))^2}{(w(i)n(i))^2}, \tag{1.4}$$

kde $w(i)$ značí zvolené váhové okno. Tyto definice použijeme pro simulaci detekce řeči při působení šumu, přičemž ji budeme uvádět decibelové škále, tj. $10\log_{10}\frac{\sum_i s^2(i)}{\sum_i n^2(i)} = SNR[\,\mathrm{dB}]$.

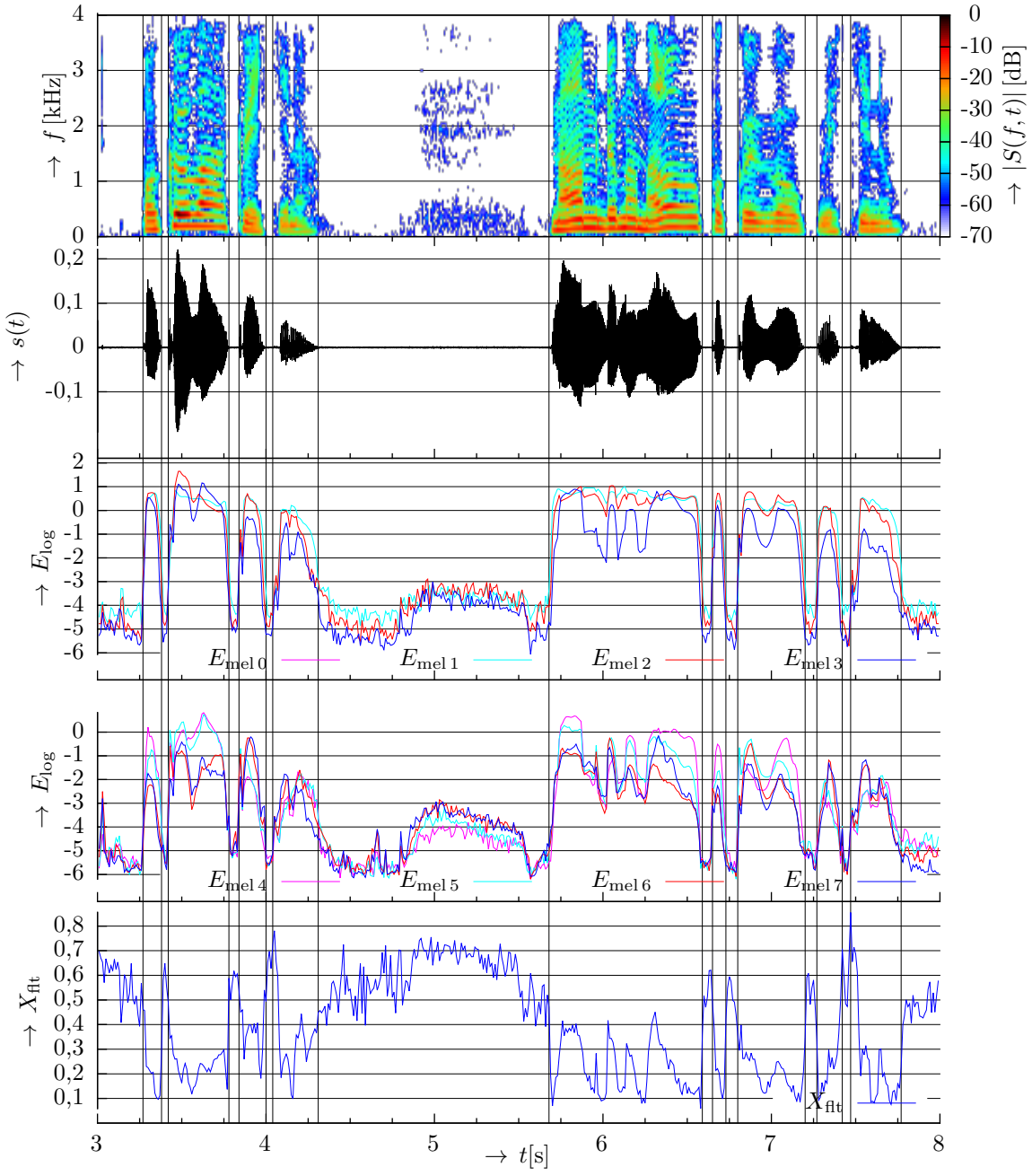
Je použit lineární směšovací model s následnou re-normalizací smíšeného signálu na původní hladinu energie, aby pokud možno neutrpěla variabilita v hlasitosti projevu mezi jednotlivými mluvčími v původní databázi. Vyjdeme-li z definice (1.3), pak směšovací vztah bude

$$x(i) = \frac{s(i) + c_{\mathrm{mix}} n(i)}{c_{\mathrm{E}}}, \tag{1.5}$$

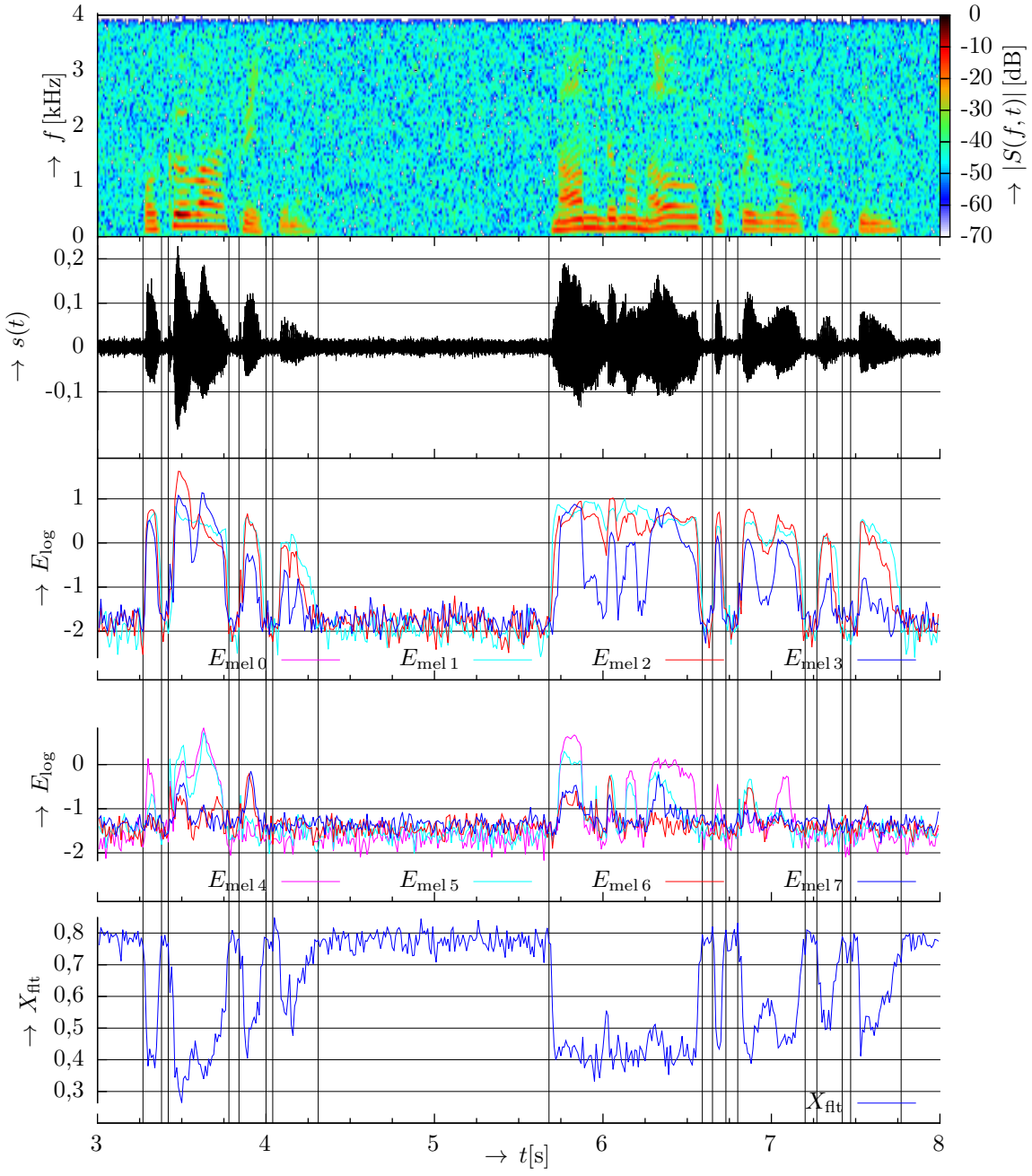
kde $c_{\mathrm{mix}} = \sqrt{10^{-\frac{SNR}{10}}}$ je směšovací konstanta, jsou-li si směšované signály energeticky rovny, tj. $E_{\mathrm{s}} = E_{\mathrm{n}}$. Nejsou-li, pak je třeba napřed dorovnat energii šumu $n(i) = \sqrt{\frac{E_{\mathrm{s}}}{E_{\mathrm{n}}}} n'(i)$, kde n' je původní šumový signál. Konstanta $c_{\mathrm{E}} = \sqrt{1 + 10^{-\frac{SNR}{10}}}$ slouží k re-normalizaci smíšeného signálu na původní hladinu energie.

Vliv specifických typů šumů na na vybrané segmentální příznaky je ukázán na následujících grafech. K čistému řečovému signálu (na obrázku **??**), nahrávanému v bezdozvukové komoře, je jednotlivě přimícháván *bílý* a *hlaholivý* šum z databáze *NOISEX-92* z [VS93] při $SNR = 12\,\mathrm{dB}$ dle směšovacího vztahu (1.5). Původní šumové nahrávky `white.wav` `babble.wav` byly před směšováním převzorkovány na 8 kHz. Všechny segmentální analýzy jsou prováděny po vážení Hammingovým oknem délky 20 ms s posuvem po 10 ms.

