AKUSTICKÉ LISTY

České akustické společnosti
www.czakustika.cz

ročník 16, číslo 2–3 říjen 2010

Obsah

ČESKÁ AKUSTICKÁ SPOLEČNOST

Vážení čtenáři,
jak jste si jistě všimli, již od roku 2001 procházejí všechny odborné příspěvky v našich Akustických listech recenzním řízením. Věříme, že to pomáhá zvyšovat kvalitu našeho časopisu, a proto jsme rozhodnuti v tomto trendu pokračovat. Od letošního roku je také náš časopis zařazen na takzvaný pozitivní seznam periodik Rady vlády pro výzkum, vývoj a inovace, tedy Seznam recenzovaných neimpaktovaných periodik vydávaných v České republice. Díky tomu se články publikované v našem časopise mohou přenést do Registru informací o výsledcích výzkumu a vývoje (RIV), což potěší zejména přispěvatele z akademického prostředí. Pokud se nám podaří nastavenou úroveň udržet, třeba se dočkáme i přechodu mezi impaktované časopisy.

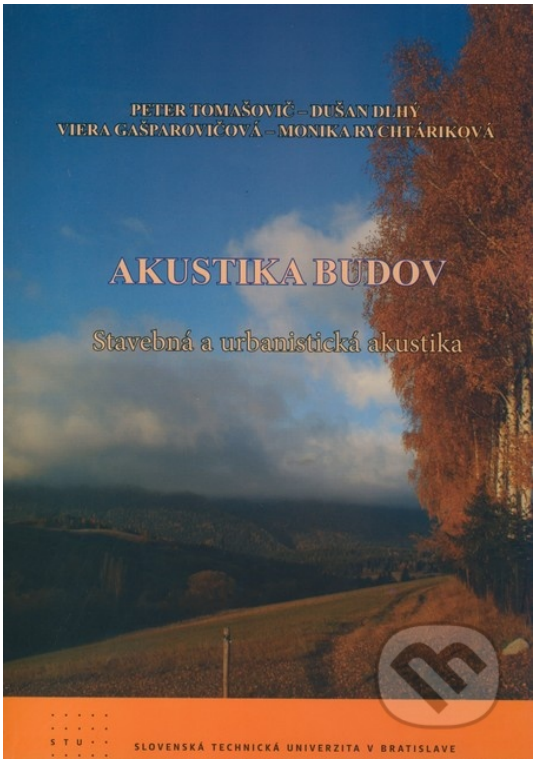
Za redakční radu

Ondřej Jiříček

Akustika budov
Stavebná a urbanistická akustika

Peter Tomašovič, Dušan Dlhý, Viera Gašparovičová, Monika Rychtáriková

Vydání: první.
Počet stran: 398.
Vazba: brožovaná.
Vydala: STU v Bratislavě.
Rok vydání: 2009.
ISBN: 978-80-227-3019-8



Na Slovensku vyšla velmi zajímavá publikace, kterou mohou využít všichni ti, kteří se zabývají akustikou budov. Publikace je dobře formálně zpracovaná. Obsahuje přehledné rejstříky věcné i bibliografické a je doprovázena množstvím obrázků, grafů a tabulek. V potřebném rozsahu autoři uvádějí nejprve základy fyzikální akustiky. Následuje seznámení s metodami hodnocení hluku ve vnějším i vnitřním prostředí. Projektanti uvítají pojednání o potřebném rozsahu dokumentace z hlediska ochrany před hlukem v jednotlivých stupních projektování.

Rozsáhlá kapitola Akustika stavebních konstrukcí charakterizuje tři základní projevy šíření zvuku v budově: zvuk šířící se vzduchem, zvuk vznikající nárazem a zvuk šířící se konstrukcí. Seznamuje s teorií vzduchové i kročejové neprůzvučnosti pro různé typy jednoduchých i násobných konstrukcí a seznamuje se základními výpočtovými a graficko-výpočtovými metodami.

V dalších částech publikace je podrobný rozbor akustických vlastností jednotlivých prvků stropních konstrukcí a podlah, akustických vlastností dveří, oken a zasklení. Velká pozornost je věnována hluku výtahů a jeho snižování, vstupním a schodišťovým prostorám i obvodovým svislým a střešním konstrukcím. Tato část obsahuje praktické návody pro návrh i posouzení různých konstrukčních typů a variant.

V kapitolách Urbanistická akustika a Predikce hladiny dopravního hluku jsou uvedeny všechny informace a nástroje potřebné pro návrh a posouzení akustiky budovy. Stejně tak tento účel splňuje kapitola Laboratorní měření vzduchové a kročejové neprůzvučnosti dělicích konstrukcí.

Publikaci autoři věnovali svému učiteli profesoru Júliu Puškášovi k 75. narozeninám.

J. Schwarz

Detekce řečové aktivity na bázi HMM a GMM modelování

Jiří Tatarinov a Petr Pollák

ČVUT–FEL, Technická 2, 166 27 Praha 6
e-mail: [tatarji1;pollak]@fel.cvut.cz

This article describes several solutions of voice activity detection which represents an important subpart of more general research in the field of speech processing and which is a subject of many contemporary research activities and many applications of speech technology. The approaches based on Gaussian mixture models and hidden Markov models are presented in this article, commonly with the study of using different speech parameterizations in GMM and HMM based VADs. Presented detectors were compared with referential heuristic algorithms based on energy and cepstral analysis, and with the VAD according to ITU-T G.729 recommendation. The testing of suggested algorithms was realized using the data from CZKCC signal database recorded in running car and the contribution of proposed statistical detectors based on GMM and HMM is evident, especially, for speech signals collected in very noisy environment.

1. Úvod

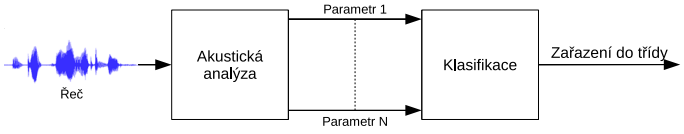
Detekce řečové aktivity představuje významnou úlohu řešenou v oblasti zpracování řeči, která je stále intenzivně rozvíjena v současném výzkumu v celosvětovém měřítku. Detektory řečové aktivity (Voice Activity Detectors – VAD) jsou využívány v mnoha aplikacích řečových technologií. Jednu z nejvýznamnějších aplikací nalezneme zejména v algoritmech potlačování šumu v řečovém signálu, kde jsou detektory řeči typicky využívány pro účely odhadu charakteristik pozadí v neřečových úsecích či v dalších souvisejících úlohách, jako je odhad SNR řečového signálu. Další aplikační oblastí jsou komunikace, kde jsou VAD detektory součástí kodeků signálu při přenosu v GSM či VoIP (Voice over Internet Protokol) sítích, snižující nutnou přenosovou kapacitu kanálu a sloužící rovněž pro odhad kódovaných parametrů řeči. V menší míře jsou detektory řeči používány i v systémech rozpoznávání řeči, nejčastěji opět v souvislosti se zpracováním řeči s aditivním šumem, kde přispívají k lepší detekci neřečových úseků. V neposlední řadě může být přímou aplikací detekce řečové aktivity automatizovaná segmentace dlouhých nahrávek na krátké dílčí promluvy, které se používají například v informačních systémech, hlášeních v různých veřejných prostorách, elektronických slovnících s hlasovým výstupem apod.

