Řečové detektory využívající ergodické Markovovské modely

J. Tatarinov, P. Pollák

České vysoké učení technické v Praze
Fakulta elektrotechnická

Abstrakt

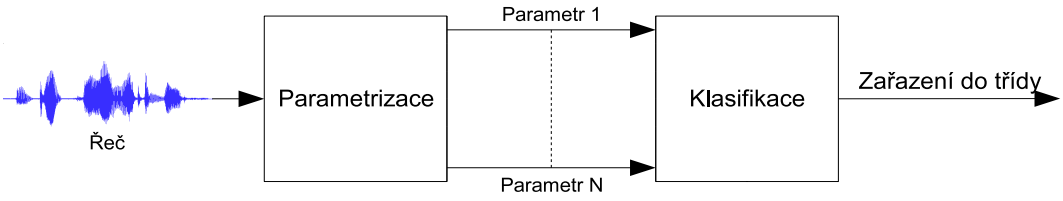
Tento článek prezentuje využití ergodických Markovovských modelů při detekci řečové aktivity. Tradiční detektory řeší tuto úlohu pomocí klasifikátorů založených na prahování vhodných řečových charakteristik. V prezentovaném článku je použit přístup založený na statistickém modelování. Byl navržen klasifikátor a na jeho základě byl sestrojen detektor řečové aktivity. Detektor byly otestován a zhodnocen na signálech z databáze CAR2CS. Detektor využívající ergodické skryté Markovovy modely dosahuje lepších výsledků než tradiční detektory. Největší přínos prezentovaných detektorů spočívá ve zlepšení klasifikace silně zarušených signálu.

1 Úvod

Detekce řečové aktivity hraje důležitou roli v oblasti zpracování řeči a je objektem současného výzkumu. Detektory řečové aktivity (Voice Activity Detector - VAD) jsou využívány nejen v mnoha různých oblastech vědy, ale i v průmyslových aplikacích. Detektory řečové aktivity jsou používány v průběhu rozpoznávání řeči a pro odhadování SNR (Signal to Noise Ratio) nebo v algoritmech zvýrazňování řeči. Jiné aplikace můžou být v oblasti komunikací, například VoIP (Voice over Internet Protokol), kde VAD slouží pro snížení nutné přenosové kapacity. To je dosaženo nepřenášením paketů obsahující ticho. Bylo také ukázáno, že VAD můžou přispět k zvětšení přesnosti rozpoznávání řeči.

2 Principy detekce řečové aktivity

Detektor řečové aktivity je algoritmus, který klasifikuje čistý řečový signál, či směs řeči a šumu do dvou tříd - řeči a šumu. Řečový signál je rozdělen do segmentů a ke každému je segmentu je přidělena příslušná třída. Často jsou řeč a šum označovaný jako “1” a “0”. Výstupem je pak “1”, je-li v i-tém segmentu přítomná řeč a “0” pro segment bez řečové aktivity. Detekci řečové aktivity můžeme provádět i na úrovni vzorků řečového signálu. Průběh detekce řeči je téměř vždy rozdělen do dvou částí. V první jsou ze segmentů řečového signálu získaný parametry řeči, které zdůrazňují rozdíl mezi řečí a šumem. Ve druhé fázi probíhá samotná klasifikace.



Obrázek 1: Struktura detektorů řečové aktivity

2.1 Tradiční detektory

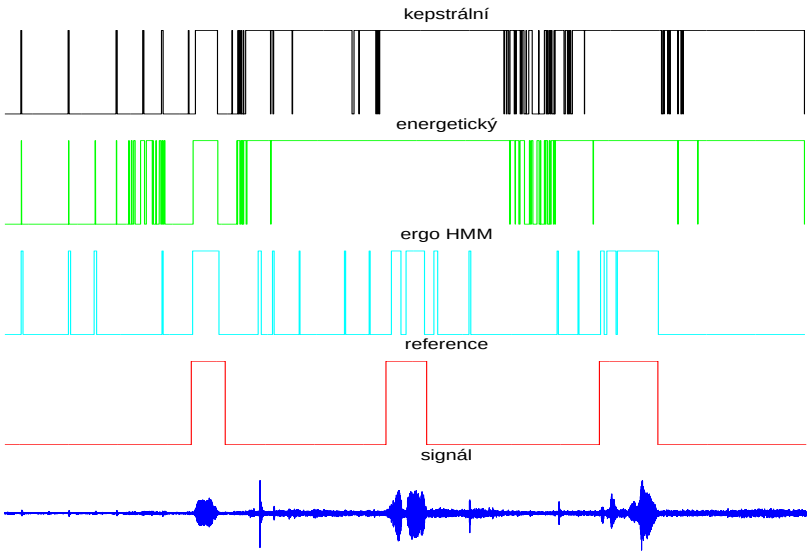
Tradiční detektory řeší tuto úlohu pomocí klasifikátorů založených na prahování vhodných řečových charakteristik. Detektory využívající prahování rozhodují o absenci nebo přítomnosti řeči na základě srovnávání hodnot parametrů segmentů s prahovou hodnotou, tj. prahem. Principiálně můžeme rozlišit dva zakladní druhy prahování - první způsob spočívá ve využití statické