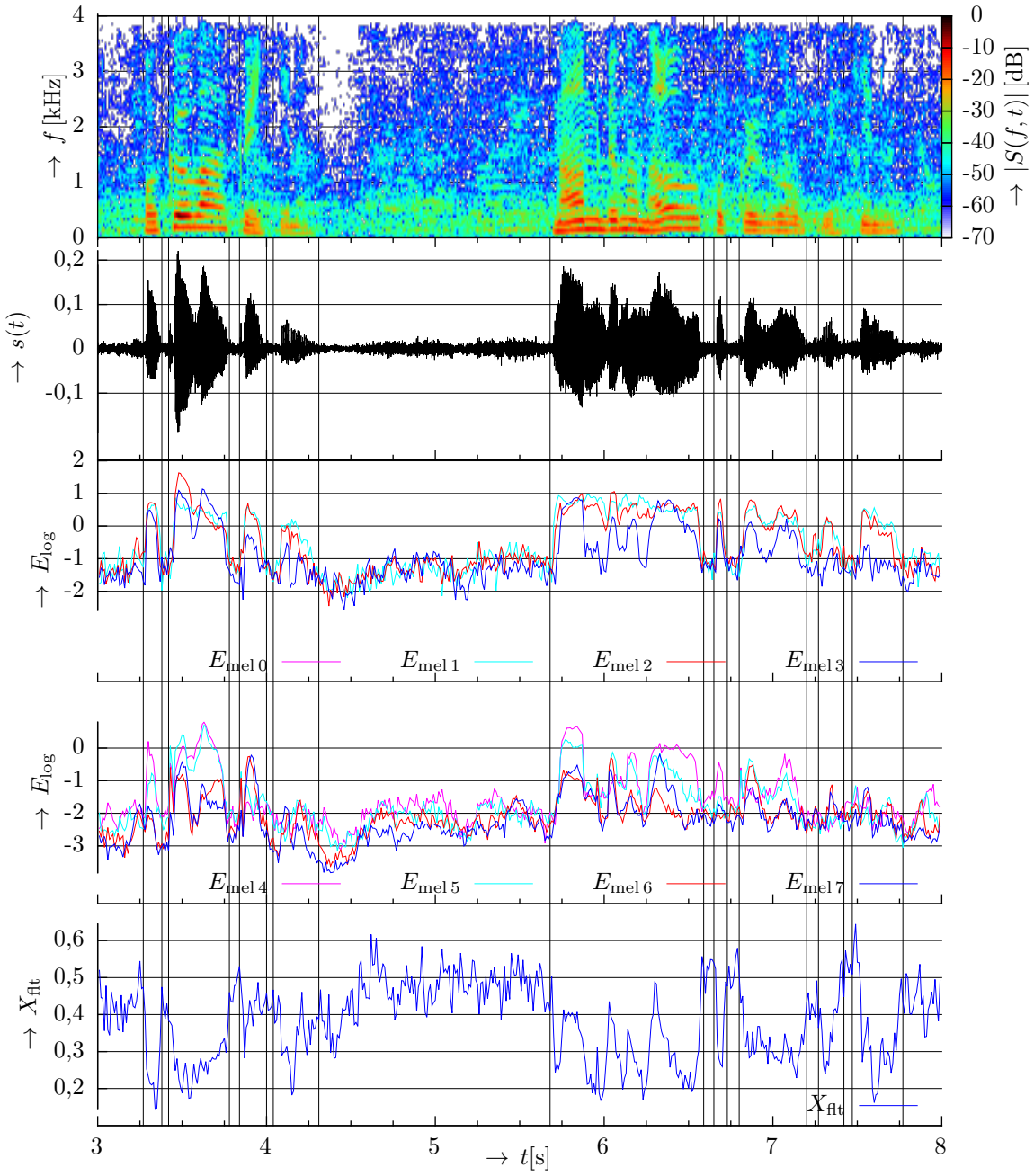
Další často se vyskytující šumy, které způsobují obtíže při detekci řeči, zahrnují např. impulzní rušení, hlasitou hudbu na pozadí. V telekomunikační praxi je časté postižení signálu technickými tóny či vyčkávacími nahrávkami spouštěnými automaticky při přepojování.



Obr. 1.4: Čistý řečový signál. Průběh příznaků mel-frekvenční banky filtrů vykazuje silnou korelaci ve všech pásmech. S rostoucím kmitočtem ale klesá energie, jak je patrné také ze spektrogramu. V úseku přibližně mezi $4{,}8-5{,}5\,\mathrm{s}$ je ve spektrogramu vidět nízkou energii rozloženou přes celé spektrum – projev sotva slyšitelného nádechu.



Obr. 1.5: Řečový signál postižený bílým šumem při $SNR = 12\,\mathrm{dB}$. Bílý šum postihuje celé spektrum, ale na rozdíl od hledané řeči má výkon rozložen přes celé frekvenční spektrum rovnoměrně a je stacionární. Při $SNR = 12\,\mathrm{dB}$ nepředstavuje bílý šum pro detekční algoritmy vážný problém. Je poměrně spolehlivě detekovatelný příznakem *spektrální plochosti*.



Obr. 1.6: Řečový signál postižený hlaholivým šumem při $SNR = 12\,\mathrm{dB}$. Hlaholivý šum, tedy nesourodá směsice hlasů na pozadí řeči, je nestacionární šum s časově-frekvenčními vlastnostmi velmi podobnými hledané řeči a je od ní ve spektrální oblasti velmi obtížně odlišitelný. Spektra se zcela překrývají, hledaná řeč se od šumového pozadí liší pouze vyšší energií a lepší časově-spektrální korelací.

2 DATABÁZE ŘEČI

Nezbytnou součástí testování výkonnosti algoritmů VAD je databáze nahrávek s referenčními časovými značkami na hranicích mezi řečově aktivními úseky a tichem. Existuje několik přístupů k obstarání takové databáze, rozhodujícím kritériem je obvykle dostupnost (cena) a dostatečné pokrytí variability pracovních podmínek aplikace, pro niž je detektor určen.

Jedná-li se o aplikaci s blíže neurčenými pracovními podmínkami, pak bývá snahou dosáhnout co největší variability slovní zásoby, stylů promluvy, mluvčích, i typů rušení. Ideální by byla databáze pokrývající *náhodným výběrem* celou populaci mluvčích, promluv a reálných typů rušení. Databáze tak velkého rozsahu by bylo obtížné i navrhnout, natož sestavit, proto se podmínky vždy v některém směru omezují. Například se použije pouze *příležitostný výběr* nebo se databáze sestaví výběrem z jiných existujících databází původně určených pro jiné účely. To je případ některých univerzálních výzkumných databází, jako je např. databáze [TP09], která vznikla úpravou a výběrem z existujících databází pro rozpoznání řeči získaných v kancelářském prostředí, v automobilu atd.

Je-li výzkum zaměřen se na známé problematické jevy, jako např. na detekce předělů obsahujících neznělá plozíva při zkreslení signálu šumy při proměnných SNR, jako v [Kac06], pak je výběr slovníku cíleně zúžen. Nahrávky se pořizují v akusticky čistém prostředí, aby byl signál co možná nejčistší pro pozdější směšování se specifickými typy rušení, protože vyšší hladiny šumu v referenčních nahrávkách před směšování ztěžují interpretaci výsledků.

Aplikačně specifické databáze vznikají až při aplikovaném výzkumu, kdy jsou již pracovní podmínky dobře známy a je možno získat dostatečně reprezentativní vzorek.

Ve všech případech by měla být snaha o *konzistentní* značení v mezích dané tolerance, aby zbytečně nedocházelo k chybám vyhodnocení výkonnosti detektoru vlivem výkyvů ve značení.

V této práci jsou použity dvě databáze, které vznikly během výzkumu problematiky VAD.

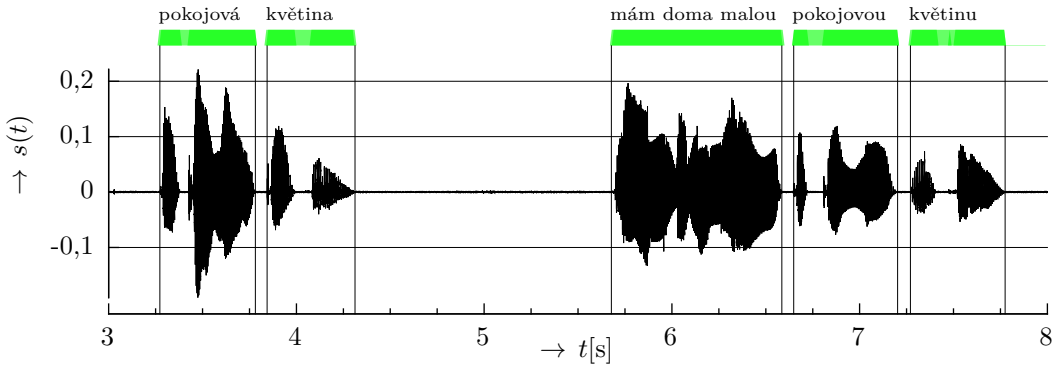
- *Laboratorní* databáze vytvořená v rámci bakalářské práce [Pel11], která byla dále přizpůsobena pro účely tohoto výzkumu.
- *Aplikačně specifická* databáze záznamů telefonních hovorů z prostředí kontaktního centra a byla použita v [Míč+10], která obsahuje typické i okrajové případy a pokrývá tedy celé spektrum reálně se vyskytujících kombinací zkreslení komunikačním kanálem, rušení na pozadí, slovníku i stylu promluvy, a to v širokém rozsahu věkových skupin.

2.1 Laboratorní databáze

Nahrávky pro laboratorní databázi byly pořízeny v bezdozvukové komoře s využitím kvalitního mikrofonu pro záznam hlasu při vzorkování 44 100 Hz. Jedná se tedy o ideální, v reálném akustickém prostředí nedosažitelné podmínky, kdy analyzovaný audio materiál není postižen rušivým šumem okolí, zkreslení a zašumění vlivem mikrofonu je zanedbatelné, řeč je zaznamenána v celém kmitočtovém spektru, nedochází ani k přeznívání vlivem akustického přenosového kanálu, protože doba dozvuku T_{60} je prakticky nulová.

V grafu a obrázku 2.1 je ukázka průběhu a značení jednoho vzorku databáze [Pel11]. V tomto krátkém úseku je zaznamenána mužská promluva v řečnickém stylu *umělecký přednes*, který je charakteristický klidným tempem, zřetelnou artikulací a výraznou intonací. V grafu je dále schematicky zaznamenán lingvistický obsah promluvy a rovněž vyznačení hranic řeči tak, jak je provedl tvůrce databáze.

Z kontury okamžité výchylky v kontextu s přepisem a s vyznačenými hranicemi je patrné:



Obr. 2.1: Ukázka průběhu a původního značení laboratorní databáze

- Hranice mezi jednotlivými slovy (lingvisticky) obecně nekorespondují s pauzami v řeči; v úseku tří slov „mám doma malou“ není v řeči ani jedna zřetelná pauza; naopak v obou slovech „pokojová“, „pokojovou“ je vždy dobře patrný akusticky neaktivní úsek (zřejmě na předělu „po–ko“) a kratší neaktivní úsek se vyskytuje také v obou realizacích slova „květina“.
- Autor databáze žádný z těchto akusticky neaktivních úseků uvnitř slov neoznačil jako pauzu v řeči, ačkoliv srovnatelně dlouhé pauzy mezi slovy označoval.

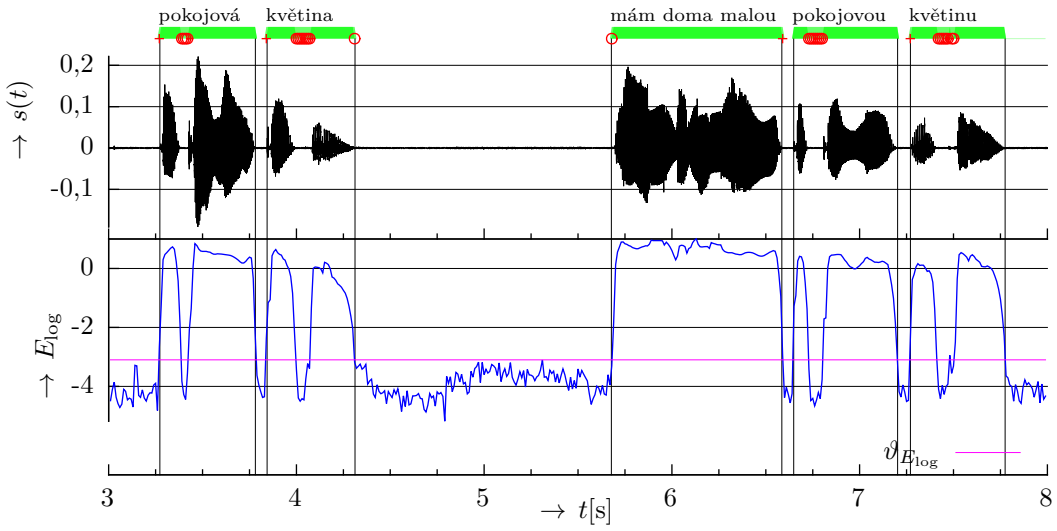
Pozorným vizuálním i poslechovým rozborem dalších nahrávek této databáze tvořené pro účely detekce hlasové aktivity je možné se přesvědčit, že toto „opomíjení“ zřetelných pauz uvnitř slov je systematické, zato jsou důsledně značeny pauzy srovnatelné délky, pokud se vyskytují mezi slovy. Vysvětlení tohoto jevu tedy zřejmě není v ledabylosti při značení hranic řeči, ale spíše v subjektivním vnímání toho, co ještě klasifikovat jako řečovou aktivitu a co již označit jako pauzu.

