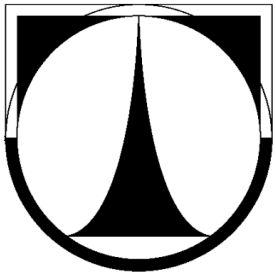
TECHNICKÁ UNIVERZITA V LIBERCI

Fakulta mechatroniky a mezioborových inženýrských studií



METODY DETEKCE ZMĚNY MLUVČÍHO V AKUSTICKÉM SIGNÁLU

DISERTAČNÍ PRÁCE

2005 JINDŘICH ŽĎÁNSKÝ

METODY DETEKCE ZMĚNY MLUVČÍHO V AKUSTICKÉM SIGNÁLU

AUTOREFERÁT DISERTAČNÍ PRÁCE

Disertant:	Jindřich Žďánský
Studijní program:	2612V Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Pracoviště:	Katedra elektroniky a zpracování signálů
	Fakulta mechatroniky a mezioborových inženýrských studií
	Technická univerzita v Liberci
	Hálkova 6, 461 17, Liberec
Školitel:	Prof. Ing. Jan Nouza, CSc.

ROZSAH DISERTAČNÍ PRÁCE A PŘÍLOH:

Počet stran:	103
Počet obrázků:	21
Počet tabulek:	25
Počet vzorců:	145
Počet příloh:	2

 listopad 2005

Anotace

Disertační práce je zaměřena na metody detekce změny mluvčího v akustickém signálu. Autor se v práci zabývá teoretickými východisky a formuluje úlohu detekce jednoho bodu změny jako testování hypotéz změny parametrů gaussovského procesu. Z rozboru problematiky vyplývá, že běžně používaný přístup k testování jednoho bodu změny na základě Bayesovského informačního kritéria není zcela v souladu s principy testování hypotéz a ukazuje, že teoreticky více ospravedlnitelné postupy vedou k lepším výsledkům. Při aplikaci teorie jednoho bodu změny na problém detekce více bodů změny se vymezuje vůči nejběžnějšímu přístupu - metodě fixních oken - a navrhuje algoritmus metody binárního dělení, dobře známý z jiných oblastí change-point analýzy, jako základní algoritmus detekce změny mluvčího. Jako alternativu nabízí vylepšení on-line metody s adaptivním oknem a zcela původní algoritmus přímé multiple change-point analýzy - metodu globální maximalizace BIC. Kapitoly 1 a 2 jsou věnovány úvodnímu slovu a krátce popisují hlavní motivaci pro tuto práci, jíž je media mining systém vyvíjený na Technické univerzitě v Liberci. Kapitola 3 se věnuje teoretickému rozboru úlohy detekce změny řečníka a popisu stávajících i nově navržených metod vhodných k jejímu řešení. Kapitola 4 shrnuje základní údaje o databázích užitých v této práci a popisuje metodiky vyhodnocení obdržených výsledků. Kapitoly 5 až 7 se zabývají návrhem implementace, možnostmi trénování a vyhodnocením metody binárního dělení, metody globální maximalizace Bayesovského informačního kritéria a metody s adaptivním oknem. Kapitola 8 hodnotí výsledky dosažené v této práci a porovnává dílčí metody z hlediska jejich praktického využití.

Annotation

The dissertation thesis is focused on the issue of speaker change detection in acoustic signals. The author interprets the single change-point problem in terms of the hypothesis testing theory. After the theoretical analysis he shows that the common approach used to solve the single change-point detection problem via the Bayesian information criterion (BIC) is not always in accord with principles of hypothesis testing. It is shown that the methods based on proper theoretical assumptions provide better results. To solve the multiple change-point problem, the author analyses all known methods and proposes several own ones. First of all he discusses the most popular method that is based on a fixed window length scenario and demonstrates its weak points (namely many free parameters). In order to overcome them, he proposes a binary segmentation technique (well-known from other branches of change-point studies) as a fundamental approach to the speaker change detection problem. Furthermore, he suggests a modification of the method that uses an adaptive window length scenario in order to provide an on-line solution. Finally he proposes a novel approach named *Global BIC Maximization* which can be characterized as an attempt to solve the multiple change-point problem globally, i.e. not via a sequential application of single change-point detection. Chapter 2 briefly describes the main motivation for this work, which is the developement of a media mining system providing automatic transcription of broadcast news. Chapter 3 deals with the theoretical analysis of the speaker change detection task and outlines principles of existing and newly proposed methods. Chapter 4 describes databases that were created and utilized for evaluation tests. Chapters 5–7 are devoted to the theory, implementation, training and evaluation of the proposed methods: *binary segmentation*, *global BIC maximization* and *a modified version of the adaptive window length method*. Chapter 8 concludes the thesis and compares the individual methods from the practical point of view.

Obsah

ÚVOD

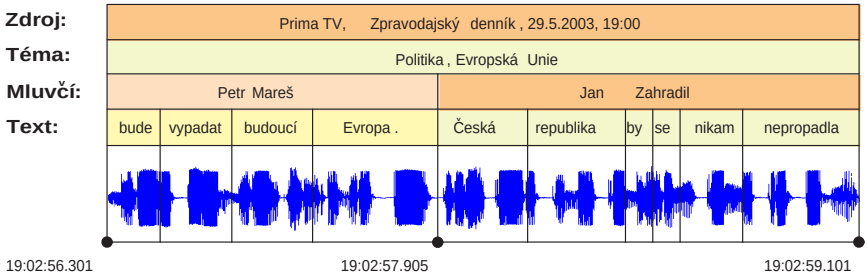
Na přelomu 20. a 21. století žijeme v období tzv. *informační společnosti*. Jejími hlavními rysy jsou převaha práce s informacemi, interaktivita, integrace a globalizační tendence. Z technologického hlediska lze informační společnost považovat za společnost s *vysokou mírou* využívání informačních a komunikačních technologií založených na prostředcích výpočetní techniky. V souvislosti s obrovským množstvím informací šířených elektronickými médii, z nichž jen nepatrná část je pro konkrétního jedince relevantní, vznikají prostředky pro jejich rychlé a přesné zpracování. V současné době existuje značné množství nástrojů umožňujících efektivní třídění a vyhledávání potřebných informací, takřka bezvýhradně ovšem vyžadují zdroje v *textové podobě*. Dychtivost současné společnosti po rychlém a snadném přístupu k informačnímu obsahu aktuálních zpráv poněkud komplikuje fakt, že jedním z nejvýznamnějších prostředků vzájemné komunikace a výměny informací je *lidský hlas*. Převod lidské řeči do textové podoby se jeví být nejpřirozenějším postupem jak využít rozvinutých technologií zpracování textových informací k automatickému vytěžování informačního obsahu hlasových nahrávek.

Jedním z nejožehavějších témat současného výzkumu v oblasti rozpoznávání řeči je *automatický přepis* zvukových nahrávek různých televizních či rádiových pořadů, jako jsou politické debaty či zpravodajství. Tento přepis slouží jako základní materiál pro automatickou *indexaci* multimediálních archivů. Komplexní systémy umožňující vytěžování informací z těchto multimediálních archivů se nazývají *media mining systémy*. Pro nejrozšířenější světové jazyky (angličtina, japonština, francouzština a němčina) jsou intenzivně vyvíjeny již několik let. Obdobný media mining systém pro češtinu je vyvíjen i na Technické univerzitě v Liberci.

1.1 Detekce změny řečníka v media mining systémech

Vzhledem k obrovskému množství informací šířených masmédii, existuje velké množství společností, které se zabývají monitorováním a archivací moderních elektronických médií. Jejich úkolem je každodenní sběr co největšího množství např. zpravodajských pořadů, ať už v podobě textů nebo ve formě audio, audiovizuálních či obecně multimediálních záznamů, a jejich zařazení do databázových systémů. Získat tyto multimediální záznamy je z technického hlediska velmi snadné, ovšem jakmile jsou jednou zařazeny do archivu, je velmi obtížné se v nich orientovat a vyhledávat, což je hlavní příčinou, proč je nutné u každého pořadu provést důkladnou *transkripci*. Pod pojmem transkripce rozumíme víceúrovňový přepis multimediálního záznamu (s ohledem na typ zpracovávaných dat). Následné vytvoření rejstříku jednotlivých popisků se nazývá *indexace*. Na obrázku 1.1 je znázorněn typický příklad transkripce zpravodajského pořadu, kde předmětem zájmu je především text, mluvčí, téma, zdroj, datum a čas záznamu.