prahu a druhý způsob využívá práh dynamický. Mezi nejznámější typy pracující na tomto principu jsou detektory kepstrální [2],[6] a energetické [1],[10].

2.2 Nevýhody tradičních detektorů

V případě, kdy je použit statický výpočet prahu je nevýhoda tohoto způsobu detekce řeči přímočará, při významnější změně prostředí začíná takto postavený detektor selhávat. Tento problém se snaží řešit druhý přístup využívající dynamické nastavování prahu. Výhodou je, že jsou tyto algoritmy ustálené a dlouhodobě spolehlivé. Další výhodou tohoto postupu je, že uživatel má plnou kontrolu nad rozhodováním, protože rozhoduje nad pravidly, která slouží k vypočtu hodnoty prahu. Naopak nevýhodou je, že nacházení prahu může být velmi složité, a to zvláště v případech, kdy je k dispozici rozsáhlá množina parametrů signálu. Možnosti zlepšení algoritmů detekce řeči je v hledání nových postupů pro získávání aktuální prahové hodnoty nebo využití jiných algoritmů, které prahování vůbec nepoužívají. Další výhodou by bylo využití algoritmu, který jsou schopny nastavit parametry detektorů řeči tak, že jsou z určitého hlediska optimální. Tyto body splňuje využití statistických klasifikačních algoritmů, mezi které patří například skryté Markovovy modely (HMM) [7],[4].

Na obrázku 2 vidíme výstupy diskutovaných tradičních detektorů, tj kepstrálního a energetického současně s výstupem detektoru na bazi ergodických HMM. Jde o signál s vysokou úrovní šumu v jedoucím automobile. Je vidět, že kepstrální i energetický detektor chybně detekují hlučné pozadí jako řeč. Pravděpodobně selhalo nastavování prahové hodnoty. Naopak detektor na bázi ergodických HMM si v uvedeném případě vedl o něco lépe.



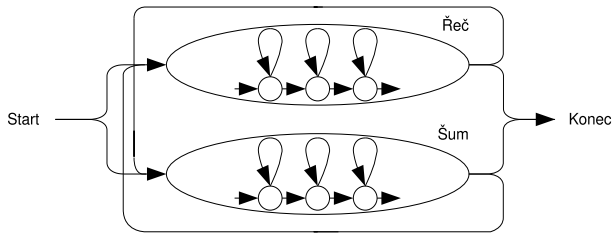
Obrázek 2: Příklad detekce řečového segmentu

2.3 Detektory na bazi HMM

Skryté Markovové modely (HMM) jsou statistické stavové automaty, které modelují prvky řeči pomocí stavů a přechodů mezi nimi. Stavy modelů jsou charakterizovány pravděpodobnostním rozložením parametrů určité části signálu, která je daným stavem modelována. Částmi signálu rozumíme například fonémy, slabiky nebo slova. V prezentovaném článku je zkoumaná pouze jedna skupina Markovovský modelů, tzv. ergodické modely. Tyto modely jsou specifické tím, že umožňují přechody mezi libovolnými stavy. Tato struktura je velmi vhodná pro modelování nestacionárních šumových složek signálu [8].

2.4 Rozpoznávací síť

Hlavní výhodou Markovovských modelů je jejich schopnost se zřetězovat dohromady, čímž se modelují větší celky řeči, v ideálním případě dokonce řeč samotná. Způsob, jakým je řečový signál modelován je předmětem současného výzkumu. Mezi jeho úkoly patří volba vhodného tvaru a vlastností modelů, či způsob jakým jsou jednotlivé modely spojeny. Způsob propojení jednotlivých modelů je vyjádřen pomocí jazykového model. Výsledná rozpoznávací síť je získána kombinací tvaru jazykového modelu a struktury použitých skrytých Markovovských modelů. Pro úlohu detekce řeči je řečový signál modelován tak, že část stavů odpovídá řeči a jiná část odpovídá šumu. Jazykový model je velmi jednoduchý a umožňuje libovolné přechody mezi modely řeči a šumu, můžeme tedy pozorovat tyto sekvence: řeč - šum, šum - řeč, šum - šum a řeč - řeč. Z hlediska stavů pozorujeme libovolné přechody mezi stavy uvnitř jednotlivých modelů. Vše si je možno představit na základě následujícího obrázku 3.