Obecně jsou algoritmy detekce řeči založeny na různých přístupech, v principu však je možné všechny modelovat dvěma základními bloky: akustickou analýzou řečového signálu řešící extrakci vhodných příznaků popisujících řečový signál a následným klasifikačním algoritmem rozlišujícím mezi řečovými a neřečovými úseky, viz obrázek 1.

Blok akustické analýzy řeči je navrhován vždy pro dané prostředí, ve kterém je řečový signál detekován, přičemž různé charakteristiky mají rozdílnou výpočetní náročnost detektoru i odolnost vůči případnému rušivému pozadí. Typickými výstupy akustické analýzy mohou být jednotlivé parametry jako výkon (energie) signálu, intenzita, počet průchodů nulou, entropie, kepstrální (spektrální) vzdálenost od pozadí, průměrná koherence apod. nebo vektory více příznaků, jako jsou LPC koeficienty, kepstrální koeficienty, či koherenční funkce.

Nejstarší a stále používané VAD jsou detektory výkonové (energetické). Popularita a časté používání těchto algoritmů je dána především velmi malou výpočetní náročností, kde přítomnost hlasového signálu je detekována na základě vyšší energie oproti úseku neřečovému [4, 12, 26]. Nevýhodou je pak pokles spolehlivosti detekce při intenzivnějším šumovém pozadí, pro nízkoenergetické neznělé úseky řeči už i pro nižší úrovně rušivého pozadí. Doplňkovou charakteristikou energetických detektorů může být počet průchodů nulou (ZCR) [2, 7, 9], jehož aplikace přináší zlepšení detekce neznělých úseků u energetických detektorů. Hodnota ZCR ovšem velmi závisí na šumovém pozadí a i zde dochází ke sbližování hodnot pro šum a pro neznělé hlásky. Někteří autoři [10, 24] využívají jako další doplňkové kritérium i detekci periodických složek řeči.

Druhou výraznější skupinu tvoří detektory založené na analýze spektrálních charakteristik řeči, aproximovaných nejčastěji pomocí kepstrálních koeficientů. Princip těchto detektorů vychází ze spektrální (kepstrální) odlišnosti řečového signálu a jeho pozadí. Kepstrální detektory řeči [3] jsou poměrně spolehlivé a hranice použitelnosti pro detekci řeči v zarušeném prostředí je výrazně nižší než u detektorů energetických.



Obrázek 1: Struktura detektorů řečové aktivity

Další používané algoritmy jsou založeny na měření entropie, která vyjadřuje míru neuspořádanosti soustavy, neboť analýzou spektra velmi zašuměné řeči bylo zjištěno, že oblasti obsahující řeč jsou více organizované než oblasti šumové. Experimenty potvrdily, že VAD používající k vý-

Přijato 3. června 2010, akceptováno 13. září 2010.

počtu entropii pracují v prostředí s nestacionárním šumem spolehlivěji než detektory čistě energetické. Přínosem je také, že tyto detektory nejsou citlivé na změny dynamiky šumu, ale reagují pouze na změny spektrální povahy [19].

V systémech s vícekanálovým řečovým signálem lze zlepšení detekce silného a nestacionárního šumu dosáhnout použitím charakteristik na bázi koherenční funkce [20], která přináší principiálně novou informaci o podobnosti korelovanosti signálů ve dvou kanálech. Vyšší počet vstupních kanálů a s tím související nároky na použitý hardware jsou ovšem často limitujícími faktory použití těchto algoritmů, zejména v případech aplikace v jednoduchých a snadno implementovatelných systémech s hlasovým vstupem.

Druhým principiálním blokem algoritmu detekce řeči je klasifikace na základě některých výše zmíněných akustických parametrů. V tichém prostředí a zejména při použití jednotlivých akustických příznaků jako je energie či kepstrální vzdálenost se užívají nejčastěji jednoduché heuristické klasifikační algoritmy na bázi adaptivního či fixního prahování a dosahují akceptovatelné přesnosti detekce. Druhou skupinu pak tvoří algoritmy vycházející z teorie rozpoznávání, které využívají klasifikační metody na bázi statistického modelování, resp. strojového učení; tj. Markovovské modely [13], neuronové sítě [8], diskriminační analýza (LDA) [11], či SVM [18].

Příkladem algoritmu, který využívá více výše diskutovaných parametrů současně a heuristické rozhodování, je detektor dle doporučení G.729, navržený a optimalizovaný pro práci s řečovým kodekem ITU-T G.729 8 kbit/s ASCELP pro přenos hovorového signálu telekomunikačním kanálem. Parametry signálu jsou zde energie z celého spektra, nízkofrekvenční energie, ZCR a LSF koeficienty (Line Spectral Frequencies). Rozhodnutí o přítomnosti řeči (primární rozhodnutí) je získáno pomocí lineárních rozhodovacích funkcí, které rozdělují vektory příznaků na řečové a šumové regiony. Sekundární rozhodnutí je zjednodušené a v případě, kdy je detekován šum, slouží k aktualizaci parametrů detektoru. Rychlé změny v rozhodování o přítomnosti řeči jsou filtrovány pomocí heuristicky zjištěných pravidel [1, 5, 6].

Tato práce se zaměřuje na detektory, které jsou založeny na statistické klasifikaci, a zejména pak na jejich přínos při detekci řeči snímané ve více zarušených prostředích. Mezi statistické klasifikátory patří Gaussovské směsové modely (GMM) [14, 23], či skryté Markovovské modely (HMM). Parametry HMM mohou být i časově proměnné a adaptované v průběhu detekce [13]. Dříve diskutované heuristické detektory na bázi energie, kepstra a detektor dle G.729 [2–4, 12, 16, 26] jsou v této práci použity jako reference pro srovnání výsledků detekce řeči na bázi GMM a HMM modelování.

2. Statistické detektory řeči

Zjednodušeně se dá říci, že statistická klasifikace je postup, kdy jsou jednotlivé prvky zařazeny do určité třídy na základě podobných statistických vlastností zařazovaného prvku a dané třídy, přičemž statistické vlastnosti třídy jsou získány z trénovací množiny, ve které je známá příslušnost prvku k dané třídě. Pro případ detekce přítomnosti řeči je nutné zjistit statistické vlastnosti řeči a šumového pozadí, které jsou vyjádřeny prostřednictvím parametrů a struktury modelů řeči a šumu. Detekce řeči je pak výstupem zhodnocení rozdílů mezi modely řeči, šumu a vyhodnocované promluvy.

Většina tradičních VAD algoritmů předpokládá, že je šum stacionární v delších úsecích, než je tomu u řeči. Tento předpoklad umožňuje vystihnout charakteristiky měnícího se šumu, a to i v případě občasného výskytu řeči [21]. Většinou je šum ovšem nestacionární, proto se v čase mění jeho statistické parametry. Pro modelování nestacionárních procesů je možné použít skryté Markovovy modely. Stacionární šum lze modelovat pomocí HMM, které obsahují pouze jeden stav, nebo pomocí GMM. Nestacionární šum vystihuje lépe vícestavový HMM, kde jsou změny charakteristik šumového signálu modelovány konečným počtem stacionárních stavů [25].