Je-li archiv oindexován, je zřejmé, že již není problém rychlým způsobem pomocí *fulltextového vyhledávače* zodpovědět například otázku: *co řekl Jan Zahradil v médiích během uplynulých dvou měsíců o Evropské unii*? Na druhé straně je pochopitelné, že „ruční“ přepis a popis těchto záznamů vyžaduje enormní množství lidského úsilí, což je hlavní důvod, proč se laboratoře počítačového zpracování řeči na celém světě snaží o tvorbu co nejdokonalejších automatizovaných transkripčních systémů. Pojmem *media mining systém* (MMS) je pak označován soubor softwarových a hardwarových nástrojů, jež umožňují automatizovanou tvorbu, správu, aktualizaci a prohledávání multimediálních archivů.



Obrázek 1.1: *Názorná ukázka víceúrovňové transkripce záznamu zpravodajského pořadu pro účely media miningu.*

1.2 Architektura automatického media mining systému

Typická architektura automatického media mining systému pro zpracování multimediálních dat je znázorněna na obrázku 1.2. Přijímaný audiovizuální signál je digitalizován a komprimován do některého ze standardizovaných formátů. Jeho audio složka[1] pak vstupuje do *modulu automatické transkripce*, jehož úkolem je poskytnout víceúrovňový časový popisu audio záznamu co možná nejvíce se blížícího ukázce na obrázku obr. 1.1. Následnou indexací je vytvořen rejstřík, jenž obsahuje odkazy na *meta data*, což jsou v podstatě dílčí multimediální záznamy spolu s korespondujícími přepisy. Meta data jsou uchovávána a organizována prostřednictvím *media mining serveru*. Pro vyhledávání v archivu, tj. komunikaci se serverem, pak slouží klientská aplikace s implementovaným *uživatelským rozhraním*.

1.3 Modul automatické transkripce

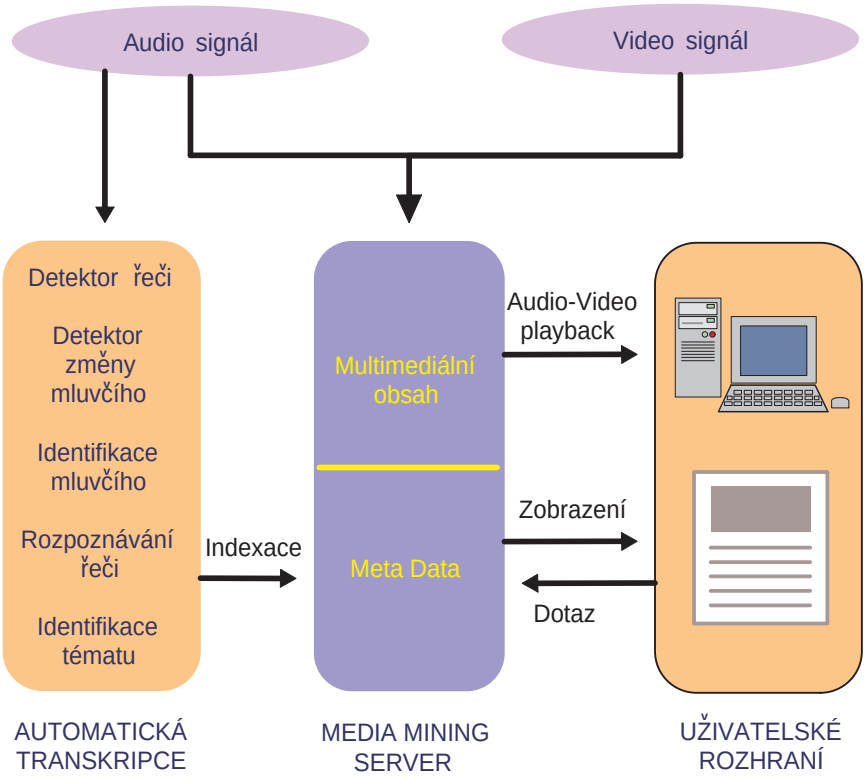
Proces automatické transkripce začíná buď aplikací *detektoru řeči* nebo *detektoru změny mluvčího*, záleží na konkrétní implementaci systému. V prvním případě jsou nejprve odstraněny neřečové části záznamu, jimiž není pouze ticho, ale také hudba či jiné ruchy, a následně se hledají místa změny řečníka. Nalezením těchto bodů změny vznikají segmenty, jež obsahují řeč právě jednoho mluvčího. Tyto segmenty jsou poté podrobeny procesu *identifikace řečníka*, čímž vybereme nejpravděpodobnějšího mluvčího z existující databáze osob. Následná *verifikace* potvrdí či zamítne hypotézu, zda se opravdu jedná o odpovídající osobu.

V druhém případě se nejprve využije detektoru změny řečníka k nalezení akusticky homogenních segmentů. Místo označení detektor změny řečníka se pak používá pojmu *detektor akustických změn*, využívá však naprosto stejných principů. Detekované segmenty jsou následně odesílány do jednotky *identifikace segmentu*, která pomocí principů podobných identifikaci mluvčího umožňuje rozlišit i jiné třídy dat, jako jsou hudba, ticho...atd.

V obou případech je výsledkem audio signál segmentovaný dle jednotlivých mluvčích spolu s informací o identitě mluvčího. V ideálním případě, jedná-li se o osobu frekventovanou v médiích, známe přímo její jméno. V ostatních případech je užitečnou pomůckou alespoň pohlaví dané osoby. Toto je důležité nejen z hlediska transkripce jako takové, ale především pro robustnost převodu řeči do textové podoby.

Rozpoznávače řeči pro tyto účely jsou v současné době založeny výhradně na technologii *skrytých Markovových modelů* (HMM). Mimo technologie samé jsou dalšími důležitými

[1] Audio signál je před vstupem konvertován do vhodného příznakového prostoru, v této práci se jedná o dobře známe mel-frekvenční kepstrální koeficienty (MFCC).



Obrázek 1.2: *Typická architektura automatizovaného media mining systému.*

součástmi rozpoznávače *akustický a jazykový model*. Akustický model je primárně koncipován jako *na mluvčím nezávislý*. Je-li k dispozici totožnost mluvčího, lze použít tzv. *na mluvčím závislý* akustický model, což vede k vyšší přesnosti přepisu. Obdobně je tomu i s jazykovým modelem a *identifikací tématu*.

1.4 Cíle práce

- Prozkoumat a jednotným způsobem popsat principy, z nichž lze vycházet při řešení úlohy detekce změny řečníka;
- modifikovat či nově vytvořit metody vhodné pro zadanou problematiku;
- vypracovat jejich efektivní algoritmizaci;
- navrhnout metodu pracující v on-line režimu;
- porovnat jednotlivé metody na uměle vytvořených datech simulujících změny mluvčích a ověřit nejslibnější přístupy na reálných nahrávkách;
- realizovat modul pro segmentaci skutečných zpravodajských pořadů využitelný jako součást vyvíjeného systému pro přepis televizních zpráv.