Obrázek 3: Gramatika rozpoznávání

2.5 Inicializace a trénování modelů

Zásadní vlastností detektoru řeči je nejen tvar rozpoznávací sítě, ale i číselné hodnoty jejich parametrů. Pro jejich získání je nutné inicializovat a natrénovat jednotlivé modely sítě. V oboru zpracování řeči, jsou nejčastěji používány levo-pravé (LP) modely. K jejich inicializaci jsou používané standardní metody založené na shlukování. K jejich inicializaci se předpokládá, že můžeme přiřadit úseky signály jednotlivým stavům v pořadí z leva do prava. U ergodických modelů tento postup pochopitelně selhává, protože jsou umožněny i jiné přechody než ve směru z leva do prava. Je tedy nutné navrhnout alternativní postup inicializace modelů. Navržený alternativní způsob spočívá s využitím LP modelů a jejich následnou transformací na modely ergodické. Alternativní postup inicializace modelů by se dal shrnout v následujících bodech

- Navržení jazykového modelu složeného z LP modelů.
- Inicializace těchto modelů a jejich natrénování.
- Pro model řeči - spojení řečových složek jazykové modelu do jednoho velkého modelu.
- Pro model šumu - spojení neřečových složek jazykové modelu do jednoho velkého modelu.
- Přidání chybějících přechodů tak, aby vznikly větší ergodické modely šumu a řeči.

Jsou-li získány inicializované ergodické modely šumu a řeči je možné pokračovat v jejich dalším trénování pomocí Baum-Welchova algoritmu. Tento algoritmus již nemá problém s ergodičností skrytých Markovovských modelů.

2.6 Algoritmus detekce řeči

Princip detekce spočívá v najití nejpravděpodobnější cesty skrze rozpoznávací síť. Tato cesta je nalezena pomoci token passing algoritmu. Najití této cesty vede k získání posloupnosti sestávající s modelů řeči a šumu. Je tedy možno říci ke každému úseku detekované řeči, kterým modelem byl tento úsek nejpravděpodobněji generován. Zaznamenáním času přechodů mezi modelem řeči a šumu získáme požadovaný vystup detektoru řečové aktivity.

3 Experimenty

S výše prezentovaným detektorem byla provedená řada experimentů. Pro tyto experimenty byla použita databáze CAR2CZ [5]. V této databázi se vyskytují signály nahrané v různém prostředí vyskytující se během jízdy automobilu - tiché prostředí, stojící automobil se zapnutým motorem a jedoucí automobil. Všechny experimenty vyžadovaly úvodní natrénování HMM. K tomu byla použita trénovací množina skládající se z 9259 promluv o celkové délce 8,69 hodin.

Porovnání výsledků prezentovaného detektoru bylo uděláno s ručně olabelovanými signály. Pro testování byla použita množina obsahující 195 signálů o průměrné délce 8,39 s. Celkem bylo použito 27,25 minut testovacích dat. Každý signál obsahoval nejčastěji 4 promluvy v jednom ze tří prostředí - prostředí jedoucího automobilu, prostředí stojícího automobilu se zapnutým motorem a tiché prostředí. Testovací a trénovací množiny byly odlišné.

V testovací i trénovací množině byla použity mel-kepstrální koeficienty - MFCC_EDA, tj. 12 mel-kepstrálních koeficientů včetně energie, 12 akceleračních koeficientů a 12 delta koeficientů.

Výsledky prezentovaných detektorů byly porovnány s výsledky kepstrálního detektoru. Pro porovnání detektorů byly zvoleny tyto parametry: *správně detekovaná řeč* - $P(A/S)$, *správně detekované ticho* - $P(A/N)$, *rozlišení řeč/ticho* definované

$$P(A) = P(A/S)P(S) + P(A/N)P(N) \tag{1}$$

a *správná detekce* definována

$$P(B) = P(A/S)P(A/N), \tag{2}$$

kde $P(S)$ a $P(N)$ jsou poměr počtů segmentů řeči a ticha k celkovému počtu segmentů.