2.1. Detektor na bázi Gaussovských směsí – GMM VAD

Základem pro použití GMM je Gaussovské rozložení vektoru parametrů signálu a dále předpoklad, že prvky stejných tříd mohou mít některé podobné statistické vlastnosti. Principem detekce řeči pomocí GMM je modelování řeči a šumu pomocí akustického modelu, který je tvořen dvěma různými GMM, které umožňují klasifikovat množinu vektoru parametrů. Jednotlivé třídy $l \in \{ticho, řeč\}$ jsou pak charakterizovány pomocí M-složkových směsí hustot normálních rozdělení $b_l(\mathbf{o}_t)$, pro každý vektor parametrů $\mathbf{o}_t$, délky n, v čase t tedy platí

$$b_l(\mathbf{o}_t) = \sum_{m=1}^{M} c_{lm} \mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}_{lm}; \boldsymbol{\Sigma}_{lm}), \tag{1}$$

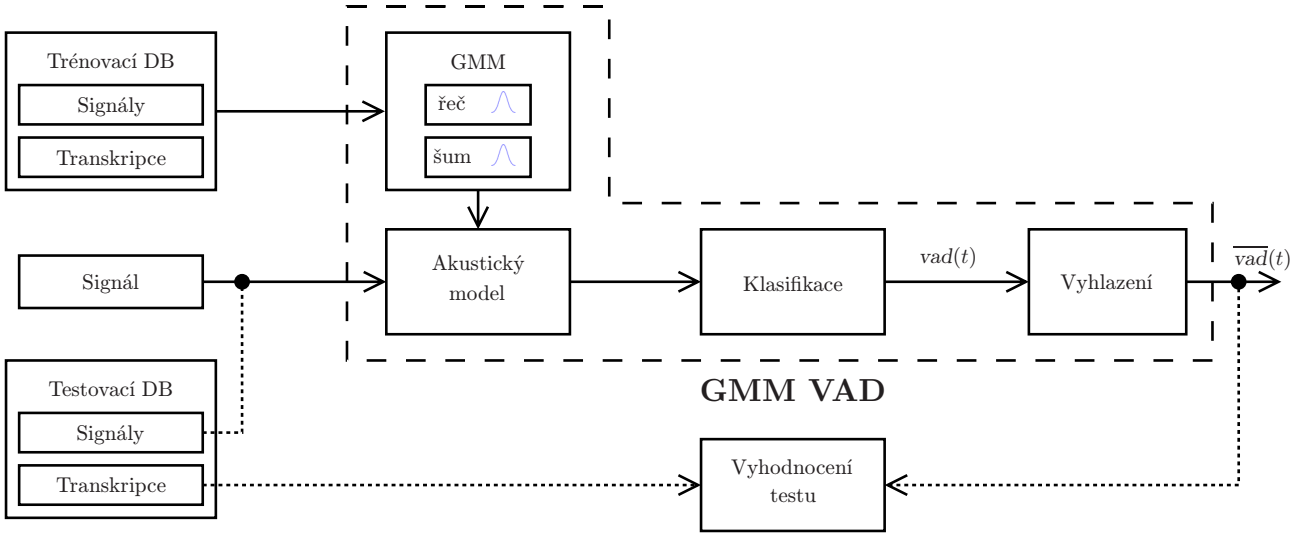
kde $\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}; \boldsymbol{\Sigma})$ je známá hustotní funkce normálního rozložení, tj.

$$\mathcal{N}(\mathbf{o}_t; \boldsymbol{\mu}; \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \mathrm{e}^{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{o}_t - \boldsymbol{\mu})}, \tag{2}$$

přičemž c_m jsou váhové koeficienty směsi, $\boldsymbol{\mu}$ je střední hodnota a $\boldsymbol{\Sigma}$ je diagonální kovarianční matice. Známe-li pravděpodobnostní rozložení řeči $b_\mathrm{s}(\mathbf{o}_t)$ a šumu $b_\mathrm{n}(\mathbf{o}_t)$, je klasifikace poměrně jednoduchá, pozorovaný vektor je přiřazen do třídy, ve které se vyskytuje s větší pravděpodobností. Výstupem klasifikace je vektor $vad(t)$ obsahující informaci o přítomnosti, nebo absenci řeči, tj.

$$vad(t) = \begin{cases} 1 & \text{pro } b_\mathrm{s}(\mathbf{o}_t) \geq b_\mathrm{n}(\mathbf{o}_t), \\ 0 & \text{pro } b_\mathrm{s}(\mathbf{o}_t) < b_\mathrm{n}(\mathbf{o}_t). \end{cases} \tag{3}$$

Parametry každého GMM, tj. přesné hodnoty parametrů hustotních funkcí řeči $b_\mathrm{s}(\mathbf{o}_t)$ a šumu $b_\mathrm{n}(\mathbf{o}_t)$, se nastaví



Obrázek 2: Algoritmus GMM VAD

v rámci trénovacího procesu tak, aby nejlépe vystihovaly danou třídu signálu. Pokud se prvky třídy shlukují kolem center, je vhodné modelovat pravděpodobnostní rozložení pomocí více směsí. Optimální počet směsí není znám a je typickým předmětem optimalizace nastavení detektoru. Počet směsí by měl odpovídat počtu složek, ze kterých je signál složen.

Pro trénovací účely je nutné mít k dispozici množinu signálů, u kterých je již dopředu známá klasifikace v jednotlivých časových okamžicích. Vhodným postupem trénování, který je použit i v našem řešení, je použití Baum-Welchova algoritmu (konkrétně implementovaného v sadě HTK nástrojů), který lze považovat za implementaci EM algoritmu.

Jelikož uvedený algoritmus pracuje na bázi krátkodobé analýzy a klasifikace bez kontextu, výstupní detekce obsahuje velké množství krátkých a chybných zákmitů. Ty je možné odstranit pomocí vyhlazení na bázi mediánové filtrace, což je i poslední krok tohoto algoritmu. V našem řešení pracujeme s mediánovým filtrem desátého řádu.

Jednotlivé kroky resp. funkční bloky trénování a detekce algoritmu na bázi GMM jsou znázorněny v přehledovém schématu na obrázku 2.

2.2. Detektory řeči se skrytými Markovovými modely – HMM VAD

Detektory řeči na bázi HMM v principu rozšiřují výše diskutované GMM detektory o použití kontextové informace mezi analyzovanými segmenty. Skryté Markovovské modely (HMM) jsou statistické stavové automaty, které kromě jediné statistiky u GMM modelují klasifikované třídy pomocí více stavů a pravděpodobnostních přechodů mezi nimi. Jednotlivé stavy modelů jsou pak charakterizovány opět směsí hustotních funkcí normálního rozdělení podle vztahu (1). Modelovanými částmi signálu jsou v tomto případě fonémy, slabiky nebo slova.