KAPITOLA 2

DATABÁZE A METODY VYHODNOCENÍ ÚSPĚŠNOSTI SEGMENTACE

Trénování a testování navržených metod detekce změny řečníka bylo realizováno na pěti různých databázích. Zdrojem prvních čtyř „uměle namíchaných“ databází se staly záznamy různých pořadů českých televizních a rozhlasových stanic. Tyto pořady byly pečlivě anotovány a údaje o počátku a konci promluvy jednotlivých mluvčích byly využity jako zdroj pro vznik umělé databáze ART. Jelikož tato databáze obsahovala „příliš reálná data“, což znamená, že se v ní vyskytovaly i různé neřečové části a aditivní rušení na pozadí řečového signálu, byly vytvořeny navíc „ideální“ databáze S-ART, kde byly tyto jevy v maximální míře potlačeny. Konkrétně se jedná o databázi FS-ART, složenou pouze z ženských hlasů, MS-ART, jež obsahuje pouze mužskou řeč a smíšenou databázi S-ART. Dalším dobrým důvodem tvorby umělých databází byla potřeba mít trénovací a testovací data, u nichž jsou zcela jednoznačně známy přesné pozice bodů změny. Pro testování navržených metod na reálných datech pak posloužila panevropská databáze televizních zpráv, která vznikla v rámci projektu Evropské Unie COST 278.

2.1 Metody vyhodnocení výsledků

Vyhodnocení výsledků je běžně praktikováno formou obousměrného hledání nejbližšího souseda mezi referenčními a vypočtenými body změny. *i-tý* vypočtený bod změny t_{ci} považujeme za správně nalezený (HIT) a odpovídající *j-tému* referenčnímu bodu změny t_{rj} tehdy a jen tehdy, když

1. t_{ci} je vypočtený bod změny nejblíže referenčnímu t_{rj},
2. t_{rj} je referenční bod změny nejblíže vypočtenému t_{ci},
3. vzdálenost mezi nimi je menší než určitá mez τ_{max}, typicky $|t_{ci} - t_{rj}| < 1\ s$.

O takovýchto dvojicích pak tvrdíme, že tvoří *pár* a jejich počet označíme H. Všechny vypočtené body, jež netvoří pár, označujeme jako INZERCE a jejich počet I. Obdobně pak pro referenční body změny, jež nebyly nalezeny (netvoří pár), se používá označení DELECE, pro jejich počet D.

Označíme-li celkový počet referenčních bodů $N = H + D$, lze nadefinovat tři základní míry pro vyhodnocení úspěšnosti detekce změny řečníka:

$$R = \frac{H}{N} \times 100\% \tag{2.1}$$

$$P = \frac{H}{H+I} \times 100\% \tag{2.2}$$

$$F = \frac{2 \times R \times P}{R+P}. \tag{2.3}$$

Míra R se nazývá *recall* a značí procento správně nalezených ze všech hledaných bodů změny. V mezním případě, umístí-li detektor bod změny do každého diskrétního časového

okamžiku, se bude recall blížit ideálním 100%, což ovšem neznamená, že máme k dispozici kvalitní detektor. Z tohoto důvodu se navíc používá míra zvaná *precision* (P), která vyjadřuje procento správně nalezených ze všech nalezených bodů změny. Tyto dvě míry jsou protichůdné, tj. roste-li jedna, klesá druhá a naopak. Avšak ani jedna z těchto měr nemá lokální maximum, pročež nejsou vhodné jako kritéria pro trénování detektoru. Toto je důvodem zavedení míry zvané *F-rate* (F), která tuto podmínku splňuje.

Dalším hlediskem vyhodnocení detektoru změn je přesnost poloh správně nalezených bodů změny. Pro tento účel využijeme histogram chyb časového zarovnání

$$\Delta t_h = |t_{ci} - t_{rj}|, \tag{2.4}$$

kde $h = 1, \ldots, H$ a H je celkový počet správně detekovaných změn. Z histogramu odečteme tři hodnoty vypovídající o přesnosti detektoru:

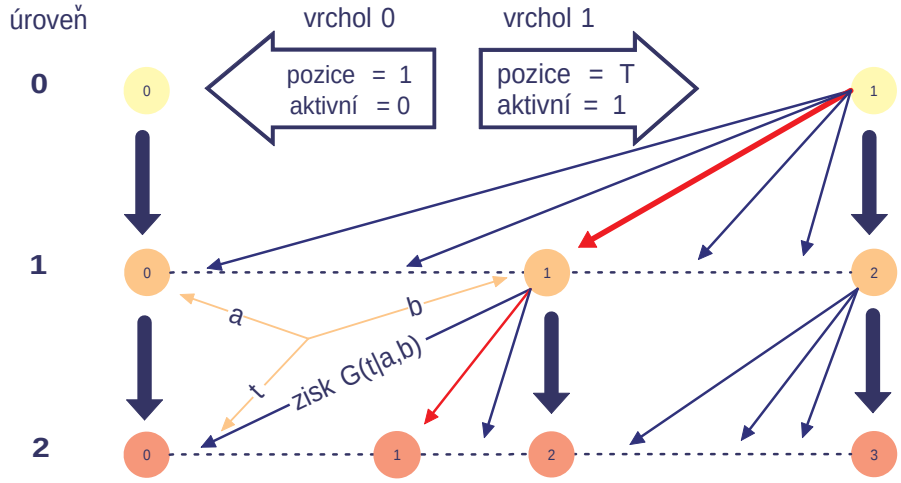
1. $\Delta_{2/3}$ označíme maximální chybu časového zarovnání pro 2/3 všech správně detekovaných bodů změny;

2. $\Delta_{0.95}$ označíme maximální chybu časového zarovnání pro 95% všech správně detekovaných bodů změny;

3. δ_{10} označíme procento správně detekovaných bodů změny, jejichž chyba zarovnání je menší než 10 ms.

METODA BINÁRNÍHO DĚLENÍ

Princip metody binárního dělení je znázorněn na obrázku 3.1 a vychází z práce [Vostrikova, 1981]. Nadefinujeme si nultou *úroveň* procesu detekce změny řečníka se dvěma *vrcholy* odpovídajícími počátku a konci akustického signálu. Každému vrcholu přiřadíme dvě *vlastnosti*. První je pozice vrcholu, druhou pak vlastnost, zdali je vrchol *aktivní*, tj. je-li možné přejít z daného vrcholu do následující úrovně. Tenké šipky na obrázku značí *cestu* z jedné úrovně do úrovně následující a *zisk* daný tímto přechodem odpovídá výpočtu rovnic uvedených v části 3.1. V souladu s teorií popsanou v disertační práci vybereme tu cestu, jež přináší nejvyšší zisk. Když bude zisk větší než určitá kritická hranice K, ustanovíme nový vrchol v následující úrovni a nastavíme mu následující vlastnosti: *pozice* bude odpovídat místu, kam vedla nejvýhodnější cesta a *aktivní* = 1. Poté zkopírujeme testovaný vrchol do následující úrovně. Pakliže zisk nejlepší cesty nepřekročí kritickou hranici K, nebude nový vrchol ustaven, dojde pouze ke zkopírování původního do nové úrovně a změní se nastavení vlastnosti *aktivní* = 0. Algoritmus ukončíme na té úrovni, kde již nebude žádný *aktivní* vrchol.



Obrázek 3.1: *Grafické znázornění metody binárního dělení.*

3.1 Výpočet zisku cesty

V disertační práci byly odvozeny tři přístupy k detekci bodu změny pomocí testování hypotéz o změně parametrů gaussovského procesu - viz např. [Lehmann, 1986]. V souvislosti s terminologií zavedenou v této kapitole zavedeme na jejich základě zisk cesty $G(t|a,b)$.

3.1.1 Metoda maximální věrohodnosti

Z přístupu metodou maximální věrohodnosti vyplynula testovací statistika maximálního typu [Horváth, 1993], na jejímž základě lze zisk G cesty z bodu b do bodu t odvodit ve tvaru

$$G_{MLLR}(t|a,b) = \tag{3.1}$$

$$\alpha\sqrt{\left[(b-a+1)\log|\hat{\Sigma}| - (t-a+1)\log|\hat{\Sigma}_1| - (b-t)\log|\hat{\Sigma}_T|\right]} - \beta,$$

když $a-d > t > b+d$ a d je rozměr příznakového vektoru. Proměnné α a β lze je vyjádřit jako:

$$\alpha = (2\log\log(b-a+1))^{\frac{1}{2}} \text{ a} \tag{3.2}$$

$$\beta = 2\log\log(b-a+1) + d\log\log\log b - a + 1) - \log\Gamma(d). \tag{3.3}$$

Matice $\hat{\Sigma}$, $\hat{\Sigma}_1$, $\hat{\Sigma}_T$ jsou kovariance dat $\{x_a,\ldots,x_b\}$, $\{x_a,\ldots,x_t\}$, $\{x_{t+1},\ldots,x_b\}$. Metoda využívající funkci G_{MLLR} bude v dalším textu označována zkratkou MLLR (Maximum Log-Likelihood Ratio).