Ke kořenů tohoto jevu se můžeme dobrat, pohlédneme-li na značkování databáze jako na psychoakustický experiment (viz též [Mel05], kde jsou popsány i metody které vedou k lépe definovaným a konzistentnějším výsledkům experimentů, přičemž podrobný rozbor přesahuje rámec této práce). Zhruba je možno příčiny nepřesného značkování shrnout jako kombinaci vlivu obtížné rozlišitelnosti hranic některých řečových předělů (subjektivní vliv) a nepozornosti z únavy (další občasné chyby ve značení). Pečlivou přípravou psychoakustického experimentu je možno některé z těchto vlivů omezit, je však prakticky vyloučeno zcela je potlačit. Z podstaty vzniku těchto nekonzistencí navíc plyne obtížná kvantifikovatelnost chyby.

Z vyhodnocení energetického detektoru na obrázku 2.2, kde se pracovalo 10 ms segmenty s 10 ms posuvem při vážení obdélníkovým oknem, je možno si udělat představu, jak může nekonzistentní značení může ovlivnit další vývoj. Přestože celková délka těchto neoznačených pauz je v poměru k délce signálu malá, není vhodné ji zcela zanedbat. Může totiž jednak zkreslit vyhodnocení přesnosti testovaného algoritmu označením správně detekovaných pauz za FN, jednak může při použití databáze jako trénovací množiny přímo ovlivnit klasifikační schopnosti detektoru.

2.1.1 Objektivizace značení

Bez použití vyhlazovacího schématu by výsledek detekce vypadal jako v grafu 2.2, některé pauzy označené, jiné nikoliv – každé vyhodnocení účinnosti by tím bylo zkreslené. Naopak po zahrnutí vyhlazovacího schématu by bylo možné dosáhnout toho, že nebudou označené žádné z těchto krátkých pauz, tedy ani ty původně správné. Značení by bylo konzistentní, ale již se zahrnutím ad-hoc schématu, což by zkreslovalo výsledky během vývoje, kdy se obvykle žádné vyhlazovací schéma



Obr. 2.2: Výsledek detekce a vyhodnocení energetického detektoru oproti původnímu referenčnímu značení dle [Pel11] – a základ metody pro objektivizované referenční značení, které je použito pro některé simulace v této práci.

nepoužívá. Třetí možností je dodatečné označení těchto pauz – za předpokladu, že výsledek bude subjektivně v pořádku. Pozorným poslechem tohoto pětisekundového úseku při *binárním maskování* detektorem označených míst je ovšem možno se přesvědčit, že percepčně je ztráta těchto FN úseků zcela nepostřehnutelná.

Bylo tedy navrženo spojit jednoznačně lepší lidskou schopnost rozpoznat, co do řeči nepatří, dohromady s algoritmickou detekční schopností u krátkých pauz, takže algoritmu je dovoleno změnit vybrané původní P značky na N, nikoliv naopak. Tato metoda byla ověřena poslechem a s její pomocí pak bylo upraveno značení celé trénovací databáze.

Jako objektivní test přijatelnosti byla vyhodnocena výkonnost standardního detektoru G.729B, , při němž byla databáze s původním a poté s modifikovaným značením použita jako testovací množina. Pro tento účel byly použité nahrávky převzorkované na 8000 Hz. Žádné další úpravy signálu prováděny nebyly.

Z grafu 2.3 je dobře vidět, že proti modifikovaným referenčním značkám ve všech případech vzrostlo procento FP chyb, významně pokleslo procento FN a celková chyba je vždy menší. Vzhledem k vychýlení detektoru G.729B směrem k pozitivním detekcím, se mírný vzrůst hlášených FP proti modifikované referenci dal očekávat, protože navrženou metodou jsou potlačeny krátké pauzy s nádechy, které G.729B detekuje zpravidla jako řeč.

Podstatné je, že pokles FN detekcí je výrazně větší než vzrůst FP. Toto je ovšem podmíněno provedením nahrávek trénovací databáze v idealizovaných poměrech. Testovací databáze, které byly nahrávány v reálném provozu, není možné tímto způsobem upravit k tréninkovým účelům – u aplikačně specifických databází to však ani není žádoucí, neboť tam je součástí aplikačního zadání i způsob vyznačení pauz (zda je potřeba rozlišovat na úrovni kolem 50 ms, nebo zda stačí jednotky sekund).

Z tabulky 2.2 je vidět, že modifikace značek s použitím prostého energetického detektoru se příliš neliší od modifikace doplněné ještě mediánovým filtrem. Poslechem rozdílového signálu částí označených za FN doplněného vizuální kontrolou nebyly nalezeny subjektivně vadné úseky v žádné z těchto modifikací. Krátké úseky označené jako FP byly při kontrolním poslechu téměř vždy po-

Tab. 2.1: Srovnání výsledků G.729B VAD s modifikovanými a s původními referenčními značkami testovací databáze na části *umělecký přednes*

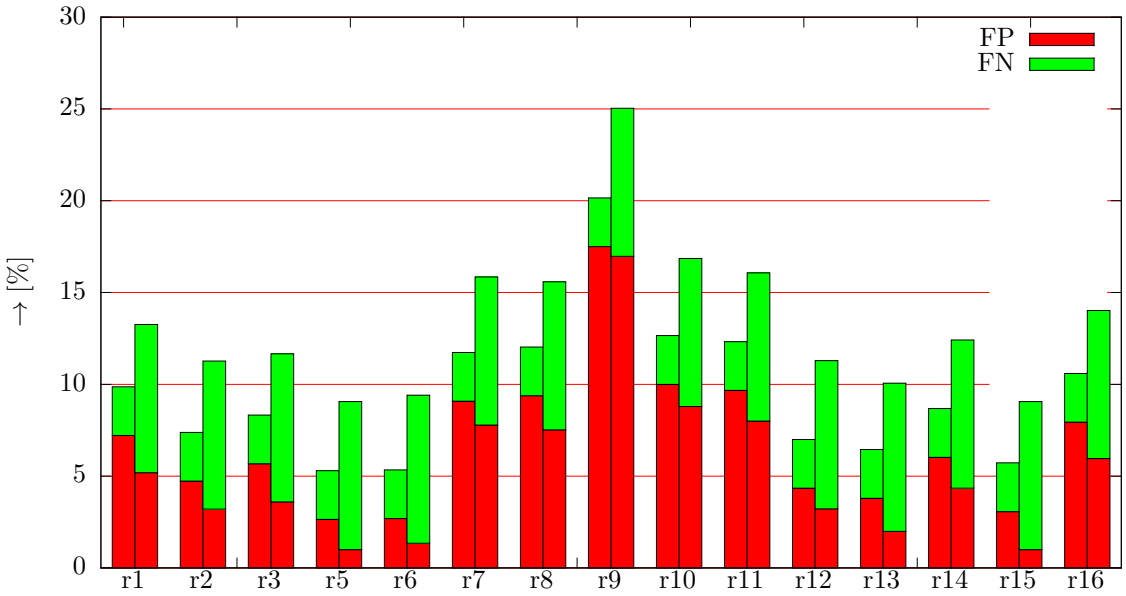
ID	DÉLKA [S]	ŘEČ [%]	CHYBA [%]	FP [%]	FN [%]
r1	149,9	63,24	9,87	7,21	2,66
		70,68	13,26	5,19	8,07
r10	173,3	64,70	10,62	10,00	0,62
		68,34	11,84	8,79	3,05
r11	145,8	68,98	11,56	9,67	1,89
		74,34	13,59	8,00	5,59
r12	160,5	64,64	5,46	4,34	1,12
		68,18	6,77	3,22	3,55
r13	128,9	64,26	4,99	3,79	1,19
		69,02	6,15	1,99	4,16
r14	123,5	71,29	7,54	6,02	1,51
		76,40	9,30	4,35	4,95
r15	144,0	59,09	5,19	3,07	2,12
		67,09	9,04	0,99	8,05
r16	129,9	70,18	9,78	7,94	1,84
		76,39	12,03	5,96	6,07
r2	151,5	62,66	5,87	4,73	1,15
		67,21	7,37	3,20	4,17
r3	138,4	63,18	7,76	5,67	2,09
		70,27	10,71	3,60	7,11
r5	134,6	65,54	4,08	2,64	1,44
		70,63	5,88	1,00	4,89
r6	143,8	62,69	4,15	2,69	1,47
		67,25	6,02	1,34	4,68
r7	122,1	65,38	9,63	9,08	0,55
		68,50	10,15	7,78	2,37
r8	128,5	70,52	10,51	9,38	1,13
		75,67	11,94	7,52	4,42
r9	125,0	70,62	17,89	17,50	0,39
		72,07	18,29	16,97	1,32
CELKEM	2099,8	65,62	8,26	6,84	1,42
		70,64	10,11	5,26	4,85

Tab. 2.2: Porovnání „chyby“ modifikovaného značení proti původním „referenčním“ značkám nahrávek *uměleckého přednesu*. U každého ID jsou v horním řádku výsledky algoritmu po dodatečném průchodu mediánovým filtrem a v dolním jsou výsledky bez této dodatečné úpravy.

ID	DÉLKA [S]	ŘEČ [%]	CHYBA [%]	FP [%]	FN [%]
r1	149,9	70,68	7,43	0,00	7,43
			7,41	0,00	7,41
r10	173,3	68,34	3,63	0,00	3,63
			3,72	0,00	3,72
r11	145,8	74,34	5,40	0,02	5,38
			5,47	0,00	5,47
r12	160,5	68,18	3,56	0,01	3,55
			3,69	0,00	3,69
r13	128,9	69,02	4,78	0,01	4,77
			4,84	0,00	4,84
r14	123,5	76,40	5,13	0,01	5,12
			5,21	0,00	5,21
r15	144,0	67,09	8,01	0,01	8,01
			8,10	0,00	8,10
r16	129,9	76,39	6,24	0,02	6,22
			6,21	0,00	6,21
r2	151,5	67,21	4,55	0,00	4,55
			4,65	0,00	4,65
r3	138,4	70,27	7,10	0,00	7,10
			7,23	0,00	7,23
r5	134,6	70,63	5,11	0,01	5,10
			5,15	0,00	5,15
r6	143,8	67,25	4,58	0,01	4,57
			4,79	0,00	4,79
r7	122,1	68,50	3,18	0,02	3,15
			3,33	0,00	3,33
r8	128,5	75,67	5,17	0,01	5,16
			5,24	0,00	5,24
r9	125,0	72,07	1,47	0,01	1,46
			1,61	0,00	1,61
CELKEM	2099,8	70,64	5,03	0,01	5,03
			5,12	0,00	5,12

tvrzeny jako správně označené experimentátorem. Výjimečně se vyskytující percepčně vnímatelné výpadky původního značení u neznělých plozív byly v souladu s navrženou metodou ponechány v původním značení a jejich existence byla pouze zaznamenána.

Ve prospěch modifikace s mediánovým filtrem bylo rozhodnuto na základě vizuální kontroly – obvykle se jednalo o výpadky v ojedinělých 10 ms rámcích, kde sice energie klesla pod nastavený práh, ale obvykle jen velmi těsně, a tato FN detekce byla často jen na vrub volby počátku segmentace, tedy objektivně neopodstatněná.



Obr. 2.3: G.729B VAD proti modifikované (vlevo) a původní (vpravo) sadě referenčních značek – nahrávky uměleckého přednesu

Objektivizovaná databáze je použita pro některé simulace v této práci.