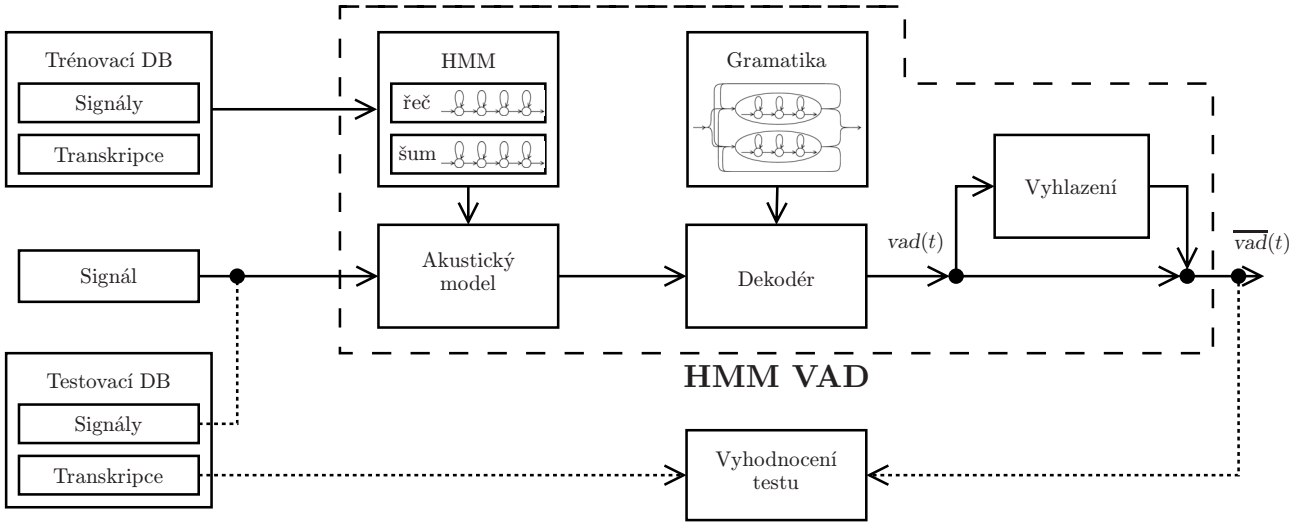
V detektorech řeči na bázi HMM se tedy předpokládá, že signál je tvořen posloupností řečových a neřečových úseků a k modelování jednotlivých úseků promluv jsou použity v nejjednodušším případě pouze 2 modely, model řeči λ_s a model šumu λ_n. Jedná se tedy v principu o zjednodušený rozpoznávač řeči pouze se dvěma různými modely oproti typickému modelování 30–40 fonémů.

Označme tedy $W = \{w_1, w_2, \ldots, w_N\}$ jako posloupnost úseků řeči w_s a šumu w_n, kterou se snažíme rozpoznat na základě známé posloupnosti vektorů příznaků analyzovaného řečového signálu $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$. Principem detekce řeči je potom nalezení takové posloupnosti úseků řeči a šumu $\hat{W}$, která maximalizuje pravděpodobnost $P(W|\mathbf{O})$, tedy pravděpodobnost posloupnosti úseků řeči a šumu při pozorované posloupnosti vektorů. Pomocí Bayesova pravidla je možné odvodit, že výpočet pravděpodobnosti $P(\hat{W}|\mathbf{O})$ je ekvivalentní výpočtu $P(\mathbf{O}|\lambda_W)$, tj. pravděpodobnosti, že výstupní posloupnost $\mathbf{O}$ byla generována akustickým modelem λ_W vytvořeným pomocí spojení modelů řeči λ_s a šumu λ_n [17]. Pro posloupnost $\hat{W}$ tedy platí

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|\mathbf{O}) \approx \underset{W}{\operatorname{argmax}}\, P(\mathbf{O}|\lambda_W)\,. \qquad (4)$$

Výpočet pravděpodobnosti $P(\mathbf{O}|\lambda_W)$ je pak transformován na úlohu nalezení nejpravděpodobnější cesty skrze rozpoznávací síť, která je získána pomocí token passing algoritmu, který je hojně využíván při rozpoznávání. Výstupem tohoto algoritmu je nejen nejpravděpodobnější posloupnost úseků řeči a šumu $\hat{W}$, ale v případě detektoru řeči i posloupnosti začátků $T_\mathrm{b}(\hat{W})$ a konců $T_\mathrm{e}(\hat{W})$ těchto úseků, tj.

$$\hat{W} = \{\hat{w}_1, \hat{w}_2, \ldots, \hat{w}_N\}, \qquad (5)$$
$$T_\mathrm{b}(\hat{W}) = \{t_\mathrm{b}(\hat{w}_1), t_\mathrm{b}(\hat{w}_2), \ldots, t_\mathrm{b}(\hat{w}_N)\}, \qquad (6)$$
$$T_\mathrm{e}(\hat{W}) = \{t_\mathrm{e}(\hat{w}_1), t_\mathrm{e}(\hat{w}_2), \ldots, t_\mathrm{e}(\hat{w}_N)\}. \qquad (7)$$



Obrázek 3: Algoritmus HMM VAD

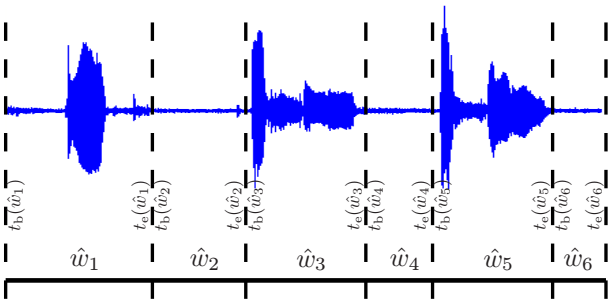
Význam těchto posloupností je ilustrován na obrázku 4. Jejich znalost pak vede k získání posloupnosti $vad(t)$, která v čase t klasifikuje pozorovanou promluvu jako řeč w_s nebo šum w_n, tj.

$$vad(t) = \begin{cases} 1 & \text{pro } t_\mathrm{b}(\hat{w}_i) < t \leq t_\mathrm{e}(\hat{w}_i)|_{w_i=w_\mathrm{s}}, \\ 0 & \text{pro } t_\mathrm{b}(\hat{w}_i) < t \leq t_\mathrm{e}(\hat{w}_i)|_{w_i=w_\mathrm{n}}, \end{cases} \quad (8)$$

přičemž $\hat{w}_i$ je označení rozpoznaného i-tého úseku časově ohraničeného okamžiky $t_\mathrm{b}(\hat{w}_i)$ a $t_\mathrm{e}(\hat{w}_i)$, kde $1 \leq i < N$.

Výstup dekodéru $vad(t)$ v tomto bodě neobsahuje velké množství krátkých a chybných zákmitů v důsledku průchodu vícestavovým HMM modelem, i přesto je vhodné toto menší množství zákmitů vyhladit. V navrhovaném algoritmu byl použit opět mediánový filtr desátého řádu. Optimální počet směsí a stavů HMM modelů směsi a šumu je předmětem optimalizovaného nastavení. Jednotlivé bloky algoritmu jsou opět znázorněny v blokovém schématu 3.

Při trénování modelů řeči a pauzy je nutné zajistit správnou inicializaci relativně obecných modelů, aby statisticky odpovídaly částem signálu obsahujícím šum a řeč, tj. v první fázi je nutné inicializační trénování s přesným přiřazením řečových a neřečových úseků jednotlivým částem signálu. Tímto způsobem jsou odhadnuty první přibližné hodnoty parametrů HMM. Konečný odhad získáme iteračním trénováním tzv. Baum-Welchovým algoritmem. Nastavení inicializačních modelů ze známých úseků je velmi důležité, neboť konvergence k lokálnímu maximu kriteriální funkce na bázi EM algoritmu je významně ovlivněna právě inicializační fází, která musí být co nejpřesnější.



Obrázek 4: Ilustrace hranic řečových úseků u HMM VAD

3. Implementace

Implementace zmiňovaných detektorů řečové aktivity byla provedena v jazyce C/C++. Byla zvolena modulární struktura tak, aby bylo umožněno snadné doplňování vytvořeného kódu o další moduly. Implementace v jazyce C/C++ byla zvolena v první řadě s ohledem na větší rychlost při běhu programů, ale i vzhledem k portování na jiné platformy, například ARM XScale. Trénování parametrů HMM i GMM modelů bylo realizováno pomocí sady nástrojů HTK Toolkit [27], vytvořených pro vývoj rozpoznávačů řeči založených na skrytých Markovových modelech. V sadě jsou obsaženy programy pro předzpracování řečových signálů, trénování, stavbu Markovovských modelů, rozpoznávání řeči a jiné nástroje. HTK Toolkit byl také využit i pro extrakci příznaků a pro dekódování řeči pomocí token passing algoritmu.