3.1.2 SIC s fixní hranicí kritického regionu

Budeme-li uvažovat přístup k testování hypotéz pomocí *Schwarzova informačního kritéria* (SIC) [Schwarz, 1978] s pevnou hranicí kritického regionu, zisk G cesty z bodu b do bodu t za podmínky omezení předchozím bodem změny a lze vyjádřit jako

$$G_{FTSIC}(t|a,b) = \tag{3.4}$$
$$(b-a+1)\log|\hat{\Sigma}| - (t-a+1)\log|\hat{\Sigma}_1| - (b-t)\log|\hat{\Sigma}_T| - D\log(b-a+1),$$

kde D je rozdíl v dimenzi hypotéz - viz [Chen J., 2000]. Tato metoda bude v následujícím textu označována zkratkou FTSIC (Fixed Threshold SIC).

3.1.3 SIC s fixní váhou penalizační funkce

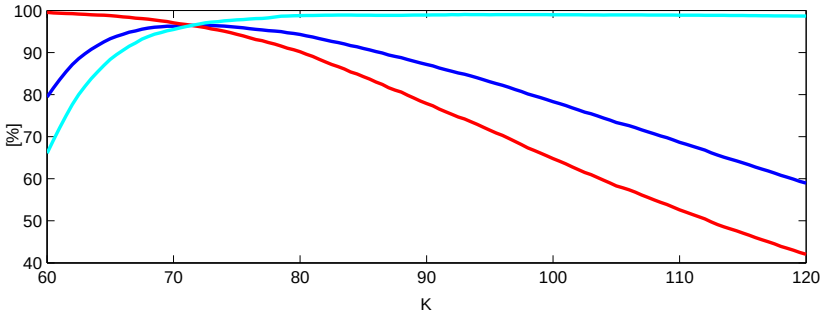
Posledním ověřovaným přístupem je z teoretického hlediska nejméně opodstatněná, leč nejvíce využívaná, metoda s pevnou váhou penalizační funkce. Zisk pak nabývá podoby

$$G_{FPWSIC}(t|a,b) = \tag{3.5}$$
$$\frac{(b-a+1)\log|\hat{\Sigma}| - (t-a+1)\log|\hat{\Sigma}_1| - (b-t)\log|\hat{\Sigma}_T|}{D\log(b-a+1)},$$

kde D je opět rozdíl v dimenzi hypotéz. V dalším textu ponese označení FPWSIC (Fixed Penalty Weight SIC).

3.2 Experimentální výsledky

V této části jsou shrnuty výsledky detekce změny řečníka na všech dostupných databázích. Základem vyhodnocení dílčích metod je idealizovaná databáze S-ART, neboť obsahuje pouze řeč od jednotlivých mluvčích, tj. neobsahuje žádné ticho ani rušení. Trochu realističtější umělá databáze ART pak umožňuje trénování dílčích metod pro účely jejich aplikace na reálná data, jež jsou reprezentována databází COST 278. Databáze FS-ART a MS-ART umožňují zodpovědět otázku, jak se liší detekovatelnost změn v závislosti na pohlaví mluvčích.



Obrázek 3.2: *Graf závisloti měr recall (červěně), precision (zeleně) a F-rate (modře) na hranici kritického regionu K pro metodu MLLR a trénovací část databáze S-ART.*

Databáze	Trénovací		Testovací		
Princip	K	F_{max} [%]	F [%]	R [%]	P [%]
MLLR	72.5	96.51	96.27	95.73	96.82
FTSIC	570	96.32	96.05	95.73	96.73
FPWSIC	2.05	94.39	94.19	92.92	95.49

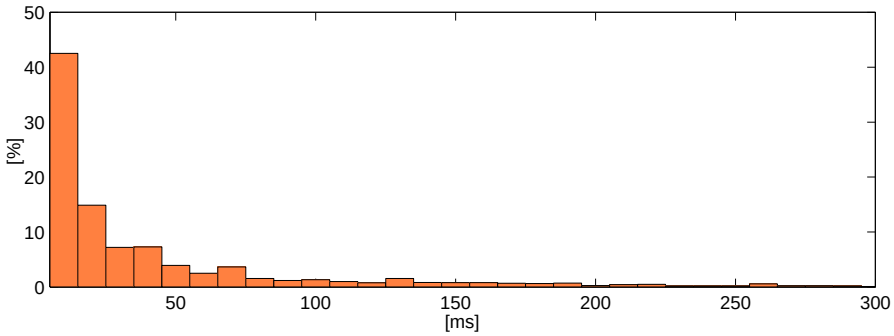
Tabulka 3.1: *Vyhodnocení úspěšnosti detekce změny řečníka metod MLLR, FTSIC a FPWSIC na databázi S-ART.*

3.2.1 Databáze S-ART

Typický graf měr *recall* (R), *precision* (P) a *F-rate* (F) v závislosti na hranici kritického regionu K je pro trénovací část databáze S-ART uveden na obrázku 3.2. Jelikož jsou tyto křivky velmi podobné pro všechny tři testované přístupy, je zobrazen průběh pouze pro metodu MLLR.

Jak již bylo uvedeno dříve, autor zvolil jako kritérium, které má být maximalizováno, míru *F-rate*. Kritická hranice K odpovídající maximu F_{max} dosaženému na trénovací databázi byla použita k vyhodnocení metody na testovací části databáze. Výsledky všech tří přístupů jsou shrnuty v tabulce 3.1. Z této tabulky vyplývá, že za daných podmínek lze odhad parametrů u všech metod nazvat konzistentním, tj. na trénovací a testovací části dosahuje *F-rate* velmi podobných hodnot. Nejhůře ($F = 94.19\%$) dopadla metoda využívající váhování penalizační funkce Schwarz-Bayesova kritéria FPWSIC. Nicméně tento přístup je běžně využívaný, což je možná dáno tím, že absolutní rozdíl oproti zbylým přístupům $\Delta F \approx 2\ \%$ není nijak dramatický. Rozdíl mezi přístupy metodou maximální věrohodnosti a SIC je z praktického hlediska velmi malý $\Delta F \approx 0.2\ \%$. Ovšem z hlediska statistické signifikance je metoda MLLR oproti FTSIC výrazně lepší, neboť byla její menší chybovost potvrzena na úrovni $\alpha_0 = 0.1\%$.

Dalším zajímavým pohledem na vyhodnocení navržených metod binárního dělení je přesnost časového určení správně nalezených hranic. Histogram těchto chyb pro metodu MLLR je uveden na obrázku 3.3. Srovnání parametrů histogramů ostatních metod pak nalezneme v tabulce 3.2. Není překvapující, že se tyto parametry příliš neliší, neboť pozice každého bodu je určována u všech tří metod stejnou rovnicí. Drobné rozdíly nelze přičítat rozdílu v kvalitě dílčích metod. Z tabulky 3.2 vyplývá, že detektor změn mluvčích založený na principu binárního dělení určí 2/3 resp. 95 % všech změn s přeností větší než 40 resp.



Obrázek 3.3: *Histogram chyb časového určení pozic správně nalezených bodů změny metodou MLLR na testovací části databáze S-ART. Histogram je dělen po intervalech 10 ms.*

250 ms [1]. Míra δ_{10} byla zavedena v souvislosti s použitou parametrizací, kdy je vektor příznaků získáván každých 10 ms. Z tohoto pohledu lze cca 42.5 % změn považovat za určené zcela přesně.