2.2 Aplikačně specifická databáze

Během řešení projektu byla vytvořena databáze nahrávek z kontaktního centra a z nich byl proveden výběr obtížných nahrávek na nichž standardní detektory často výrazně chybovaly. Přibližně polovina nahrávek je zkreslená průchodem telekomunikačním kanálem – to jsou nahrávky ze strany zákazníků. Druhou část tvoří nahrávky na straně operátorů, které jsou více či méně postižené hlaholivým šumem prostředí kontaktního centra – SNR je kolísavé podle toho, jako blízko mikrofonu daného operátora se momentálně projevují rušivé hlasy kolegů. Databáze obtížných nahrávek sestává ze 42 vzorků a je heterogenní téměř po všech myslitelných stránkách:

- délky nahrávek jsou v řádech jednotek až stovek sekund,
- procentuální zastoupení řeči kolísá od 0% až po 75%, jsou zde i několika minutové čistě šumové nahrávky
- je velká variabilita šumů a SNR mezi nahrávkami,
- operátoři i zákazníci jsou všech věkových skupin,
- vyskytují se velmi nepříznivé případy, jako hudba či vyzváněcí nebo technické tóny na pozadí

Tato databáze tvoří jádro srovnávacích testů.

3 TECHNIKY DETEKCE

Jedna z prvních prací, věnující se zcela klasifikaci signálových segmentů na třídy znělá řeč, neznělá řeč, ticho, modeluje Gaussovou distribucí pět segmentálních příznaků a je založena na strojovém učení s učitelem [AR76]. Trénování i testování je prováděno pouze na čistých laboratorních nahrávkách. Neuplatňuje však *naivní bayesovský model*, jako mnoho pozdějších prací založených na parametrickém modelu, ale modeluje příznakový prostor i s kovariancemi.

Velký rozmach statistických detektorů postavených na testu poměrem věrohodností s měkkou klasifikací přinesly práce [SS98; SKS99], které byly dále upravovány např. aplikací vyhlazovacího schématu na logaritmickou věrohodnost [CK01] nebo zahrnutím příznakového vektoru přímo do modelu věrohodnosti [Ram+05]. Tato skupina statistických detektorů dosahuje poměrně dobrých výsledků i v méně příznivých podmínkách, ale podmínkou správné funkčnosti je možnost algoritmu natrénovat vnitřní modely šumu vždy na počátku signálu, kde se nevyskytuje řeč.

Obvykle se v detekčních algoritmech řeč i šum modeluje Gaussovým rozdělením [Gór+08], mimo jiné pro relativní jednoduchost takových modelů. Bylo však ukázáno [GZ03], že řeč má spíše charakter Laplaceova rozdělení, a tak se objevily i v oblasti VAD algoritmy používající výstižnějších modelů [CKM06], čímž bylo dosaženo dalšího zlepšení, ovšem za cenu zvýšení složitosti modelu.

Byly zaznamenány také pokusy o využití podstatně složitějších statistických modelů šumů i řeči [TR07; TR08] – experimentální detektory postavené na detekci okamžiku změny parametrů podmíněného heteroskedastického modelu, známého spíše z ekonomické statistiky [Bol86] – avšak pro příliš vysokou algoritmickou složitost, která přinesla jen mírné zlepšení v laboratorních případech se tento směr již dále nerozvíjel.

Soustavná tendence ke zjednodušování algoritmů při současné snaze o nalezení pokud možno univerzálně využitelného modelu pro detekci řeči se projevuje také v [Gór+06], kde se již využívá shlukové analýzy C-means nad množinou příznaků sub-pásmových logaritmických energií. V této práci jsou však ještě použity šumové prototypy, které se vytváří v počátečních segmentech signálu – tedy je zde opět implicitní předpoklad čistě šumového počátku.

Na myšlence využití shlukové analýzy staví také práce [Yin+11], kde jsou řeč i šum parametrizovány Gaussovým modelem se sekvenční aktualizací v několika pásmech mel-banky a se závěrečným hlasováním. Zavedený model již ale není závislý na čistě šumovém počátku, a tak se tento přístup řadí k metodám strojového učení bez učitele.

3.1 G.729B

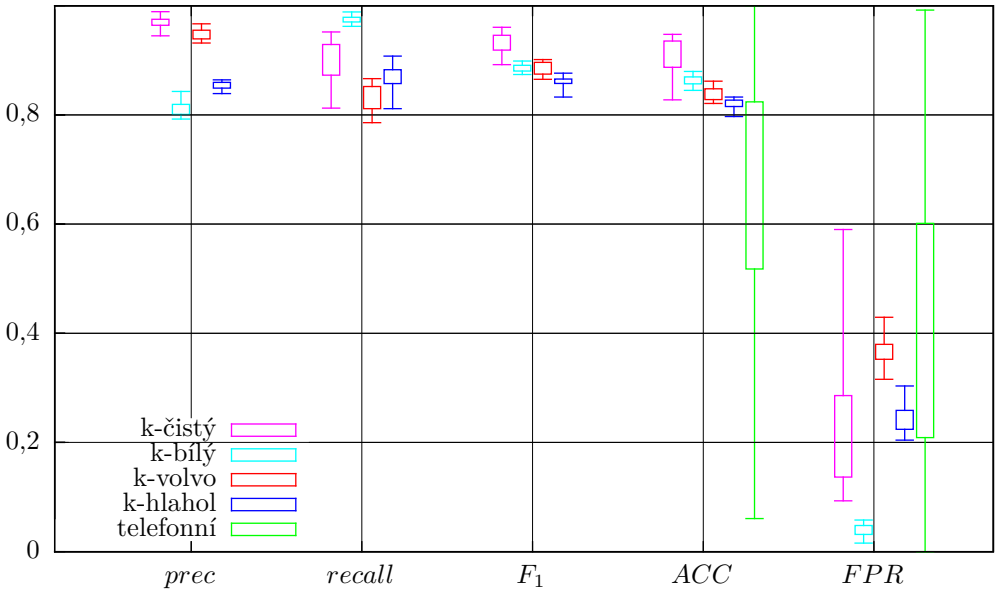
Jako referenční detektor pro srovnání výkonnosti původních a optimalizovaných detekčních algoritmů v různých podmínkách je použit standard G.729B VAD v revizi z roku 2012 [IT12].

3.1.1 Výkonnost detekce

Výsledky testů výkonnosti detekce na laboratorní i aplikačně specifické databázi jsou na obrázku 3.1.

Za povšimnutí či hlubší prozkoumání stojí zejména:

- vysoká hodnota maxima FPR na čisté databázi, která překonává všechny uměle zašumělé nahrávky
- poměrně slabé hodnoty všech výkonnostních měr na přirozené databázi
- extrémně špatné hodnoty správnosti a FPR na telefonní databázi
- neexistující hodnoty vybavení, přesnosti a F_1 na telefonní databázi



Obr. 3.1: Vyhodnocení výkonnosti detektoru G.729B na simulovaných a reálných nahrávkách. Značeno je vždy rozpětí a mezikvartilové rozpětí výsledků v jednotlivých nahrávkách. Jsou užity všechny dříve diskutované míry výkonnosti, tj. vybavení (recall), přesnost (prec), F_1 hodnota, správnost (acc) a podíl falešně ozitivních. První čtyři zkušební sady značené "k-"jsou nahrávky z bezdozvukové komory; čistá a pak uměle zašumělé bílým, automobilovým a hlaholivým šumem při $SNR = 12\,\mathrm{dB}$. Následují výsledky testů na na obtížných vzorcích telefonní databáze.

Poslední jmenovaný fakt lze vysvětlit zvoleným postupem při zobrazování – pokud některá z měr vyšla jako *nedefinovaná hodnota*, čili `NaN`, pak pochopitelně nebyla do grafu zanesena – to je případ nahrávek, v nichž není žádná označená řeč a tedy vybavení vychází bez ohledu na výsledek detektoru vždy $recall = \frac{0}{0}$, potažmo i odvozená míra F_1 vyjde taktéž `NaN`. Obdobně u přesnosti, jeli nahrávka označena jako čistě šumová a přitom nedojde k žádné falešně pozitivní detekci, vychází $prec = \frac{0}{0} =$ `NaN`. Šance, že podobným způsobem selže FPR je nízká – nutnou podmínkou by byla čistě řečová nahrávka. Jediná míra z uvedené pětice, která nemůže selhat nikdy, je *správnost*.

K pochopení příčin špatných výsledků detekce v ostatních případech je třeba s rámcovou znalostí algoritmu prozkoumat detailněji vybrané případy, kdy detektor v některé hodnocené míře vážně selhal.

3.1.2 Algoritmus

Algoritmus je detailně popsán v [IT12]. Zde uvedeme pouze informace podstatné pro pochopení příčin selhávání algoritmu ve sledovaných případech.

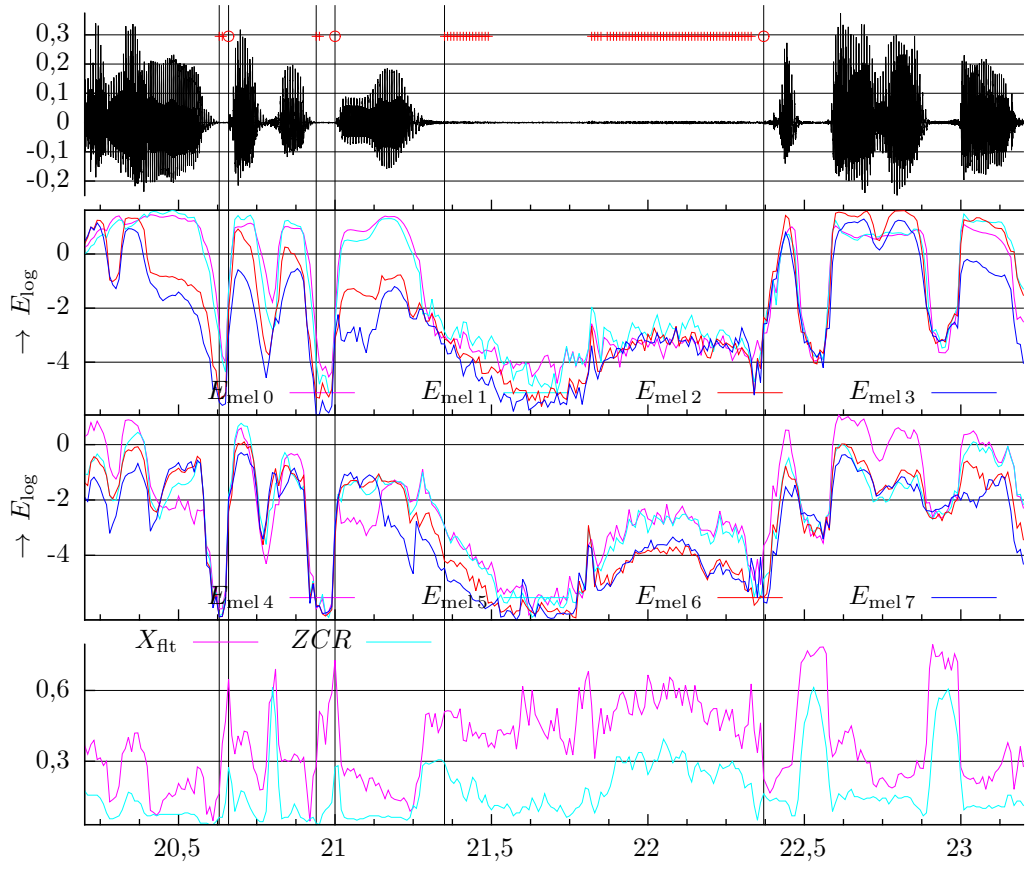
Předpokládá se vzorkovací kmitočet 8 kHz.