Referenční algoritmy, tj. energetický a kepstrální detektor, byly napsány rovněž v jazyce C/C++. Implementace VAD detektorů podle G.729 a G.729 A.III jsou součástí přílohy B a dodatku III doporučení ITU-T [5]. Všechny výše diskutované algoritmy jsou součástí balíčku, který je dostupný na webových stránkách [22].

4. Experimenty

Popisované navrhované a referenční detektory řeči byly otestovány na reálných řečových signálech snímaných

v prostředí automobilu. Byl analyzován vliv použité parametrizace a nastavení jednotlivých parametrů detektorů na přesnost detekce řeči a na základě uvedené analýzy byly stanoveny optimální hodnoty parametrů použitých detektorů.

4.1. Kritéria

Prezentované detektory řeči byly hodnoceny pomocí objektivních kritérií, jimiž lze objektivně posoudit typ a velikost chyby detekce přítomnosti řeči. Byla použita následující kritéria [20], jejichž význam je ilustrativně vysvětlen na obrázku 5:

ERR (*Error Decision Rate*) – celková relativní četnost chyby klasifikace,

ERS (*Error Decision in Speech*) – relativní četnost chyby detekce řečových segmentů,

ERN (*Error Decision in Noise*) – relativní četnost chyby detekce šumových segmentů.

Rozšířené varianty uvedených kritérií pak jsou:

SDN (*Speech Detected as Noise*) – relativní četnost chyby uprostřed řečového segmentu,

NDS (*Noise Detected as Speech*) – relativní četnost chyby uprostřed šumového segmentu,

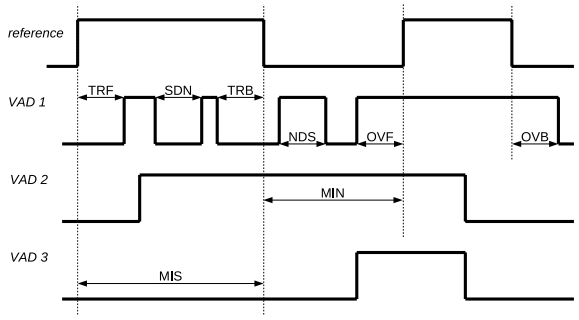
OVF (*Overlap at the Front*), ***OVB*** (*Overlap at the Back*), ***TRF*** (*Truncation at the Front*), ***TRB*** (*Truncation at the Back*) – četnosti chyb na přechodech mezi šumem a řečí (oříznutí/přesah na začátku/konci řečového úseku).

4.2. Množiny signálů

Signály pro experimenty byly použity z databáze snímané v prostředí automobilu, viz [15]. Promluvy jsou nahrávané za různých podmínek v automobilu, tj. v tichém prostředí, ve stojícím automobilu se zapnutým motorem a nakonec především v jedoucím automobilu.

Vyhodnocení výsledků detekce diskutovaných algoritmů bylo provedeno na množině ručně olabelovaných signálů. Tato testovací množina měla dvě základní části, podmnožinu obsahující nahrávky z tichého prostředí (129 signálů o celkové délce 13 min 3 s) a druhou podmnožinu se signály z prostředí jedoucího automobilu (celkem 36 signálů o celkové délce 6 min 43 s). Signály obsahovaly typicky čtyři promluvy oddělené delšími pauzami.

Pro trénovací množinu, která již byla výrazně větší, nebylo možné manuálně vytvořit přesné hranice začátků a konců řečových a neřečových úseků. Tyto hranice je ovšem nutné znát, neboť automatická uniformní segmentace takto obecných modelů nevede k jejich dobrému natrénování. Poměrně přesné hranice je však možné získat pomocí automatické fonetické segmentace na bázi zarovnání natrénovaných HMM modelů fonémového rozpoznávače řeči. Pomocí tohoto postupu byla vytvořena trénovací množina signál zahrnující 5 048 promluv o celkové délce 4 h 28 min 32 s, která byla použita pro trénování v realizovaných experimentech a která nalézá uplatnění i v jiných úlohách.



Obrázek 5: Ilustrace významu kritérií

4.3. Volba příznaků pro GMM a HMM detektory

Jako příznaky popisující řečový signál byly pro GMM a HMM detektory použity různé parametrizace většinou používané při rozpoznávání řeči vždy s délkou okénka 32 ms a krokem 10 ms

- MFCC – 12 mel-kepstrálních koeficientů a energie, spolu s dynamickými a akceleračními koeficienty,
- RASPLP – 21 spektrálních percepčně lineárně prediktivních koeficientů RASTA,
- RACPLP – 13 kepstrálních percepčně lineárně prediktivních koeficientů RASTA zahrnujících nultý koeficient,
- SPLP – 21 spektrálních percepčně lineárně prediktivních koeficientů,
- CPLP – 13 kepstrálních percepčně lineárně prediktivních koeficientů zahrnujících nultý koeficient,
- DCTC – 13 kepstrálních koeficientů vypočtených pomocí diskrétní kosinovy transformace (DCT) zahrnujících nultý keoficient,
- LPC – 13 kepstrálních koeficientů vypočtených pomocí lineární predikce zahrnujících nultý koeficient,
- LPA – 12 koeficientů lineární predikce,
- F0ZCRE – 1 koeficient základního hlasivkového tónu, 1 koeficient počtu průchodů nulou a 1 koeficient energie.

4.4. Optimalizace počtu směsí GMM VAD

První experiment byl zaměřen na analýzu optimálního počtu směsí v GMM detektoru. Nejprve byl zvyšován počet směsí od 0 do 64, stejně u modelu šumu i řeči. Zjištěné hodnoty chyb pro všechny použité parametrizace jsou v tabulkách 1 a 2. Je vidět, že zvyšování počtu použitých směsí má vliv na výslednou chybovost detektorů. Dalo se očekávat, že použití většího počtu směsí povede k přesnějšímu modelování pravděpodobnostních rozložení, a tím ke snížení celkové chybovosti ERR. Z výsledků ovšem vyplývá,

Mix	MFCC		RACPLP		RASPLP		SPLP		CPLP		DCTC		LPC		LPA		F0ZCRE	
	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)
0	12,4 ±7,9	4,0 ±3,2	12,1 ±7,2	1,9 ±2,0	15,3 ±6,2	10,2 ±4,9	27,0 ±12,2	26,2 ±12,4	14,0 ±10,5	5,9 ±4,5	22,9 ±14,7	8,0 ±5,7	21,9 ±14,6	7,7 ±5,7	24,8 ±10,8	23,5 ±10,9	19,2 ±13,3	14,5 ±11,2
2	11,9 ±6,9	6,8 ±4,1	13,8 ±6,6	5,8 ±3,1	21,6 ±9,0	11,3 ±5,1	21,9 ±11,9	18,6 ±10,0	12,2 ±8,3	7,3 ±5,0	17,1 ±11,2	8,2 ±5,5	21,7 ±15,1	6,8 ±4,8	23,1 ±13,3	16,6 ±9,1	16,9 ±10,8	12,6 ±8,4
16	11,3 ±7,4	6,8 ±4,4	13,9 ±8,1	3,2 ±2,3	18,1 ±8,8	9,3 ±4,5	20,9 ±12,8	12,2 ±8,4	12,5 ±9,8	6,5 ±5,0	16,7 ±13,5	9,6 ±9,4	14,2 ±11,5	7,7 ±5,4	28,8 ±18,2	11,9 ±7,3	26,4 ±13,7	19,3 ±12,2
32	11,2 ±7,3	6,4 ±4,3	13,0 ±7,4	4,2 ±2,7	16,1 ±7,8	8,8 ±4,2	26,2 ±16,5	10,8 ±8,0	13,6 ±12,3	5,2 ±4,3	16,8 ±14,1	8,2 ±7,8	14,7 ±12,1	6,7 ±4,6	25,9 ±15,0	11,7 ±7,1	23,1 ±13,3	19,3 ±11,9
64	11,5 ±7,4	7,4 ±4,7	12,7 ±7,2	5,1 ±3,1	14,1 ±7,7	6,8 ±3,7	26,8 ±16,9	8,9 ±6,2	12,1 ±9,5	6,2 ±4,6	17,1 ±11,6	7,8 ±6,5	16,2 ±12,9	5,6 ±4,2	29,5 ±10,7	10,7 ±6,8	20,9 ±12,6	18,7 ±11,4