Princip	K	$\Delta_{2/3}$ [ms]	$\Delta_{0.95}$ [ms]	δ_{10} [%]
MLLR	72.5	40	250	42.53
FTSIC	570	40	240	42.63
FPWSIC	2.05	40	220	42.64

Tabulka 3.2: *Tabulka parametrů histogramů chyb určení časových pozic správně detekovaných bodů změny pro dílčí metody aplikované na testovací část databáze S-ART.*

3.2.2 Databáze FS-ART a MS-ART

Úspěšnost detekce změny mluvčího z pohledu porovnání dílčích metod binárního dělení potvrdilo výsledky testů na databázi S-ART, tj. signifikantně nejlepším byl opět přístup pomocí MLLR. Srovnání výsledků získaných pro čistě ženskou, čistě mužskou a smíšenou databázi metodou MLLR je uvedeno v tabulce 3.3.

Část	Trénovací		Testovací		
Databáze	K	F_{max} [%]	F [%]	R [%]	P [%]
FS-ART	68.5	93.25	92.78	92.54	93.02
MS-ART	71.5	95.68	95.57	95.79	95.39
S-ART	72.5	96.51	96.27	95.73	96.82

Tabulka 3.3: *Srovnání úspěšnosti detekce změny řečníka metodou MLLR pro čistě ženskou* FS-ART*, čistě mužskou MS-ART a smíšenou databázi S-ART.*

Z této tabulky je patrný pokles úspěšnosti detekce na obou „unisex“ databázích oproti databázi smíšené. Tento jev není překvapivý, neboť skutečnost, že u osob stejného pohlaví jsou rozdíly v hlasových charakteristikách menší, je zjevná. S tímto zřejmě souvisí i pokles

[1] Pro srovnání: 250 ms je doba trvání některých delších fonémů.

optimální hranice kritického regionu, čímž se nastavuje vyšší „citlivost“ detektoru. Zajímavým jevem je prudší pokles úspěšnosti FS-ART vs. S-ART ($\Delta F \approx 3.5\%$) oproti poklesu MS-ART vs. S-ART ($\Delta F \approx 0.7\%$). Můžeme usuzovat, že MFCC příznaky zajišťují větší separabilitu u mužských hlasů než u ženských. Na druhou stranu je toto pouze domněnka a provedené testy nelze považovat za důkaz.

3.2.3 Databáze ART

Důvodem tvorby databáze ART bylo mít možnost spolehlivě odhadnout volné parametry detektorů tak, aby je bylo možné používat v praxi. Z přehledu výsledků uvedených v tabulce 3.4 vyplývají dvě nepříjemné skutečnosti. Obě jsou zapříčiněny tím, jak se trénovací i testovací data blíží reálným signálům. Nejprve si povšimněme rostoucí optimální hranice kritického regionu odhadovaného na trénovacích datech. Tento vzestup je zapříčiněn tím, že reálnější signály obsahují různé typy rušení, které, nezačíná-li synchronně s počátkem promluvy, způsobuje chybu typu inzerce. Další inzerce jsou produkovány v místech, kde jsou např. delší úseky ticha. Vzhledem k tomu, že při trénování maximalizujeme *F-rate*, je přirozenou reakcí trénovacího algoritmu potlačit množství inzercí zvýšením kritické hranice K. Toto snížení citlivosti detektoru má ovšem také přímý dopad na množství delecí, neboť přestanou být detekovány některé méně významné změny mluvčích.

Databáze	Trénovací		Testovací		
Princip	K	F_{max} [%]	F [%]	R [%]	P [%]
MLLR	76.0	92.84	93.36	93.25	93.48
FTSIC	690	92.83	93.16	92.89	93.43
FPWSIC	2.20	91.59	91.83	90.68	93.01

Tabulka 3.4: *Vyhodnocení úspěšnosti detekce změny řečníka metod MLLR, FTSIC, FPWSIC na databázi ART.*

3.2.4 Databáze COST 278

Vzhledem k tomu, že databáze COST 278 není natolik rozsáhlá, aby ji bylo možné rozdělit na dostatečně velkou trénovací a testovací část, přejal autor metodiku vyhodnocení užívanou v rámci projektu COST 278 - viz [3]. Testování navržených metod probíhalo podle dvou scénářů. První scénář předpokládal trénování systému na externí databázi a následné testování na celé databázi COST. Pro natrénování systému autor využil databáze ART, základní shrnutí výsledků je znázorněno v tabulce 3.5.

Databáze	ART		COST 278		
Princip	K	F_{max} [%]	F [%]	R [%]	P [%]
MLLR	76.0	92.84	68.27	84.56	57.24
FTSIC	690	92.83	69.65	82.88	60.07
FPWSIC	2.20	91.59	70.80	79.06	64.10

Tabulka 3.5: *Vyhodnocení úspěšnosti detekce změny řečníka metod MLLR, FTSIC, FPWSIC trénovaných na databázi ART a testovaných na DB COST 278.*

Z tabulky 3.5 je patrný znatelný pokles úspěšnost u všech tří metod o cca $\Delta F \approx 20\%$. Je způsobený především poklesem míry *precision*, což odpovídá prudkému nárůstu chyb

typu inzerce. Důvody tohoto jevu jsou z části vysvětleny v sekci 3.2.3 věnované databázi ART, z části jsou způsobeny tím, že databáze je neustále ve stádiu vzniku a obsahuje značné množství chyb. Zajímavějším momentem je absolutně opačné pořadí dílčích metod vzhledem k úspěšnosti, než jaké bylo dosaženo při tzv. *matched testech*[2] uvedených v předchozích částech. Zdá se, že metody FPWSIC a FTSIC vykazují vyšší odolnost proti nesourodým trénovacím a testovacím podmínkám než metoda MLLR. Na druhé straně si je však nutné uvědomit, že počet testovaných položek $P = 57$ je vcelku nízký. Vzhledem k tomu, že test significance za daných podmínek není příliš přesvědčivým důkazem, neodvažuje se autor z těchto výsledků vyvozovat žádný jednoznačný závěr.

Zvýšení věrohodnosti výsledků podle druhého scénáře bylo dosaženo pomocí principu rotace trénovacích a testovacích dat. Algoritmus byl natrénován vždy pro jednu komponentu databáze a následně otestován na zbylé části. Tento postup byl opakován pro všechny komponenty, tj. 10 krát. Úspěšnost detektoru pak byla vyhodnocena jako průměr dílčích výsledků a jejich hodnoty jsou uvedeny v tabulce 3.6.

Část	Trénovací		Testovací		
Princip	ϕK	ϕF_{max} [%]	ϕF [%]	ϕR [%]	ϕP [%]
MLLR	90.95	74.16	70.74	74.53	68.62
FTSIC	1057	73.51	70.52	75.13	67.72
FPWSIC	2.30	73.31	70.31	72.39	69.47

Tabulka 3.6: *Výsledky cyklického testu na databázi COST 278.*

Z těchto testů vychází nejlépe opět metoda MLLR následovaná FTSIC a FPWSIC. U metod MLLR a FTSIC lze pozorovat zlepšení (2.5%, 0.8%) oproti předchozímu testu, kdy byly trénovány na externích datech. U metody FPWSIC nastává naopak mírné zhoršení (0.5%), což naznačuje nekonzistenci odhadu její kritické hranice.

3.3 Shrnutí

Z pohledu praktického využití detektoru změn mluvčího založeného na metodě binárního dělení jsou zajímavé údaje o úspěšnosti detekce změn a výpočetních nárocích. Výpočetní náročnost všech tří porovnávaných verzí je zhruba stejná a závisí jak na délce celého signálu, tak na počtu změn. Na běžném osobním počítači (P IV, 2.4 GHz) a používaných databázích byl každý signál rozsegmentován v průměru 10 krát rychleji, než byla jeho doba trvání. Paměťové nároky jsou determinovány především poli z_1 a z_2 - viz DP. Jejich velikost závisí na použité parametrizaci a délce signálu. Při použité parametrizaci MFCC, kdy byl příznakový vektor o rozměru 12 emitován každých 10 ms, je to zhruba 40 kB na jednu sekundu záznamu. Z pohledu statistického je nejúspěšnější verzí metody binárního dělení princip MLLR, nejhorší pak FPWSIC. Z praktického hlediska je však mezi nimi tak malý rozdíl, že je lze považovat za stejné.