1. *Předzpracování:* dělení signálu dvěma a následná horní propust 2. řádu s mezním kmitočtem 140 Hz.
2. *Váhovací okno:* 30 ms nesymetrické; 0 . . . 199 vzorků je poloviční Hammingovo, 200 . . . 239 je první čtvrtina cosinového cyklu; posouvá se po 10 ms,
3. Jsou použity 4 skupiny segmentálních příznaků: *energie celého pásma*, *energie dolního pásma*, *spektrální páry* vypočtené z LPC a *míra průchodů nulou*.
4. Prvních 32 segmentů se použije pro inicializaci klouzavých průměrů šumových charakteristik. Během této inicializace detektor rozhoduje pouze na základě energie celého pásma – překročíli 15 dB, prohlásí se za řečový, jinak je považován za šumový. Přitom se počítá se celočíselnou interpretací 16 bit hodnot, takže 15 dB odpovídá řádově hodnotě 10^{-3} v desetinném vyjádření.
5. Dále si detektor udržuje informaci o dlouhodobém minimu energie, počítá se z posledních 128 segmentů.
6. Následuje fáze výpočtu rozdílových příznaků, tj. rozdílu mezi aktuálními parametry a klouzavým průměrem šumů.
7. Na základě množiny rozdílových příznaků se v klasifikátoru provede předběžné rozhodnutí o hlasové aktivitě.
8. Finální rozhodnutí se provede ve vyhlazovacím bloku.
9. Na podkladě aktuální hodnoty energie v celém pásmu a dlouhodobého minima energie se nakonec rozhodne o aktualizaci klouzavých průměrů šumových charakteristik, k čemuž se pak používá autoregresní model prvního řádu.

3.1.3 Rozbor nepříznivých případů

Vysoká maximální hodnota FPR na čisté databázi – vyšší než na všech zašuměných nahrávkách – patří k nahrávce, jejíž výsek je uveden na obrázku 3.2. Uvedeny jsou rovněž příznaky, které nejsou sice využívány detektorem G.729, uplatní se však v algoritmu dle [Yin+11] a jeho rozšíření, které jsou testovány taktéž na těchto problematických nahrávkách.

Vysvětlení tohoto jevu, kdy výsledky na čisté nahrávce jsou horší než na šumové, spočívá ve způsobu výpočtu příznaků pro klasifikátor rozdílem aktuálních hodnot a dlouhodobých klouzavých průměrů šumových charakteristik.



Obr. 3.2: Typická příčina falešných detekcí VAD G.729B je slyšitelný nádech či výdech. Zde jsou oba případy pohromadě. Kolem 21,5 s vydechuje mluvčí zbylý vzduch, následuje klidová chvíle a potom svižný, ale hluboký nádech.

Je-li dlouhodobě velmi nízká hladina šumu pozadí, pak musí být velmi nízké i dlouhodobé klouzavé průměry energií. V tom případě stačí i relativně nízká energie slyšitelného nádechu rozložená v celém spektru k tomu, aby jej detektor klasifikoval jako řečovou aktivitu.

V přítomnosti širokopásmového bílého šumu jsou klouzavé průměry energií nastaveny odpovídajícím způsobem, tedy mnohem výše než je akustická hladina nádechu. Proto se falešně pozitivní detekce u nahrávek degradovaných bílým šumem vyskytují zcela ojediněle.

Zvýšenou míru FPR v případě hlaholu na pozadí nelze tak snadno vysvětlit. Směsice hlasů na pozadí je už z podstaty obtížně odlišitelná od jednoho hlasu na popředí – rozložení energie ve spektru i časově-frekvenční průběhy jsou si velmi podobné, a tak je obtížné definovat příznaky s dostatečnou diskriminační silou. Zdá se, že nejlépe v tomto případě vyhoví příznaky založené na energii, jejichž diskriminační síla ovšem je dána aktuálním SNR.

3.2 Detektor učící se bez učitele

Správná inicializace metod odvozených od práce [SS98] je explicitně podmíněna neřečovým počátkem signálu, což je ovšem ekvivalentní k ručnímu označení počátečních segmentů za neřečové a dle [Yin+11] tedy spadá mezi metody *učení s učitelem*. Následné aktualizace statistik jsou již prováděny automaticky a spadají tedy mezi metody *učení bez učitele*. Tento kombinovaný přístup je tedy možná zařadit mezi metody *částečně dozorovaného učení* [**CSZ06-intro**].

Předpoklad *neřečového počátku* však v praxi není možno zaručit, v čemž spočívá základní nedostatek těchto algoritmů. Je-li takový algoritmus aplikován na signál začínající řečí, inicializace je provedena chybně, což vede k větší chybovosti detekce v následující fázi s učením bez učitele. Návrhem metody v [Yin+11] učení bez učitele, která staví na shlukové analýze, má být především odstraněn tento omezující předpoklad.

Nejedná se o první pokus o aplikaci shlukové analýzy na problém detekce řeči. Teprve tato práce se však snaží řešit jak otázku identifikovatelnosti, tak i problém algoritmického rozhodnutí, mají-li být v dané situaci hledány dva, nebo jenom jeden shluk, a tak umožňuje bez dalších omezení aplikovat učení bez učitele. Současně řeší také efektivitu sekvenční aktualizace parametrů.

Tento flexibilní, moderní přístup byl zvolen jako základ pro optimalizovaný detektor.

3.2.1 Model

Jako segmentální příznaky jsou použity logaritmované výstupy z mel-banky, *normované* energií pásma. Z praktických důvodů, zejména z důvodů identifikovatelnosti složek je použit *naivní* pravděpodobnostní model, každé pásmo je tedy bráno zcela samostatně až do závěrečného hlasování.

Základní idea je ve stručnosti popsána v *pásmu s vysokým SNR* a za předpokladu, že se vyskytují řečově aktivní i neaktivní úseky, kde má logaritmická energie *bimodální rozdělení* (předpokládá se směs Gaussových rozdělení). Dále se předpokládá, že řeč se vyskytuje nezávisle na šumu.

- Řečový mod logaritmické energie má vyšší hodnotu než šumový, protože je v něm obsažena energie řeči i šumu dohromady, obě třídy jsou tedy dobře separovatelné. Rozdíl mezi středními hodnotami řečového a šumového modu představuje aposteriorní SNR. Tím se zároveň po shlukové analýze řeší *identifikace* složek [**AMR09**]. Máme tedy po úvodní inicializaci modelu dva dobře vyjádřené shluky mel-příznaků, reprezentované dvěma složkami modelu gaussovských směsí (GMM), šumový s nižší energií a řečový s vyšší.
- *Nově příchozí hodnoty* pak můžeme klasifikovat podle optimálního prahu, který minimalizuje celkovou chybu danou překrytím distribucí příznaků podmíněných řečovou či neřečovou třídou v daném pásmu, viz obrázek 3.3.

- Obdobně se klasifikuje v každém pásmu, čímž se omezí vliv úzkopásmových vysoce energetických rušení a zároveň se lépe projeví nízkoenergetické hlásky nezasahující všechna pásma. Souhrnné (nevyhlazené) rozhodnutí z celého segmentu se pak získá *hlasováním* ve všech pásmech.
- *Průběžné přizpůsobování modelu* měnícím se pracovním podmínkám je umožněno sekvenční aktualizací parametrů GMM v každém pásmu samostatně.

Zahrnutí **krajních případů** do modelu, jako jsou pásma s *nízkým SNR* nebo čistě šumový počátek, je popsáno v následujícím podrobném rozboru původního algoritmu:

1. Signál je vážen 20 ms dlouhým Hammingovým oknem s posuvem po 10 ms a z těchto segmentů se počítá FFT.
2. Tvar filtrů mel-banky s $N = 8$ pásmy, aplikované následně na modulové spektrum, je na obrázku **??**.
3. Poté jsou výstupy ve všech pásmech vyhlazeny mediánovým filtrem délky 5.
4. Model se inicializuje na prvních $M = 60$ segmentech (počáteční odhad GMM parametrů).
5. Z parametrů GMM se vypočítají optimální prahy.
6. Optimální prahy se upraví dle aplikačních preferencí.
7. Klasifikuje se celý inicializační úsek v každém pásmu zvlášť.
8. Odhlasuje se předběžná klasifikace v celém inicializačním úseku.
9. Aplikuje se vyhlazovací schéma a přejde se z *inicializačního* dávkového režimu do fáze *sekvenčního vyhodnocování* v reálném čase...

Počáteční odhad GMM. Dvousložkové GMM jsou použity v každém pásmu k modelování bimodální distribuce energie ve směsi řeči a šumu. V pásmech s *nízkým SNR* nebo bez přítomnosti řeči má však logaritmická energie *unimodální rozdělení* a nastávají potíže jednak s interpretací výsledků, jednak často i s konvergencí EM algoritmu pro dvousložkový GMM, takže je nutno algoritmus EM dodatečné podmínit, aby vyhověl i krajním případům.

V každé iteraci EM, se před fází *Expectation* kontroluje splnění následujících podmínek:

$$\mu_{k,1} > \mu_{k,0} + \delta, \tag{3.1}$$

což je podmínka bimodálního rozdělení, δ je zvolený práh; Není-li podmínka splněna, tak se rozdělení v daném pásmu považuje za *unimodální* a čistě šumové. V tom případě se v tomto pásmu vytvoří *virtuální řečová složka* se střední hodnotou

$$\mu'_{k,1} = \mu_{k,0} + \delta \tag{3.2}$$

a všechny rámce v tomto pásmu se označí za šumové.

Obdobně se hlídá, zda platí

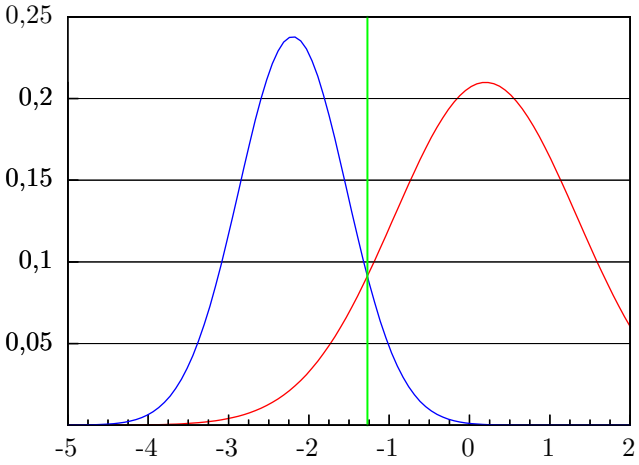
$$\sigma^2_{k,1} > \sigma^2_{k,0}, \tag{3.3}$$

kde se vychází z předpokladu, že šum má stacionárnější povahu než řeč. Při porušení této podmínky se nastaví

$$\sigma^{2\prime}_{k,1} = \sigma^2_{k,0}. \tag{3.4}$$

Prosazování těchto dvou podmínek vede v pásmech s nízkým SNR nevyhnutelně ke zhavarování algoritmu EM – jedna složka je soustavně vytlačována z husté oblasti, což vede rychle k poklesu prioru této složky na nulu. Proto je zavedena třetí podmínka, která zajišťuje, že řečový prior neklesne pod minimální stanovenou hodnotu

$$\pi_1 > \epsilon, \tag{3.5}$$



Obr. 3.3: Průběh hustot dvou složek GMM s vyznačením teoreticky optimálního prahu maximalizujícího správnost detekce.

kde ϵ je zvolená minimální hodnota apriorní pravděpodobnosti výskytu řeči. Dojde-li během iterací EM algoritmu k porušení této podmínky, pak se nastaví

$$\pi_1' = \epsilon, \tag{3.6a}$$

$$\pi_0' = 1 - \pi_1' \tag{3.6b}$$

a algoritmus EM se ukončí.