3.1 Výsledky experimentů

Výsledky experimentů jsou prezentovány v tabulce 1, kde jsou vypočteny střední hodnoty a standardní odchylky z výsledků dosažených na testovacích množinách. Výsledky jsou vždy shrnuty pro každé zkoumané prostředí, které je vyznačeno v levém sloupci. HMM detektor dosáhl velmi dobré výsledky převážně ve správné detekce šumu. V této kategorii překonává výsledky všech ostatních referenčních detektorů. Platí zde závislot, že čím je prostředí hlučnější, tím jsou dosahování výsledky ergodického HMM detektoru lepší, tedy nejznatelnější rozdíl je pozorován během jízdy automobily. Naopak srovnatelné výsledky jsou dosahovány ve správné detekci řeči a to ve všech zkoumaných prostředích. Porovnáme-li dosažené výsledky z hlediska celkové úspěšnosti, tj. správné detekce, je vidět, že detektor využívající ergodické Markovovské modely překonává kepstrální i energetický detektor.

4 Závěr

V článku byl prezentován detektor řečové aktivity založený na statistickém modelování - skrytých Markovových modelech. Zkoumána byla pouze jedna skupina Markovovský modelů, tzv. ergodické modely. Detektor byl testován a zhodnocen na signálech z databáze CAR2CS. Výsledky jsou zhodnoceny pro různé prostředí vyskytující se během jízdy automobilu - tiché prostředí, stojící automobil se zapnutým motorem a jedoucí automobil. Výsledku prezentovaného detektoru byly srovnány s energetickým a kepstrálním detektorem. Nejdůležitější závěry jsou shrnuty v následujících bodech

- Detektory využívající ergodické skryté Markovovy modely dosahují lepších výsledků než tradiční detektory.
- Při srovnání s levo-pravými modely je též výhodou menší závislost na využité datové množině.

Prostředí	VAD	$P(B)$		$P(A)$		$P(A\|S)$		$P(A\|N)$	
		μ	σ	μ	σ	μ	σ	μ	σ
jízda	Energetický	0.531	0.114	0.750	0.119	0.697	0.098	0.775	0.098
jízda	Kepstrální	0.566	0.116	0.757	0.114	0.737	0.080	0.772	0.080
jízda	Ergo HMM	0.601	0.117	0.885	0.065	0.625	0.118	0.966	0.118
zapnutý motor	Energetický	0.594	0.147	0.796	0.127	0.725	0.111	0.825	0.111
zapnutý motor	Kepstrální	0.678	0.128	0.845	0.102	0.784	0.075	0.863	0.075
zapnutý motor	Ergo HMM	0.688	0.096	0.909	0.036	0.709	0.093	0.970	0.093
ticho	Kepstrální	0.646	0.110	0.908	0.038	0.654	0.108	0.987	0.108
ticho	Energetický	0.745	0.161	0.870	0.101	0.832	0.088	0.893	0.088
ticho	Ergo HMM	0.765	0.159	0.884	0.096	0.842	0.088	0.906	0.088
všechny	Energetický	0.680	0.180	0.830	0.129	0.796	0.105	0.852	0.105
všechny	Kepstrální	0.691	0.186	0.823	0.149	0.823	0.082	0.840	0.082
všechny	Ergo HMM	0.720	0.161	0.888	0.085	0.779	0.135	0.927	0.135

Tabulka 1: Zhodnocení podle prostředí

- Největší přínos prezentovaného detektoru spočívá ve zlepšení klasifikace silně zarušených signálu.

5 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".

Reference

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

[2] J. A. Haigh and 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.

[3] Q. Li, J. Zheng, A. Tsai, and Q. Zhou. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transaction on Speech and Audio Processing*, 10(3):146–157, 2002.

[4] B. McKinley and G. H. Whipple. Model based speech pause detection. *IEEE Transaction on Speech and Audio Processing*, 1997.

[5] P. Pollák, P. Sovka, V. Hanžl, and J. Vopička. CAR2 - Czech Database of Car Speech. *Radioengineering*, 8(4):1–6, 1999.

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

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

[8] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Transaction on Speech and Audio Processing*, 8(4), 2000.

[9] Damjan Vlaj, Bojan Kotnik, Bogomir Horvat, and Zdravko Kačič. A computationally efficient Mel-Bank VAD algorithm for distributed speech recognition systems. *EURASIP Journal on Applied Signal Processing*, 2005(1):487–497, 2005.

[10] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park, and Chungyong Lee. Robust voice activity detection algorithm for estimating noise spectrum. *Electronics Letters*, 36(2), 200.

Jiří Tatarinov
České vysoké učení technické v Praze
Fakulta elektrotechnická
Technická 2
166 27 Praha 6
jiri.tatarinov@atlas.cz

Petr Pollák
České vysoké učení technické v Praze
Fakulta elektrotechnická
Technická 2
166 27 Praha 6
pollak@fel.cvut.cz