[2] Test, kdy trénovací a testovací data jsou obdobného charakteru.

METODA GLOBÁLNÍ MAXIMALIZACE BIC

Všechny ostatní metody popisované v této práci jsou založeny na předpokladu, že analyzovaná část signálu obsahuje pouze jeden bod změny. Multiple change-point problém je pak řešen rozkladem na sérii single change-point úloh za pomoci vhodných metod. Je patrné, že tyto metody zajišťují pouze lokálně optimální segmentace signálu. Z tohoto důvodu se autor pokusil nadefinovat problém detekce více bodů změny jako globální optimalizační úlohu s jasným kritériem, jímž je nalezení nejlepší sekvence stavů skrytého markovovského modelu. Vzhledem ke skutečnosti, že předem není znám počet stavů ani parametry dílčích stavů, pokusil se autor tuto nepříjemnost řešit pomocí dobře známých aproximací - viz [Chickering & Heckerman, 1996] - směřujících ke globální maximalizaci Bayesovského informačního kritéria.

Z teoretických úvah nastíněných v disertační práci vyplývá, že optimální sekvenci bodů změn lze získat nalezením takové segmentace $\mathbf{t}_i^S$, jež bude maximalizovat logaritmickou věrohodnost dat, potažmo BIC:

$$\log p(\mathbf{x}|\mathbf{S}_i^S) \approx \mathcal{BIC}(\mathbf{x}|\mathbf{S}_i^S) = \sum_{s=1}^{S} \ell(t_i^s, t_i^{s-1}) - \lambda \frac{Sc}{2} \log T, \tag{4.1}$$

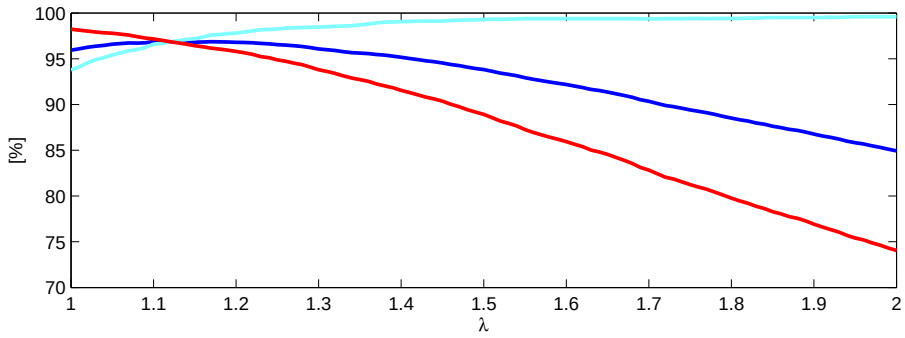
kde $\ell(t_i^s, t_i^{s-1})$ je záporný logaritmus maximální věrohodnosti vypočtený z dat v pozicích t_i^{s-1} až t_i^s a c je počet volných parametrů gaussovského modelu. První člen rovnice (4.1) vyjadřuje věrohodnost segmentace $\mathbf{t}_i^S$. Jeho maximalizací pro daný počet mluvčích S získáme body změn, tj. segmentaci $\mathbf{t}_{opt}^S$. Segmentaci $\mathbf{t}_{opt}^S$ určíme pro všechny možné počty mluvčích $S = 1, ..., S_{max}$. Aplikace druhého členu rovnice (4.1) pak umožní vybrat správný řád segmentace S a tím i optimální segmentaci $\mathbf{t}_{opt}$. Efektivní maximalizace rovnice (4.1) je řešena pomocí dynamického programování.

4.1 Experimentální výsledky

Z obrázku 4.1 je patrné, že průběh měr *recall(R)* , *precision(P)* a *F-rate(F)* je velmi podobný odpovídajícímu grafu získaného z trénovacích dat databáze S-ART metodou binárního dělení. Pro ostatní databáze jsou průběhy křivek trénování metody natolik podobné, že je zbytečné je zde dále uvádět.

Z hlediska vyhodnocení metody globální maximalizace BIC (GMBIC) jsou zajímavé údaje shrnuté v tabulce 4.1, kde jsou uvedeny výsledky testů na dílčích databázích. Pro porovnání zde také nalezneme výsledky nejlepší verze metody binárního dělení - metody MLLR. Z této tabulky je patrné, že metodou GMBIC lze dosahovat lepších výsledků než metodou MLLR, což je potvrzeno konzistentním zvýšením míry *F-rate* na všech využívaných databázích.

Test na databázi S-ART přinesl pouze nepatrné zvýšení úspěšnosti detekce bodu změny, $\Delta F \approx 0.5\%$. Ze statistického testu ale vyplývá, že se jedná o signifikantní zlepšení - dokonce na úrovni signifikance $\alpha_0 = 0.1\ \%$. Poněkud rozporuplných výsledků však bylo dosaženo při porovnání přesnosti detekce bodů změny z hlediska jejich časových poloh. Z tabulky 4.2 je možno vypozorovat, že dochází k výraznému zlepšení míry $\Delta_{2/3}$ o 70 ms.

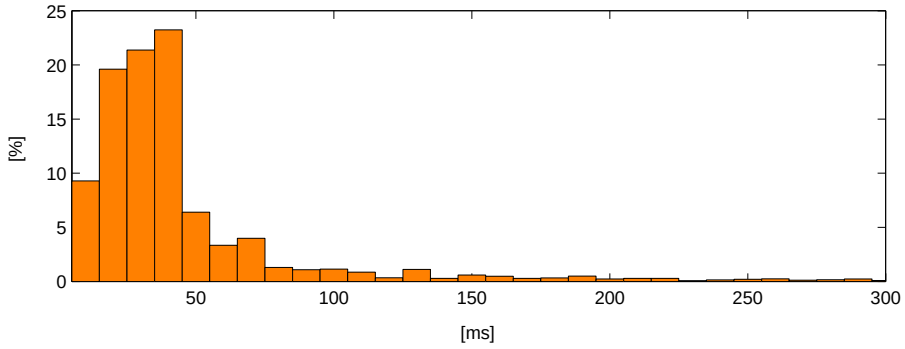


Obrázek 4.1: *Graf závisloti měr recall (červěně), precison (zeleně) a F-rate (modře) na hranici kritického regionu K pro metodu globální maximalizace BIC a trénovací část databáze S-ART.*

Část	Trénovací		Testovací			MLLR
Databáze	λ	F_{max} [%]	F [%]	R [%]	P [%]	F [%]
S-ART	1.10	96.88	96.76	97.26	96.27	96.27
MS-ART	1.15	95.99	95.99	96.12	95.86	95.27
FS-ART	1.03	93.79	93.50	93.83	93.18	92.78
ART	1.31	93.75	94.20	94.32	94.09	93.36
ART-COST	1.31	93.75	70.03	87.43	58.41	68.27
COST-COST	1.80	77.13	72.99	74.63	73.03	70.74

Tabulka 4.1: *Vyhodnocení úspěšnosti detekce změny řečníka metodou GMBIC. Řádek ART-COST označuje test na databázi COST 278 dle scénáře s trénováním na externích datech, COST-COST značí cyklický test.*

Na druhou stranu však razantně klesá absolutní přesnost detektoru δ_{10}. Zatímco metodou MLLR se podařilo absolutně přesně detekovat 42.63 % změn, metodou GMBIC pak pouhých 9.29 %. Tento jev je ilustrován na obrázku 4.2.



Obrázek 4.2: *Histogram chyb časového určení pozic správně nalezených bodů změny metodou GMBIC na testovací části databáze S-ART. Histogram je dělen po intervalech 10 ms.*

Vyhodnocení metody GMBIC na databázích MS-ART a FS-ART potvrzuje výsledky dosažené metodou binárního dělení (popsané v části 3.2.2), tj. mírně zhoršenou detekovatelnost změn mluvčích stejného pohlaví. Nicméně i na těchto datech lze pozorovat mírné

zlepšení oproti MLLR - $\Delta F \approx 0.7\%$ - potvrzené testem statistické signifikance.