Optimální diskriminační práh se odvodí následovně; Označme v každém pásmu samostatně:

- x_k logaritmická energie v čase k,
- $z \in \{0, 1\}$ 0 pro řeč, 1 pro šum,
- $\lambda = \{\mu_z, \sigma_z^2, \pi_z | z = 0, 1\}$ sada parametrů GMM,
- $p(z)$ apriorní rozdělení řeči a šumu, je rovno π_z,

Bimodální rozdělení (gaussovská směs o dvou složkách) se zapíše jako

$$p(x_k|\lambda) = \sum_z p(x_k|z, \lambda)p(z), \tag{3.7}$$

přičemž věrohodnost

$$p(x_k|z, \lambda) = \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left\{-\frac{(x_k - \mu_z)^2}{2\sigma_z^2}\right\}. \tag{3.8}$$

Za předpokladu IID bude

$$p(\boldsymbol{x}|\lambda) = \prod_{k=0}^{M-1} p(x_k|\lambda), \tag{3.9}$$

sadu parametrů λ pak ze známé posloupnosti $\boldsymbol{x} = \{x_0, \ldots, x_{M-1}\}$ odhadneme jako

$$\arg\max_{\lambda} p(\boldsymbol{x}|\lambda). \tag{3.10}$$

Z odhadnutých parametrů modelu λ lze odvodit rozdělení energií pro řeč i pro šum a následně lze stanovit *optimální práh* x minimalizující klasifikační chybu z podmínky

$$p(\theta|z = 0, \lambda)p(z = 0) = p(\theta|z = 1, \lambda)p(z = 1), \tag{3.11}$$

jak je ukázáno v grafu křivek teoretického modelu na obrázku 3.3. Tuto podmínku je možno rozepsat explicitně pomocí parametrů modelu λ jako rovnici:

$$\frac{\pi_0}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} = \frac{\pi_1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{(\theta-\mu_1)^2}{2\sigma_1^2}\right\}, \tag{3.12}$$

což po úpravách a po logaritmování vede na kvadratickou rovnici

$$\theta^2(\sigma_0^2-\sigma_1^2) + \theta(2\sigma_1^2\mu_0 - 2\sigma_0^2\mu_1) + \sigma_0^2\mu_1^2 - \sigma_1^2\mu_0^2 - 2\sigma_0^2\sigma_1^2 \ln\frac{\sigma_1\pi_0}{\sigma_0\pi_1} = 0 \tag{3.13}$$

Optimální práh θ je ten z obou kořenů, který splňuje podmínku $\mu_1 > \theta > \mu_0$.

Aplikační preference. Práh odvozený v (3.13) je optimální z pohledu *správnosti* detekce – minimalizuje celkovou chybu. Aplikace však může mít jiné preference, např. již zmíněný standardní VAD dle G.729B upřednostňuje vysokou míru *vybavení* před *přesností*. Pomocí parametru $\gamma \in [0 \dots 1]$ je možné preference doladit.

$$\theta'_k = \gamma(\theta_k - \mu_{k,0}) + \mu_{k,0} \tag{3.14}$$

V každém pásmu se potom rozhoduje na základě posunutého prahu θ'_k. Jakýkoliv posun od vypočteného optima je však vždy za cenu snížené *správnosti*.

Klasifikace. V každém pásmu se uplatní vypočtený diskriminační práh na všechny segmenty stejně:

$$z = \begin{cases} 1 & \text{pro } x_k \geq \theta'_k, \\ 0 & \text{jinak,} \end{cases} \tag{3.15}$$

Hlasování. Rozhodování pouze na základě jednoho pásma není robustní vůči variabilitě šumů. Hlasováním ve všech pásmech se robustnost zvýší, výsledkem je předběžná klasifikace řeč/neřeč.

Vyhlazovací schéma. Je aplikováno na výsledky předběžné klasifikace, aby se dále omezily náhodné výpadky detekce. Je použito schéma s přednostním nastavením řeči – každý rámec předběžně označený za řeč bude i takto označen i po vyhlazení, ale předběžné neřečové segmenty mohou být tímto schématem přeznačeny na řečové, pokud předtím byla dostatečně dlouhá posloupnost předběžných řečových. Tím se omezí výpadky řeči na nízkoenergetických koncích řečových úseků.

Ve fázi *sekvenční aktualizace* parametrů GMM a průběžné klasifikace, následující po *inicializační* fázi, se zpracovávají nově příchozí příznakové vektory jednotlivě:

Sada parametrů v čase k se značí $\lambda_k = \left\{\mu_{k,z}, \sigma^2_{k,z}, \pi_{k,z} | z = 0, 1\right\}$. Snahou aktualizace parametrů modelu je udržování co možná nejvýstižnějšího modelu rozdělení mel-energií řečových a šumových rámců v nedávné minulosti. Autoři zavedli novou metodu sekvenční aktualizace GMM, jež je výpočetně mnohem méně náročná než tradiční schémata, která s každým nově příchozím vzorkem přepočítávají odhady K rámců zpětně, a to rovnocenně bez ohledu na časovou vzdálenost od aktuálního rámce. K tomu účelu zavádí parametr *zapomínání* α, který interpretují přes maximální zapamatovanou délku K zhruba jako $\alpha = \frac{K}{K+1}$. Následují definice metody sekvenční aktualizace pomocí parametru α:

V každém pásmu se začíná výpočtem aposteriorních pravděpodobností

$$p(z|x_k, \lambda_{k-1}) = \frac{\pi_{k-1}\mathcal{N}(x_k|\mu_{k-1}, \sigma_{k-1})}{\sum_z \pi_{k-1}\mathcal{N}(x_k|\mu_{k-1}, \sigma_{k-1})} \tag{3.16}$$

Aktualizují se priory

$$\pi_{k,z} = \alpha\pi_{k-1,z} + (1-\alpha)p(z|x_k, \lambda_{k-1}) \tag{3.17}$$

a následně se omezí podmínkou (3.5).

Aktualizují se střední hodnoty

$$\mu_{k,z} = \frac{\alpha\pi_{k-1,z}\mu_{k-1,z} + (1-\alpha)p(z|x_k, \lambda_{k-1})x_k}{\pi_{k,z}} \tag{3.18}$$

a omezí se podmínkou (3.1).

Aktualizují se odhady rozptylů

$$\sigma^2_{k,z} = \frac{\alpha\pi_{k-1,z}\sigma^2_{k-1,z} + (1-\alpha)p(z|x_k, \lambda_{k-1})(x_k - \mu_{k,z})^2}{\pi_{k,z}} \tag{3.19}$$

a omezí se podmínkou (3.3).

Poté se hlasováním v pásmech získá předběžná klasifikace a nakonec se aplikuje vyhlazovací schéma. Hodnoty konstant, které doporučují autoři [Yin+11] jsou v tabulce 3.1.

Tab. 3.1: Doporučené hodnoty konstant modelu dle [Yin+11], t_{frmlen} a t_{frmshift} jsou délka, resp. posuv váhovacího okna.

$\alpha = 0{,}99$	$\delta = 3{,}5\,\text{dB}$	$\epsilon = 0{,}03$	$M = 60$
$N = 8$	$t_{\text{frmlen}} = 20\,\text{ms}$	$t_{\text{frmshift}} = 10\,\text{ms}$	

3.2.2 Algoritmus

Náleduje kompaktní zápis původního algoritmu dle [Yin+11]. Použité příznaky jsou v každém mel-pásmu nejprve vyhlazeny mediánovým filtrem délky 5.

1. V prvních $M + 1$ segmentech se provede inicializace modelu:
   (a) V každém z N mel-pásem:
      i. Vytvoří se GMM pomocí EM s omezením ((3.1) - (3.6))
      ii. Určí se práh pomocí rovnice (3.13).
      iii. Práh se doladí na základě aplikačních preferencí dle (3.14)
      iv. Klasifikuje se prvních $M + 1$ segmentů.
   (b) Předběžná klasifikace v každém segmentu se získá hlasováním ve všech N mel-pásmech.
   (c) Závěrečná klasifikace v prvních $M + 1$ segmentech se získá aplikací vyhlazovacího schématu na předběžnou klasifikaci.
2. Pro každý nový segment (v čase $(k + 1)$)
   (a) V každém mel-pásmu:
      i. Vypočítají se posteriory dle (3.16).
      ii. Aktualizují se priory dle (3.17).
      iii. Omezí se priory dle (3.5).
      iv. Aktualizují se odhady středních hodnot dle (3.18)
      v. Omezí se střední hodnoty dle (3.1).
      vi. Aktualizují se odhady rozptylů dle (3.19).
      vii. Omezí se rozptyly dle (3.3).
      viii. Určí se optimální práh z (3.13).
      ix. Práh se doladí dle aplikačních preferencí (3.14).
      x. Klasifikuje se dané pásmo dle (3.15).
   (b) Předběžná klasifikace hlasováním ve všech pásmech.
   (c) Závěrečná klasifikace pomocí vyhlazovacího schématu.
3. Konec.

3.2.3 Výkonnost detekce *v.0*

Původní algoritmus dle [Yin+11] bude označen jako *v.0*, čili výchozí verze. Předběžně byla výkonnost algoritmu testována pouze na čistých vzorcích laboratorní databáze.

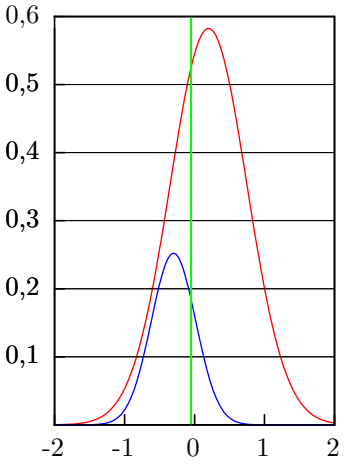
Výsledky správnosti se pohybují zhruba mezi 30 % a 70 % a míra falešne pozitivních detekcí je často téměř stoprocentní – to signalizuje vážný problém v algoritmu, který je třeba najít a odstranit před testováním a optimalizací detektoru na obtížných nahrávkách z telefonní databáze.

4 OPTIMALIZOVANÁ DETEKCE

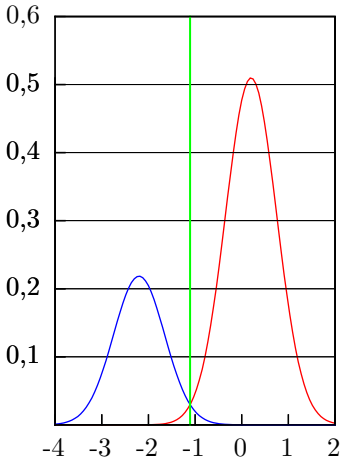
Přímá implementace detektoru učícího se bez učitele dle [Yin+11] (označeného jako verze *v.0*) vykazuje velmi nízkou výkonnost.

První závažný problém, místy se vyskytující ve všech mel-pásmech, je špatně spočítaný optimální klasifikační práh.

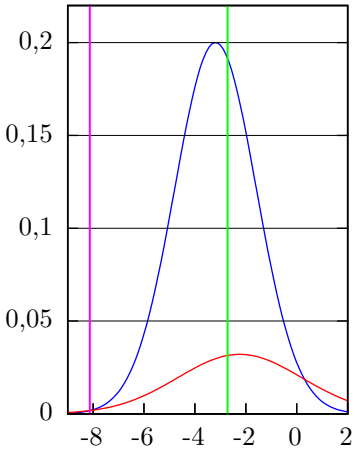
Dá se ukázat, že chybějící hodnoty optimálních prahů jsou důsledkem dvou případu, které nejsou pokryty řešením kvadratické rovnice (3.13), ale patří mezi krajní případy, jak je dále diskutováno a jak je vidět na obrázku 4.1 a), b). Třetí diskutovaný případ, je zodpovědný za vysokou míru falešně pozitivních detekcí.