Tabulka 1: Zvyšování počtu směsí u GMM VAD v tichém prostředí

Mix	MFCC		RACPLP		RASPLP		SPLP		CPLP		DCTC		LPC		LPA		F0ZCRE	
	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)
0	12,9 ±5,2	9,9 ±4,7	10,1 ±3,8	8,8 ±4,3	15,2 ±4,8	14,5 ±5,0	14,5 ±9,2	9,3 ±6,0	15,9 ±14,4	6,4 ±3,8	20,4 ±18,1	8,7 ±4,8	18,4 ±17,9	7,1 ±4,4	22,1 ±12,5	13,8 ±6,3	15,0 ±11,5	7,3 ±3,9
2	12,2 ±7,4	7,7 ±4,3	11,9 ±4,0	10,6 ±4,4	12,5 ±5,2	8,6 ±4,4	32,6 ±26,9	7,3 ±4,2	25,8 ±17,9	5,0 ±4,0	14,7 ±12,1	11,0 ±4,9	12,2 ±12,7	8,7 ±4,1	19,4 ±10,1	12,9 ±6,5	15,3 ±11,7	7,6 ±4,1
16	12,3 ±10,2	5,1 ±3,5	9,5 ±4,2	6,9 ±3,7	10,2 ±5,0	6,5 ±3,1	27,4 ±23,7	5,7 ±4,5	20,3 ±20,4	4,6 ±3,3	14,9 ±15,0	7,7 ±4,4	13,2 ±14,9	7,3 ±4,1	20,3 ±15,1	11,7 ±6,0	12,4 ±8,3	8,5 ±3,9
32	11,0 ±6,7	5,8 ±3,3	9,9 ±3,7	8,3 ±4,0	9,5 ±3,8	7,2 ±3,5	20,9 ±19,7	6,2 ±4,2	18,2 ±18,9	5,3 ±3,4	14,0 ±14,0	7,7 ±4,5	12,6 ±14,3	7,3 ±4,1	19,3 ±14,6	11,5 ±5,6	12,8 ±9,1	8,4 ±3,9
64	12,2 ±10,0	5,7 ±3,5	10,0 ±3,8	8,4 ±4,0	10,0 ±4,0	7,4 ±3,8	18,4 ±19,8	7,0 ±4,0	21,5 ±21,0	4,5 ±3,3	14,7 ±13,6	8,5 ±4,9	10,1 ±6,5	7,7 ±3,9	19,7 ±15,0	11,7 ±6,0	13,5 ±10,3	8,1 ±3,8

Tabulka 2: Zvyšování počtu směsí u GMM VAD v prostředí jedoucího automobilu

že takový trend je vidět jenom u některých parametrizací a pouze při navyšování počtu směsí v určitých mezích. Většina parametrizací žádný takový trend nevykazuje a někdy vede navýšení počtu směsí dokonce ke zvýšení chyby ERR. Takové chování je možné vysvětlit tím, že vrcholy vícesměsových rozložení jsou většinou velmi blízko u sebe, a proto se ani zásadně nemění jejich tvar. Nicméně lze nalézt optimální počet směsí, který vede k nejlepším výsledkům detekce přítomnosti řeči v obou prostředích. Nejlepších výsledků bylo dosaženo využitím kepstrálních a spektrálních koeficientů RASTA a mel-kepstrálních koeficientů. Jako optimální nastavení se jeví nepoužívat žádné směsi pro parametrizaci RACPLP, naopak použít 64 směsí pro parametrizaci RASPLP a 32 směsí pro parametrizaci MFCC. Tato nastavení byla následně použita při experimentech srovnávajících prezentované a referenční detektory řečové aktivity.

4.5. Optimalizace počtu stavů a směsí HMM VAD

Druhý experiment byl zaměřen na sledování vlivu počtu stavů modelů na chybovost VAD na bázi HMM. Byly použity vždy modely řeči a šumu se stejným počtem stavů a s emitujícími funkcemi se 16 směsmi, a to jak pro tiché prostředí, tak pro prostředí jedoucího automobilu. Počet stavů modelů byl zvyšován od 3 do 9 stavů. Výsledky těchto experimentů jsou v tabulkách 3 a 4. U většiny parametrizací je vidět, že zvětšování počtu stavů HMM vede ke snížení chyby ERR, a to především díky snižování ERS, tj. chyb v řečových segmentech. Přičemž tato závislost se projevuje jak v tichém prostředí, tak v prostředí jedoucího automobilu.

Ve třetím experimentu byl analyzován vliv počtu směsí na chybovost VAD na bázi HMM. Před začátkem experimentů byl i v tomto případě jednoznačně očekáván výsledek, že se zvyšujícím se počtem směsí se bude snižovat chybovost VAD. Tato závislost se nepotvrdila stejně jako při experimentech s GMM VAD. Určitý trend lze najít pouze pro některé parametrizace, a to především v prostředí jedoucího automobilu.

Závěrem těchto experimentů je zjištění optimálních nastavení HMM VAD pro nejperspektivnější parametrizace – MFCC, RACPLP a RASPLP. Nejúspěšnější byly VAD, které používaly nejdelší 9stavové modely, a to se 2 směsmi pro MFCC, 32 směsmi pro RACPLP a 64 směsmi pro RASPLP.