Metoda	λ/K	$\Delta_{2/3}$ [ms]	$\Delta_{0.95}$ [ms]	δ_{10} [%]
GMBIC	1.10	40	180	9.29
MLLR	72.5	40	250	42.63

Tabulka 4.2: *Tabulka parametrů histogramů chyb určení časových pozic správně detekovaných bodů změny pro dílčí metody aplikované na testovací část databáze S-ART.*

Pro vyhodnocení úspěšnosti detekce změn na reálných nahrávkách databáze COST 278 byly v části 3.2.4 nadefinovány dva druhy testů. Test, kdy byla pro trénování užita externí data (databáze ART), přinesl zlepšení detektoru o $\Delta F \approx 1.8\%$ - viz řádek ART-COST tabulky 4.1. Z řádku COST-COST téže tabulky je patrné, že u testu založeného na rotaci trénovacích a testovacích dat došlo ještě k výraznějšímu zlepšení oproti metodě MLLR - $\Delta F \approx 2.2\%$.

4.2 Shrnutí

Vyhodnocením metody globální maximalizace BIC bylo potvrzeno, že přístup, kdy je multiple-change point problém řešen jako globální optimalizační úloha, přináší lepší výsledky (viz srovnání s výsledky partnerů spolupracujících v rámci projektu COST 278 [3]), než přístup přes dílčí lokální optimalizace. Tohoto zlepšení však bylo dosaženo na úkor razantního nárůstu jak výpočetních, tak paměťových nároků. I při efektivní implementaci metodou dynamického programování spotřebuje tento algoritmus při parametrizaci s 12 příznaky emitovanými každých 10 ms v průměru 0.5 MB operační paměti na 1 s záznamu. Růstu spotřeby paměti s délkou signálu je možné zabránit implementací formou kruhového zásobníku, čímž je však nutné zavést restrikci maximální možné délky segmentu. Při omezení této délky na 10 minut je GMBIC zhruba 10x pomalejší než metoda binárního dělení. Rychlost lze samozřejmě také výrazně ovlivnit nastavením parametrů prořezávání λ_{max}, λ_{min}. Ačkoliv je GMBIC metoda velmi výpočetně náročná a její výsledky nejsou zas tak výrazně lepší než u metody MLLR, perspektiva této metody tkví především v možnosti jednoduchého zabudování apriorní pravděpodobnosti segmentace formou distribuční funkce popř. histogramu délek segmentů, což je u metod využívajících přístupu pomocí testování hypotéz podstatně komplikovanější.

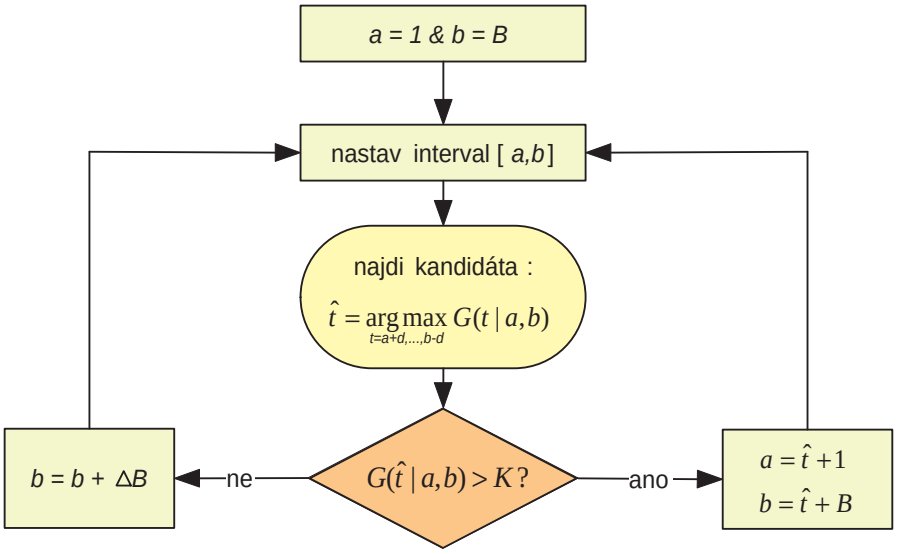
METODA S ADAPTIVNÍM OKNEM

Tato kapitola se zabývá velmi účinnou a přitom výpočetně nenáročnou metodou detekce více bodů změny, jíž autor nazývá metoda s adaptivním oknem, neboť je založena na detekci bodu změny v analyzujícím okně, jehož počátek i konec se neustále adaptuje tak, aby bylo možno rozhodovat o bodu změny s maximální možnou jistotou. Vychází přitom z dobře známého algoritmu publikovaného poprvé v [Chen & kol., 1998]. Tato metoda má však tři volné parametry a bohužel autoři ani jejich následovníci nepopisují možnosti jejich efektivního odhadu. Proto autor této práce navrhl takovou její modifikaci, že dva ze tří parametrů pak již nemají na úspěšnost metody takřka žádný vliv a třetí parametr lze odhadnout pomocí metody binárního dělení.

5.1 Originální algoritmus

Originální algoritmus metody s adaptivním oknem (AWIN) schematicky znázorňuje obrázek 5.1. Proměnnými $[a, b]$ označíme počátek a konec aktuálně analyzované části signálu. Dále zavedeme inicializační délku okna B, koeficient rozšíření ΔB a proměnnou $G(t|a, b)$ označíme zisk asociovaný s bodem změny t. Způsob jeho výpočtu je stejný jako u metody binárního dělení a definici lze nalézt v části 3.1.

Celý algoritmus funguje na velmi jednoduchém principu. Inicializujeme okno o velikosti B a umístíme ho na počátek signálu. V daném okně najdeme takový bod $\hat{t}$, jenž bude maximalizovat zisk $G(\hat{t}|a, b)$ a označíme ho jako *kandidáta* na bod změny. Je-li kandidát potvrzen, tj. $G(\hat{t}|a, b) > K$, lze ho považovat za bod změny, počátek analyzujícího okna je přesunut do polohy $\hat{t} + 1$ a velikost okna je nastavena na hodnotu B. Není-li kandidát $\hat{t}$ potvrzen, je okno rozšířeno o délku ΔB a celý postup se opakuje.

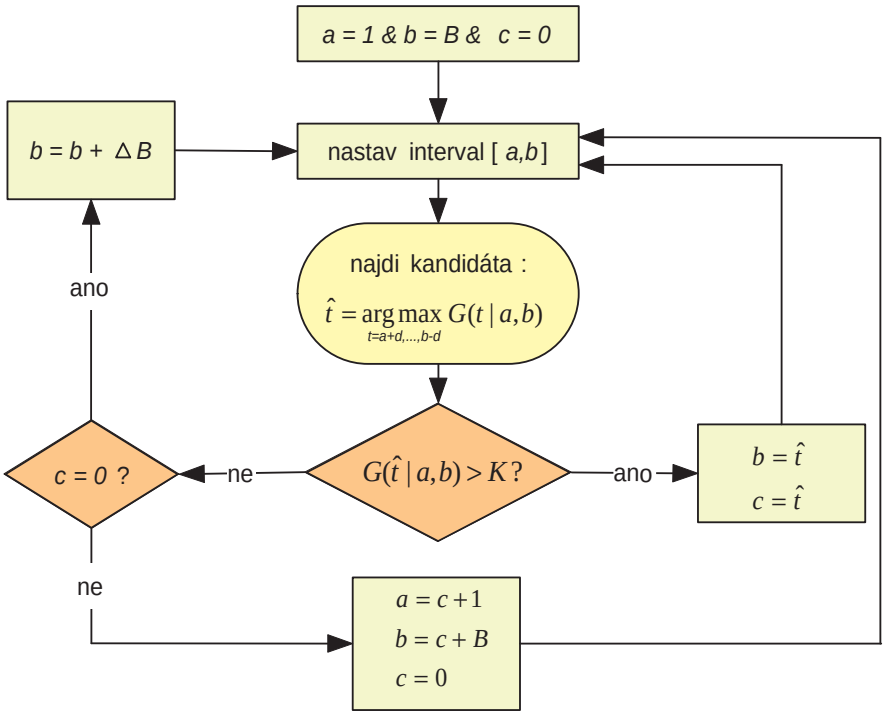


Obrázek 5.1: *Schéma originální metody s adaptivním oknem.*

5.2 Modifikovaný algoritmus

Modifikovaný algoritmus (MAWIN) funguje na velmi obdobném principu jako jeho základní verze. V aktuálně analyzované části signálu však není hledána pouze jediná změna, ale ta změna, která se nachází nejblíže počátku, čímž lze s úspěchem takřka eliminovat vliv parametrů B a ΔB na úspěšnost detekce. V reálné implementaci pak tato změna znamená přidání jedné podmínky a jedné proměnné c - viz obrázek 5.2.