(a) Kvadratická rovnice má komplexní kořeny → výpočet prahu dle (4.1)



(b) Shodné rozptyly vedou na lineární rovnici (4.2)



(c) Žádný kořen neleží mezi středními hodnotami → výpočet prahu dle (4.1)

Obr. 4.1: Krajní případy výpočtu optimálního prahu z parametrů GMM. V případě c) je kromě „optimálního“ označen také „falešný“ práh. Pro porovnání s typickým případem viz obrázek 3.3.

- V prvním případě, kdy je celá křivka neřečové složky pod křivkou složky řečové, nemá kvadratická rovnice reálné kořeny. Komplexní kořeny je těžko interpretovat a diskriminační síla příznaku je malá, takže výsledek z tohoto pásma bude vždy zatížen velkou chybou. Tento výsledek by bylo asi nejlépe interpretovat tak, že byl zvolen příliš složitý gaussovský model se dvěma složkami, zatímco by dobře vyhověl i jednodušší model s jedinou složkou. V rámci původního algoritmu je však formálně potřeba ponechat 2 složky a stanovit práh. Jako nejjednodušší řešení se jeví aritmetický průměr mezi středními hodnotami (4.1).
- V druhém případě, kdy jsou rozptyly obou složek shodné, se výpočet zjednoduší na lineární rovnici. To ovšem *musí* být v implementaci detekováno a skutečně řešeno lineární rovnicí (4.2). Při pokusu o mechanické řešení kvadratické rovnice by vycházela nula ve jmenovateli.
- V třetím případě leží řečový mod pod křivkou neřečové složky, která má menší rozptyl, takže žádný z kořenů neleží mezi středními hodnotami obou složek. Při naivní implementaci vylučovacím algoritmem, tj. „není-li první kořen mezí středními hodnotami, vybere se druhý“, je nalezen „falešný“ práh. Tento případ je dále ošetřen (4.1).

$$\theta_{\text{means}} = \frac{\mu_0 + \mu_1}{2} \tag{4.1}$$

$$\theta_{\text{lineq}} = \frac{\mu_0^2 - \mu_1^2 + 2\sigma^2 \ln \frac{\pi_0}{\pi_1}}{2(\mu_0 - \mu_1)} \tag{4.2}$$

Všechny zmíněné případy jsou postupně ošetřeny ve vývojových verzích

- *v.0.1*: detekce komplexních kořenů kvadratické rovnice → výpočet prahu dle (4.1)
- *v.0.2*: detekce shodných rozptylů → výpočet prahu dle (4.2)
- *v.0.3*: detekce „falešných" prahů → výpočet prahu dle (4.1)

Nastavením vyšší hodnoty konstanty α autoregresního modelu bylo ve verzi *v.0.4* dosaženo dalšího zlepšení, avšak neřeší to podstatu problému, že optimální práh se přizpůsobuje nejvíce aktuálním hodnotám signálu. Proto byla navržena experimentální alternativa aktualizace. Alternativní přístup zahrnuje změnu v aktualizaci odhadů parametrů GMM a současně také jiný přístup k výběru složitosti modelu.

Ve experimentální verzi *v.A* bylo upuštěno od autoregresního modelu a sekvenční aktualizace se prováděla pro každý příchozí bod novým odhadem modelu, obdobně jakoby se jednalo o inicializaci. To je extrémně výpočetně náročné a není to zamýšleno jako praktická alternativa, ale pouze pro ověření předpokládaného efektu změny sekvenčního modelu ceteris paribus.

Slibný výsledek verze *v.A* byl motivací k finálnímu návrhu algoritmu pracujícího dávkově, označeného verzí *v.opt*. Tento algoritmus zahrnuje kromě dávkového výpočtu prahu ještě další změnu, která byla vynucena častým selháním v případech, kdy signál vzorku obsahoval krátký úsek konstantních hodnot energií v některých pásmech. To způsobovalo adaptaci jedné ze složek GMM na tento mnohonásobný bod – a kolaps algoritmu. Nejprve bylo navrženo zavést do EM algoritmu další omezení: Byla stanovena nejnižší dovolená hodnota rozptylu obou složek, nikoliv pouze řečové jako v původním algoritmu. Dále byla zavedena podmínka hlídající pokles prioru neřečové složky pod stanovenou minimální hranici. Tím byl algoritmus stabilizován a bylo možno detekovat existenci těchto případů, ale to samo o sobě falešně pozitivní detekce neřešilo, protože se objevil druhý aspekt tohoto problému. Většina takových úseků, které jsou v telefonních nahrávkách způsobeny průchodem telefonním kanálem, leží totiž hluboko pod střední hodnotou energie přirozeného šumu, takže ani uměle nastavená minimální hodnota rozptylu v kombinaci s minimální hodnotou prioru nepostačí k tomu, aby se šumová složka adaptovala na skutečný šum v nahrávce, a tak je tento šum mylně považován za řečovou složku.

Tímto pozorováním je ovšem odůvodněna třetí složka GMM a bylo navrženo finální řešení, modifikací dávkového detektoru takto: Je-li v nahrávce detekována složka s rozptylem pod mez stanovenou relativně k rozptylu druhé složky, pak je zvýšena složitost modelu na tři složky, přičemž třetí složka s nízkým rozptylem je vždy považována za šumovou. Vím bylo ve verzi *v.opt* dosaženo výrazného zlepšení, kdy správnost detekce dosáhla v celé telefonní databázi 85,% správnosti a pokles falešně pozitivních detekcí na 10,9%, což je oproti standardu G.729B významné zlepšení, viz srovnávací tabulku 4.1.

Tab. 4.1: Výsledky srovnávacích testů G.729B VAD a dvou verzí optimalizované detekce s algoritmem učení bez učitele na databázi z telekomunikačního provozu.

ID	MÍRA[%]	G.729	*v.0.4*	*v.opt*
0	ACC:	83,0	39,5	39,5
	FPR:	20,1	72,3	72,3
1	ACC:	80,6	52,8	88,1
	FPR:	37,8	99,5	8,1

Tab. 4.1 – *Pokračování na další straně*

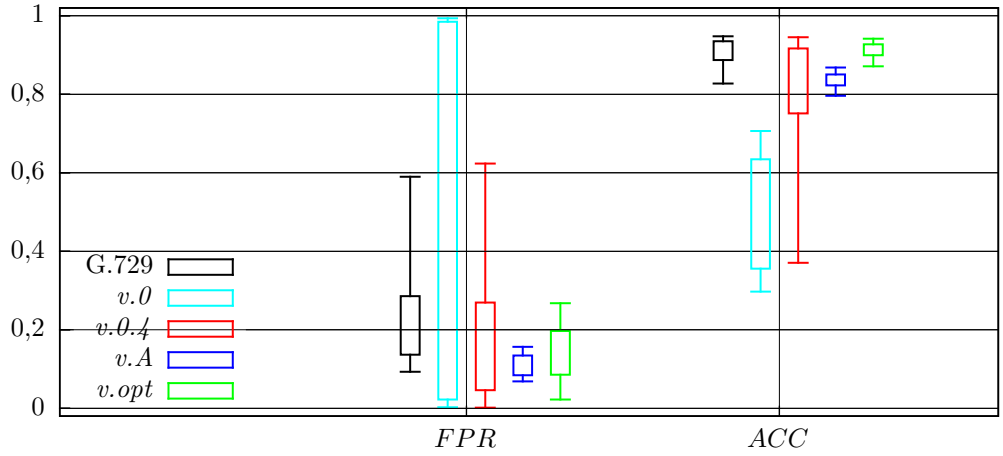
Tab. 4.1 – *Pokračování z předchozí strany*

ID	MÍRA [%]	G.729	*v.0.4*	*v.opt*
2	ACC:	76,9	92,7	92,6
	FPR:	27,4	7,6	7,4
3	ACC:	84,0	89,9	89,9
	FPR:	25,0	2,4	2,4
4	ACC:	70,8	87,8	87,8
	FPR:	39,1	10,5	10,5
5	ACC:	81,9	90,7	90,7
	FPR:	72,2	31,1	31,1
6	ACC:	91,7	94,8	79,7
	FPR:	9,5	5,9	0,0
7	ACC:	75,0	89,3	89,3
	FPR:	49,4	13,5	13,5
8	ACC:	77,4	90,3	90,3
	FPR:	29,8	12,1	12,1
9	ACC:	88,0	92,4	92,4
	FPR:	15,0	0,9	0,9
10	ACC:	90,3	90,5	90,5
	FPR:	16,6	4,6	4,6
11	ACC:	75,0	39,6	49,5
	FPR:	26,2	63,7	53,2
12	ACC:	79,0	85,7	85,7
	FPR:	64,7	23,2	23,2
13	ACC:	66,9	88,4	88,4
	FPR:	42,1	8,7	8,7
14	ACC:	63,6	94,8	94,5
	FPR:	40,0	0,0	0,0
15	ACC:	79,8	15,3	85,0
	FPR:	22,7	99,6	0,0
16	ACC:	75,4	73,2	73,2
	FPR:	49,8	0,2	0,2
17	ACC:	60,0	97,0	97,0
	FPR:	48,5	1,7	1,7
18	ACC:	82,6	82,6	82,6
	FPR:	30,2	0,2	0,2
19	ACC:	100,0	100,0	100,0
	FPR:	0,0	0,0	0,0
20	ACC:	79,7	100,0	100,0
	FPR:	20,3	0,0	0,0
21	ACC:	95,3	0,0	100,0
	FPR:	4,7	100,0	0,0
22	ACC:	89,3	100,0	100,0
	FPR:	10,7	0,0	0,0
23	ACC:	88,4	0,2	100,0

Tab. 4.1 – *Pokračování na další straně*

Tab. 4.1 – *Pokračování z předchozí strany*

ID	MÍRA [%]	G.729	*v.0.4*	*v.opt*
	FPR:	11,6	99,8	0,0
24	ACC:	65,8	32,0	100,0
	FPR:	34,2	68,0	0,0
25	ACC:	45,2	90,4	88,7
	FPR:	65,5	8,0	8,0
26	ACC:	81,5	99,4	96,5
	FPR:	19,2	0,1	0,0
27	ACC:	40,6	62,0	62,7
	FPR:	61,1	38,4	37,5
28	ACC:	46,6	28,9	100,0
	FPR:	53,4	71,1	0,0
29	ACC:	83,5	85,7	96,7
	FPR:	17,0	14,8	0,0
30	ACC:	46,7	31,3	91,9
	FPR:	77,2	99,5	6,0
31	ACC:	65,6	40,6	59,6
	FPR:	57,4	99,5	0,0
32	ACC:	6,0	5,5	94,7
	FPR:	99,2	99,8	0,0
33	ACC:	30,3	97,8	97,8
	FPR:	78,8	0,4	0,4
34	ACC:	53,4	65,0	65,0
	FPR:	46,6	35,0	35,0
35	ACC:	59,7	92,5	100,0
	FPR:	40,3	7,5	0,0
36	ACC:	36,9	16,5	16,5
	FPR:	63,1	83,5	83,5
37	ACC:	12,7	22,1	22,1
	FPR:	87,4	77,9	77,9
38	ACC:	17,3	23,2	23,2
	FPR:	82,7	76,8	76,8
39	ACC:	16,5	19,6	19,6
	FPR:	83,5	80,4	80,4
40	ACC:	53,6	3,8	100,0
	FPR:	46,4	96,2	0,0
41	ACC:	51,2	87,7	87,7
	FPR:	48,8	12,3	12,3
tot:	ACC:	72,9	69,4	85,0
	FPR:	38,1	40,6	10,9



Obr. 5.1: Srovnání výkonnosti detekčních algoritmů na čisté laboratorní databázi. *v.0* značí původní algoritmus dle [Yin+11], *v.0.4* je jeho vývojová verze po ošetření nedefinovaných stavů a po vyladění řídicích parametrů, *v.A* je experimentální varianta pracující v reálném čase a *v.opt* je optimalizovaná detekce pracující dávkově. Označeno je vždy *rozpětí* a *mezikvartilové rozpětí*.