Počet stavů	MFCC		RACPLP		RASPLP		SPLP		CPLP		DCTC		LPC		LPA		F0ZCRE	
	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)
3	10,9 ±7,7	6,2 ±4,3	13,6 ±8,4	2,3 ±2,3	18,0 ±9,1	8,9 ±4,6	21,6 ±13,6	11,7 ±8,2	12,4 ±10,1	6,3 ±4,6	16,1 ±13,2	10,1 ±9,9	13,3 ±10,6	8,0 ±5,4	30,8 ±19,1	10,9 ±7,3	18,4 ±12,8	13,3 ±10,5
5	9,0 ±7,9	4,8 ±3,8	12,7 ±7,5	6,9 ±3,7	13,2 ±7,9	6,7 ±4,2	24,5 ±16,8	9,8 ±7,5	12,0 ±10,2	6,6 ±4,9	16,4 ±12,9	10,6 ±8,1	13,8 ±12,2	6,7 ±5,1	20,6 ±12,5	12,8 ±8,1	19,4 ±13,6	14,8 ±11,0
7	9,6 ±8,7	4,7 ±3,6	12,2 ±8,2	4,3 ±3,4	10,9 ±6,9	4,7 ±3,6	15,0 ±11,3	10,4 ±7,0	12,7 ±10,5	6,2 ±4,7	15,7 ±12,9	9,8 ±8,8	13,6 ±12,5	6,7 ±5,4	19,4 ±12,0	10,2 ±6,8	18,9 ±13,6	14,2 ±10,9
9	9,4 ±7,3	5,3 ±4,3	11,2 ±6,9	3,8 ±3,1	10,5 ±6,5	3,7 ±3,4	14,7 ±10,9	10,4 ±6,7	13,5 ±12,1	6,1 ±5,2	16,4 ±13,8	10,4 ±10,3	12,8 ±10,7	6,9 ±5,6	18,9 ±12,7	9,0 ±6,8	19,0 ±14,3	13,3 ±11,6

Tabulka 3: Zvyšování počtu stavů HMM v tichém prostředí

Počet stavů	MFCC		RACPLP		RASPLP		SPLP		CPLP		DCTC		LPC		LPA		F0ZCRE	
	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)	ERR (%)	ERS (%)
3	11,8 ±10,1	4,9 ±3,4	8,6 ±4,1	5,8 ±3,2	10,2 ±5,0	6,6 ±3,2	28,8 ±23,9	6,0 ±4,4	16,2 ±18,0	5,4 ±3,5	12,3 ±11,0	7,9 ±4,4	10,9 ±9,0	7,9 ±4,0	19,0 ±14,7	11,8 ±6,0	14,3 ±11,9	7,3 ±4,1
5	10,2 ±10,7	5,2 ±3,8	8,5 ±3,8	7,1 ±3,8	9,5 ±4,2	7,0 ±3,4	21,1 ±26,0	6,0 ±4,4	15,9 ±21,6	5,3 ±3,9	18,7 ±22,7	6,1 ±4,5	16,3 ±21,6	5,8 ±4,0	18,4 ±14,6	12,1 ±6,0	13,1 ±10,3	7,2 ±4,0
7	9,9 ±12,8	4,7 ±3,8	7,4 ±4,1	5,5 ±3,6	8,0 ±4,0	6,8 ±3,9	20,6 ±24,2	5,3 ±4,3	15,7 ±22,2	4,8 ±4,2	24,2 ±26,3	4,4 ±4,4	18,3 ±23,3	5,1 ±4,2	18,6 ±15,4	11,0 ±5,7	13,1 ±10,2	6,9 ±3,8
9	9,2 ±7,7	4,9 ±3,9	7,3 ±3,8	5,9 ±3,8	7,3 ±3,4	6,0 ±3,2	21,0 ±24,3	4,6 ±4,1	15,1 ±20,4	4,4 ±3,9	22,3 ±24,6	5,7 ±5,1	20,5 ±23,9	4,7 ±4,0	17,9 ±15,8	10,3 ±5,8	13,3 ±11,3	6,1 ±3,6

Tabulka 4: Zvyšování počtu stavů HMM v prostředí jedoucího automobilu

4.6. Srovnání s referenčními algoritmy

Ve čtvrtém experimentu byly detektory na bázi GMM a HMM s optimalizovaným nastavením srovnány s nejčastěji používanými heuristickými detektory. U prezentovaných statistických detektorů byly vybrány pouze parametrizace vedoucí k nejlepším výsledkům – MFCC, RACPLP a RASPLP. Výsledky pro tiché prostředí jsou prezentovány v tabulce 5 a pro prostředí jedoucího automobilu v tabulce 6. Jsou zde vypočteny střední hodnoty a standardní odchylky.

Nejlepší výsledky dosáhl HMM detektor, a to jak pro tiché prostředí, tak pro prostředí jedoucího automobilu. V tichém prostředí jsou nejlepší výsledky dosaženy s využitím mel-kepstrálních koeficientů a v prostředí jedoucího automobilu se nejlépe osvědčily spektrální koeficienty PLP RASTA. Také GMM VAD využívající mel-kepstrální koeficienty dosahuje lepších výsledků než všechny referenční detektory. Referenční detektory chybují méně v řeči než v šumu, což je výborná vlastnost pro skutečné aplikace. U detektoru G729 bylo pozorováno časté selhání v průběhu klasifikace úvodních částí signálu, což je známý problém popsaný v [5], který by měly řešit úpravy popsané v dodatku III. Z experimentálních výsledků se ukázalo, že detektor G729 podle dodatku III sice detekuje šum lépe než G729 podle přílohy B, ale pouze v prostředí jedoucího automobilu. Nevýhodou prezentovaných detektorů je jejich větší složitost a také závislost na trénovací množině. Výborné výsledky HMM a GMM detektorů jsou převážně dosaženy pomocí výrazného snížení chybné detekce šumu. Všechny referenční detektory dosahují horší výsledky v prostředí jedoucího automobilu. Velmi překvapivé je, že prezentované detektory dosahují nepatrně lepší výsledky v prostředí jedoucího automobilu než v tichém prostředí.

5. Závěr

V článku byly popsány detektory řečové aktivity založené na statistickém modelování na bázi skrytých Markovových modelů a Gaussovských směsových modelů. Byly provedeny analýzy přesnosti detekce pro různé podmínky v prostředí automobilu – tiché prostředí a jedoucí automobil. Součástí analýzy bylo také srovnání s referenčními detektory, tj. s detektorem energetickým, kepstrálním a detektory specifikovanými v příloze B a dodatku III doporučení G.729. Nejdůležitější závěry jsou shrnuty v následujících bodech:

- Detektory využívající HMM a GMM dosahují lepších výsledků než referenční detektory energetické, kepstrální, či detektory dle doporučení G.729. Lepší výsledky jsou dosaženy při detekci neřečových úseků, což je zvláště patrné při srovnání s detektory podle ITU-T G729.

VAD	ERR (%)	ERS (%)	ERN (%)	SDN (%)	NDS (%)
HMM MFCC	8,9 ±7,0	4,1 ±3,8	4,8 ±5,4	0,8 ±2,2	0,4 ±1,7
HMM RASTASPECT	10,2 ±6,9	3,8 ±3,4	6,3 ±5,9	0,4 ±1,6	1,5 ±3,9
HMM RASTACEPST	11,0 ±7,1	3,5 ±3,0	7,6 ±7,0	0,2 ±0,9	1,8 ±4,7
GMM MFCC	11,2 ±7,3	6,4 ±4,3	4,7 ±5,5	2,1 ±2,8	1,3 ±3,2
GMM RASTACEPST	12,1 ±7,2	1,9 ±2,0	10,1 ±7,0	1,2 ±1,7	2,1 ±3,9
KEPSTRÁLNÍ	12,1 ±11,0	6,3 ±3,8	5,8 ±10,2	4,1 ±3,3	4,0 ±8,6
ENERGETICKÝ	13,0 ±10,1	7,4 ±4,6	5,6 ±8,9	4,9 ±4,1	3,7 ±7,1
GMM RASTASPECT	14,1 ±7,7	6,8 ±3,7	7,3 ±7,1	4,1 ±2,8	3,3 ±5,6
G729 A.III	30,7 ±11,6	1,1 ±1,5	29,6 ±11,6	0,8 ±1,2	8,3 ±10,3
G729	31,1 ±11,0	3,2 ±2,7	28,0 ±11,4	2,5 ±2,4	11,2 ±11,4

Tabulka 5: Tiché prostředí

VAD	ERR (%)	ERS (%)	ERN (%)	SDN (%)	NDS (%)
HMM RASTASPECT	6,7 ±3,6	5,6 ±3,6	1,1 ±2,2	0,1 ±0,3	0,5 ±1,2
HMM RASTACEPST	7,3 ±3,8	5,8 ±3,7	1,5 ±2,8	0,5 ±0,8	0,6 ±1,4
HMM MFCC	8,3 ±6,8	5,0 ±4,0	3,3 ±7,0	0,4 ±0,7	0,7 ±2,3
GMM RASTASPECT	10,0 ±4,0	7,4 ±3,8	2,6 ±3,7	3,1 ±2,1	1,6 ±2,4
GMM RASTACEPST	10,1 ±3,8	8,8 ±4,3	1,3 ±2,3	3,4 ±2,1	0,5 ±1,4
GMM MFCC	11,0 ±6,7	5,8 ±3,3	5,2 ±7,4	1,6 ±1,4	4,1 ±6,2
KEPSTRÁLNÍ	24,3 ±11,4	6,6 ±3,4	17,7 ±12,3	4,9 ±2,5	16,5 ±11,3
ENERGETICKÝ	25,0 ±11,9	7,5 ±3,6	17,6 ±13,5	5,3 ±2,8	15,0 ±9,0
G729 A.III	25,0 ±6,5	2,3 ±2,1	22,7 ±6,4	0,7 ±0,8	5,5 ±6,0
G729	32,4 ±12,5	4,6 ±2,3	27,8 ±13,2	2,6 ±1,8	12,0 ±10,0

Tabulka 6: Prostředí jedoucího automobilu

- HMM detektor pak dosahuje lepších výsledků než detektor na bázi GMM, což je dáno přesnějším modelováním signálu pomocí více stavů. Výhodou GMM VAD je na druhou stranu snadnější trénování i vlastní aplikace.
- Nevýhodou prezentovaných statistických detektorů je nutnost trénování, a tím i závislost na trénovací množině signálů. Obecná spolehlivost nezávislá na prostředí není zcela snadno dosažitelná.
- Jako nejvhodnější parametry signálů se ukázaly spektrální percepčně lineární prediktivní koeficienty RASTA.
- Největší přínos prezentovaných detektorů spočívá ve zlepšení klasifikace signálů se silnějším šumovým pozadím.

Poděkování

Tento výzkum byl podporován granty GAČR 102/08/H008 „Analýza a modelování biologických a řečových signálů“, GAČR 102/08/0707 „Rozpoznávání mluvené řeči v reálných podmínkách“ a výzkumným záměrem MSM 6840770014 „Výzkum perspektivních informačních a komunikačních technologií“. Databáze CZKCC vznikla v rámci společného projektu a za finanční podpory firmy TEMIC TELEFUNKEN GmbH se sídlem v Ulmu v roce 2001. Databáze není veřejně dostupná a jejím vlastníkem je v současné době Harman/Becker, Ulm, Německo.

Reference

[1] A. Benyassine, E. Shlimot, H. Su. A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communication Magazine*, 1997.

[2] A. Ganapathiraju, L. Webster, J. Trimble, K. Bush, P. Kornman. Comparison of energy-based endpoint detectors for speech signal processing. *Southeastcon '96, Proceedings of the IEEE*, pp. 500–503, 1996.

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

[4] W. Harrison, J. Lim, E. Singer. A new application of adaptive noise cancellation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1):21–27, 1986.

[5] International Telecommunication Union – Telecommunication Standartization Sector. *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, 1996.

[6] International Telecommunication Union – Telecommunication Standartization Sector. *A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to ITU-T V.70*, 1996.

[7] J. C. Junqua, B. Reaves, B. Mark. A study of endpoint detection algorithms in adverse conditions: Incidence on a DTW and HMM recognize. *Eurospeech*, pp. 1371–1374, 1991.

[8] J. Kačur, G. Rozinaj, S. Herrera-Garcia. Speech signal detection in a noisy environment using neural networks and cepstral matrices. *Electrical Engineering*, 55(5-6):131–137, 2004.

[9] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4), 1989.

[10] I. Lee, H. Stern, S. Mahmoud. A voice activity detection algorithm for communication system with dynamically varying background acoustic noise. *IEEE Vehicular Technology Conference*, 2(1214–1218), 1998.

[11] A. Martin, D. Charlet, L. Mauuary. Robust speech/non-speech detection using LDA applied to MFCC. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1:237–240, 2001.

[12] A. Martin, L. Karray, A. Gilloire. High order statistics for robust speech/non-speech detection. *EUSIPCO 2000*, (10):469–472, 2000.

[13] H. Othman, T. Abounasr. A semi-continuous state transition probability HMM-based voice activity detection. *Acoustics, Speech, and Signal Processing*, 5(17–21):821–824, 2004.

[14] R. Padmanabha, P. S. Krishnan, H. A. Murthy. A pattern recognition approach to VAD using modified group delay. *NCC*, 2008.

[15] P. Pollák. 300 speaker Czech database from car. Final report of the project based on Frame Agreement for the collection of Speech data Corpora. Technical report, CTU FEL, Temic Germany, 2001.

[16] P. Pollák, P. Sovka, J. Uhlíř. Cepstral speech/pause detectors. *Proceedings of IEEE Workshop on Nonlinear Signal and Image Processing*, 1995.

[17] J. Psutka. *Komunikace s počítačem mluvenou řečí.* Academia, 1995.

[18] J. Ramirez, P. Yelamos, J. M. Gorriz, J. C. Segura. SVM-based speech endpoint detection using contextual speech features. *Electronics Letters*, 42(7):426–428, 2006.

[19] P. Renevey, A. Drygajlo. Entropy based voice activity detection in very noisy condition. *EUROSPEECH'01*, pp. 1887–1890, 2001.

[20] J. Rosca, R. Balan, N. P. Fan, C. Beaugeant, V. Gilg. Multichannel voice detection in adverse environments. *EUSIPCO 2002*, 2002.

[21] J. Sohn, N. S. Kim, W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1), 1999.

[22] J. Tatarinov. VAD Toolkit. `http://noel.feld.cvut.cz/speechlab`, 2009.

[23] J. Trmal, J. Zelinka, J. Psutka, L. Müller. Comparison between GMM and decision graphs based silence/speech detection method. *Proceedings of the 11th international conference Speech and computer SPECOM'2006*, pp. 376–379, 2006.

[24] R. Tucker. Voice activity detection using a periodicity measure. *IEE Proceedings, Communications, Speech and Vision*, 139(4), 1992.

[25] S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reduction.* John Wiley and Sons, New York, Prentice-Hall, Englewood Cliffs, New Jersey, 2000.

[26] K-H. Woo, T-Y. Yang, K-J. Park, C. Lee. Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 36(2), 2000.

[27] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, V. Valtchev, P. Woodland. *The HTK Book (for HTK Version 3.1).* Cambridge University Engineering Department, UK, 2001.

Akustické listy: ročník 16, číslo 2–3 říjen 2010 ISSN: 1212-4702
Vydavatel: Česká akustická společnost, Technická 2, 166 27 Praha 6 Vytisklo: Nakladatelství ČVUT, výroba
Počet stran: 16 Počet výtisků: 200
Redakční rada: M. Brothánek, O. Jiříček, J. Kozák, R. Čmejla, J. Volín
Jazyková úprava: R. Svobodová

Uzávěrka příštího čísla Akustických listů je 30. listopadu 2010.