5 SHRNUTÍ

V práci byla popsána problematika algoritmické detekce řečové aktivity v prostředí s proměnnými typy rušení a šumů.

Byly popsány nejčastěji používané míry pro hodnocení výkonnosti detektorů a by ukázán význam testovací databází pro vývoj nových detekčních algoritmů. Podrobně byla popsána úprava laboratorní databáze s objektivizovaným značením a byla představena aplikačně specifická databáze nahrávek z reálného telekomunikačního provozu, včetně popisu obtížných jevů vyskytujících se v těchto pracovních podmínkách.

Dále byly popsány existující i historické přístupy k problému automatické detekce řeči, a to jak standardizované, tak experimentální a výzkumné – založené na algoritmech strojového učení s učitelem, bez učitele i na částečně dozorovaném učení.

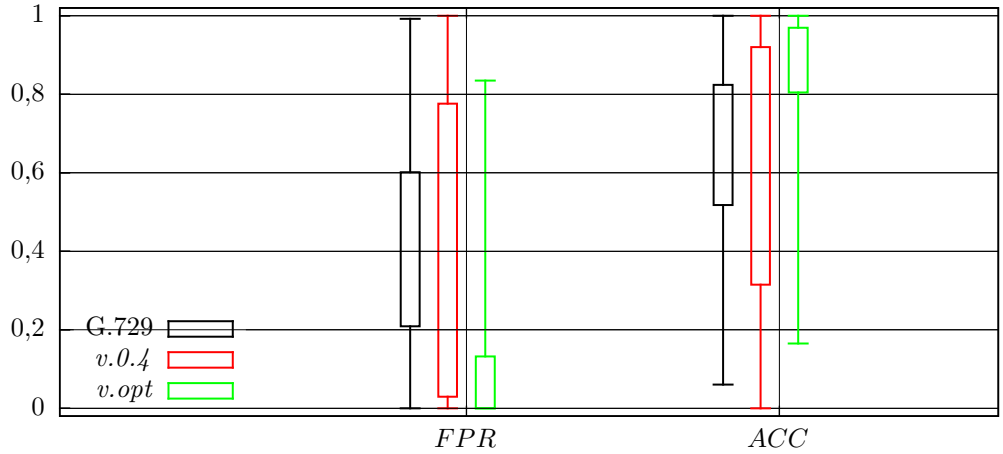
Aktuálním metodám, založeným na učení bez učitele, byla věnována největší pozornost. Jeden z možných přístupů, postavený na shlukové analýze a statistické metodě maximálně věrohodného odhadu, byl podrobně rozebrán; byl otestován v laboratorních i v reálných podmínkách, byly nalezeny algoritmické nedostatky, projevující se nižší mírou správnosti detekce a nakonec byl algoritmus optimalizován pro práci v proměnných podmínkách reálného telekomunikačního provozu.

Výsledkem optimalizace bylo zlepšení detekce v porovnání se standardizovaných algoritmem G.729B VAD

- v podmínkách simulovaných na homogenní databázi nízkošumových laboratorních nahrávek, jak je patrné ze srovnání obrázku 5.1. Zde jsou zaznamenány i významné vývojové nebo experimentální verze.
- v podmínkách reálného telekomunikačního provozu, jak je vidět na obrázku 5.2, kde jsou již pouze finální verze algoritmů.

Moderní přístup aplikace strojového učení bez učitele odstraňuje z vývojové fáze nutnost trénování detektoru na konkrétní pracovní podmínky, a tak je možno jej bez dalších úprav použít v širokém spektru aplikací a variabilních pracovních podmínek.

Optimalizovaný detektor se poměrně dobře s nepříznivými jevy v nahrávkách, jako jsou směsi



Obr. 5.2: Srovnání výkonnosti detekčních algoritmů na aplikačně specifické databázi nahrávek z telekomunikačního provozu. Označeno je vždy *rozpětí* a *mezikvartilové rozpětí*.

řeči s technickými tóny nebo s vyčkávacími signály, vč. hudby, které standardní detektory často zaměňují za řeč. I dlouhé čistě šumové nahrávky s vysokou hladinou šumu je možno s pomocí této modifikace správně klasifikovat. Standardní detektory v takových případech trpí vysokou měrou falešně pozitivních detekcí, což představuje velký problém pro následné řečové analýzy.

Detektor ve verzi *v.opt* najde uplatnění především v automatických hlasových analýzách, prováděných obvykle v dávkovém režimu. Pro aplikace pracující v reálném čase se jako perspektivní alternativa jeví varianta *v.A*, která je však příliš výpočetně náročná, takže před aplikací by ji bylo nutno nejprve algoritmicky zjednodušit.

Výkonnost algoritmu by bylo potenciálně možno dále zlepšit využitím dalších segmentálních příznaků, zavedením statistického modelu sdružených pravděpodobností nebo modelováním časových závislostí např. pomocí Markovových modelů – ovšem za cenu vyšší výpočetní náročnosti.

LITERATURA

[AR76] Bishnu S. Atal a Lawrence R. Rabiner. „A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition“. In: *IEEE Transactions on Acoustics, Speech and Signal Processing* (1976), s. 201–212.

[BF59] K. Bullington a J. M Fraser. „Engineering Aspects of TASI“. In: *The Bell System Technical Journal* (1959), s. 353–364. ISSN: 1538-7305.

[Bol86] Tim Bollerslev. „Generalized Autoregressive Conditional Heteroskedasticity“. In: *Journal of Econometrics* (1986), s. 307–327. ISSN: 0304-4076.

[CK01] Yong Duk Cho a Ahmed Kondoz. „Analysis and Improvement of a Statistical Model-Based Voice Activity Detector“. In: *IEEE Signal Processing Letters* (2001), s. 276–278. ISSN: 1070-9908.

[CKM06] Joon-Hyuk Chang, Nam Soo Kim a Sanjit K. Mitra. „Voice Activity Detection Based on Multiple Statistical Models“. In: *IEEE Transactions on Signal Processing* (2006), s. 1965–1976. ISSN: 1053-587X.

[Gór+06] Juan Manuel Górriz et al. „Hard C-means clustering for voice activity detection“. In: *Speech Communication* (2006), s. 1638–1649.

[Gór+08] Juan Manuel Górriz et al. „Jointly Gaussian PDF-based likelihood ratio test for voice activity detection“. In: *IEEE Transactions on Audio, Speech and Language Processing* (2008), s. 1565–1578. ISSN: 1558-7916.

[GZ03] Saeed Gazor a Wei Zhang. „Speech Probability Distribution“. In: *IEEE Signal Processing Letters* (2003), s. 204–207. ISSN: 1070-9908.

[IT12] ITU-T. *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP). Recommendation ITU-T G.729.* 2012.

[Kac06] J. Kacur. „The Concept of Task Specific Speech Database for VAD Systems“. In: *Proceedings ELMAR 2006.* IEEE, 2006, s. 155–158. ISBN: 9789537044039.

[Mel05] Alois Melka. *Základy experimentální psychoakustiky.* Akademie múzických umění, 2005, s. 327. ISBN: 80-733-1043-0.

[Míč+10] Ivan Míča et al. „Voice activity detection under the highly fluctuant recording conditions of call centres“. In: (2010), s. 334–336.

[Pel11] Pavel Pelikán. „Databáze nahrávek pro detekci hlasové aktivity“. Vedoucím práce byl Ing. Ivan Míča. Vysoké učení technické v Brně, 2011.

[Pre08] William H. Press. *Computational Statistics with Application to Bioinformatics: Classifier Performance: ROC, Precision-Recall, and All That.* 2008.

[Ram+05] Javier Ramírez et al. „Statistical Voice Activity Detection Using Multiple Observation Likelihood Ratio Test“. In: *IEEE Signal Processing Letters* (2005), s. 689–692. ISSN: 1070-9908.

[RS78] Lawrence R. Rabiner a Ronald W. Shafer. *Digital Processing of Speech Signals.* 1. vyd. New Jersey: Prentice-Hall, Inc., 1978. 512 s. ISBN: 0-13-213603-1.

[SKS99] Jongseo Sohn, Nam Soo Kim a Wonyong Sung. „A Statistical Model-Based Voice Activity Detection“. In: *IEEE Signal Processing Letters* (1999), s. 1–3. ISSN: 1070-9908.

[SS98] Jongseo Sohn a Wonyong Sung. „A voice activity detector employing soft decision based noise spectrum adaptation“. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. 1998, s. 365–368. ISBN: 0-7803-4428-6.

[TP09] Jiří Tatarinov a Petr Pollák. „Design and Utilization of Testing Database for VAD Classification“. In: *19th Czech-German Workshop on Speech Processing*. 2009, s. 42–47. ISBN: 978-80-86269-18-4.

[TR07] Rasool Tahmasbi a Sadegh Rezaei. „A Soft Voice Activity Detection Using GARCH Filter and Variance Gamma Distribution“. In: *IEEE Trans. on Audio, Speech and Language Processing* (2007), s. 1129–1134. ISSN: 1558-7916.

[TR08] Rasool Tahmasbi a Sadegh Rezaei. „Change Point Detection in GARCH Models for Voice Activity Detection“. In: *IEEE Transactions on Audio, Speech and Language Processing* (2008), s. 1038–1046. ISSN: 1558-7916.

[VS93] A. Varga a H. Steeneken. „Assessment for automatic speech recognition: II. NOISEX-92: A database and experiment to study the effect of additive noise on speech recognition systems.“ In: *Speech communication* (1993), s. 247–251. ISSN: 0167-6393.

[Yin+11] Dongwen Ying et al. „Voice Activity Detection Based on an Unsupervised Learning Framework“. In: *IEEE Transactions on Audio, Speech and Language Processing* (2011), s. 2624–2633.

ABSTRAKT

Tato práce se zabývá problematikou automatické detekce řečově aktivních úseků signálu. Jsou analyzovány dopady nepříznivých podmínek na spolehlivost detekce a jsou uvedeny hlavní současné i historické směry výzkumu této problematiky. Teoretický rozbor významných detekčních metod a používaných modelů je podepřen testy na laboratorní i na aplikačně specifické databázi odpovídající proměnlivým pracovním podmínkám detektorů řeči. Na základě analýzy rozebíraných algoritmů a jejich výkonnosti ve variabilních simulovaných i reálných podmínkách jsou detekční metody optimalizovány pro prostředí s proměnnými vlastnostmi.

KLÍČOVÁ SLOVA

Detekce řečové aktivity, optimalizace, proměnné pracovní podmínky

ABSTRACT

This thesis deals with the issue of algorithmic voice activity detection. Impacts of adverse conditions on the reliability of detection is analysed, and main historical and up-to-date approaches to this issue are discussed. Simulations on both synthetic, and application specific labeled speech databases are used to support the theoretical analysis of important VAD methods. Based on the theoretical analysis together with the performance results, an optimization is proposed that is capable to overcome some limitations of the current methods when dealing with variable working conditions.

KEYWORDS

Voice activity detection, Optimization, Varying environments