Algoritmus inicializujeme stejně jako v předchozím případě a najdeme kandidáta na bod změny. Poté mohou nastat dva různé případy. Kandidát není potvrzen a rozšíříme analyzující okno. Je-li kandidát $\hat{t}$ potvrzen, zmenšíme okno na velikost $[a, \hat{t}]$ a najdeme v něm dalšího kandidáta. Bude-li tento kandidát potvrzen, opět okno zmenšíme. V opačném případě jsme nalezli bod změny, jenž byl z hlediska původního okna nejvíce vlevo, posuneme analyzující okno do tohoto bodu a nastavíme jeho inicializační velikost B.



Obrázek 5.2: *Schéma modifikované metody s adpativním oknem.*

5.3 Experimentální výsledky

Metoda s adaptivním oknem je koncipována jako tzv. *on-line metoda*, což znamená, že je možné získat bod změny s minimálním zpožděním vůči kontinuálně dodávaným datům. Z tohoto důvodu je v této části - na rozdíl od předchozích experimentálních výsledků - metoda vyhodnocena také z hlediska zpoždění detekce změny vůči reálnému času. Mezi sledované údaje patří průměrné zpoždění na detekovanou změnu (AVD) a maximální zpoždění průměrované přes každou databázovou položku (MXD). Vzhledem k předpokladu, že online algoritmus by měl fungovat v reálném čase, je dalším sledovaným parametrem metody množství numerických operací. Jelikož výpočetní náročnost souvisí s délkou signálu, po-

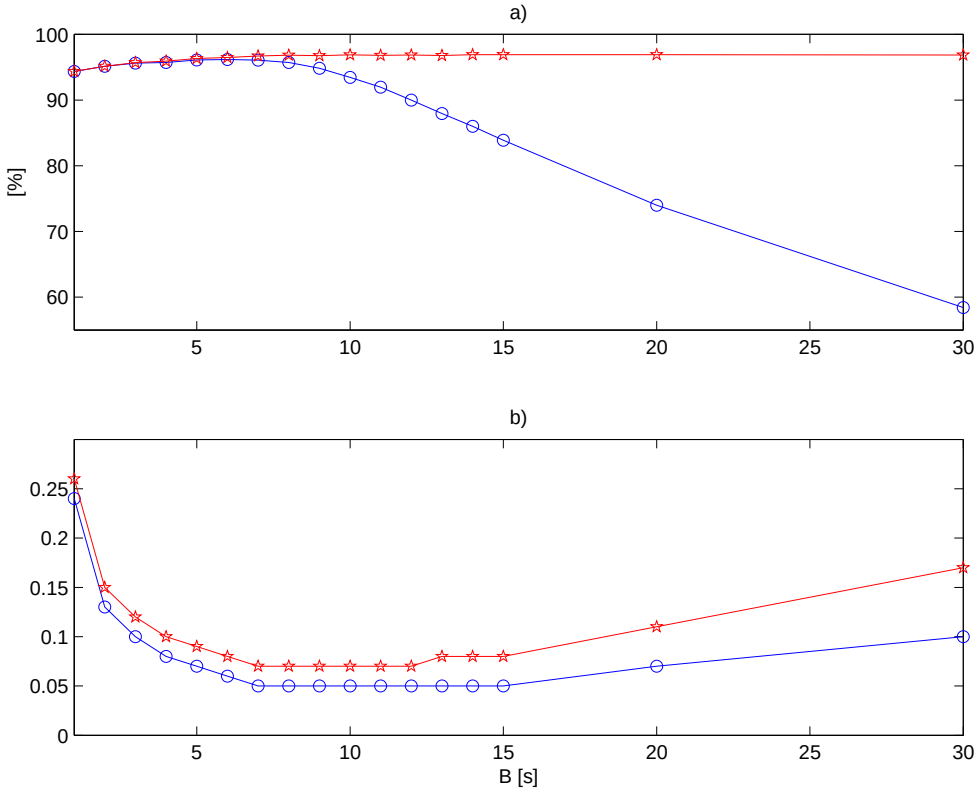
čtem změn a je určována především výpočtem determinantů kovariančních matic, definuje autor koeficient numerické náročnosti jako počet výpočtů determinantů na změnu a délku signálu (NNO).

Vzhledem k úzké spjatosti s metodou binárního dělení, pro niž bylo dosaženo nejlepších výsledků užitím verze MLLR, byla testována metoda s adaptivním oknem pouze v téže verzi, tj. zisk $G(t|a,b)$ byl počítán dle vztahu (3.1).

5.3.1 Srovnání originálního a modifikovaného algoritmu

Na obrázku 5.3 a) je znázorněn průběh míry *F-rate* v závislosti na velikosti inicializační délky okna B pro testovací část databáze S-ART. Hodnota kritické hranice zamítnutí kandidáta na bod změny $K = 72.5$ byla natrénována metodou binárního dělení na trénovací části databáze S-ART a koeficient rozšíření byl nastaven na hodnotu $\Delta B = B/2$. Z průběhu je patrné, že zatímco úspěšnost originální metody s rostoucí délkou inicializačního okna klesá, úspěšnost modifikované metody naopak mírně roste. Z podrobnějších údajů shrnutých v tabulce uvedené v dizertační práci navíc vyplývá, že modifikovaná metoda dosahuje nepatrně vyšší úspěšnosti i pro oblast počátku grafu, kde obě křivky takřka splývají.

Graf 5.3 b) zobrazuje průběh koeficientu výpočetní náročnosti (NNO) vzhledem k délce inicializačního okna B - získaného z databáze S-ART. Z tohoto obrázku je patrné, že výpočetní náročnost obou metod je takřka srovnatelná. Pro osobní počítač Pentium IV, 2.4 GHz hodnota koeficientu $NNO = 1$ značí, že algoritmus spotřebovává cca 50 % výkonu stroje.



Obrázek 5.3: *Srovnání originální (modře) a modfikované (červeně) metody s adaptivním oknem. Graf a) zobrazuje průběh míry F-rate pro různé volby velikosti okna B. Graf b) ilustruje závislost NNO na B.*

5.3.2 Vyhodnocení modifikovaného algortimu

Přehled výsledků získaných na jednotlivých databázích je uveden v tabulce 5.1. Při těchto testech byla kritická hranice K odhadována pomocí MLLR metody binárního dělení, velikost inicializačního okna B byla nastavena na 30 s a hodnota koeficientu rozšíření $\Delta B = 15\ s$. Pro srovnání jsou v tabulce uvedeny také výsledky metody MLLR a GMBIC.

Metoda	MAWIN				MLLR	GMBIC
Databáze	K	R [%]	P [%]	F [%]	F [%]	F [%]
S-ART	72.5	95.76	97.96	96.85	96.27	96.76
MS-ART	71.5	95.86	97.04	96.46	95.27	95.99
FS-ART	68.5	92.64	94.80	93.71	92.78	93.50
ART	76.0	93.14	94.88	93.90	93.36	94.20
ART-COST	76.0	84.33	59.11	69.50	68.27	70.30
COST-COST	90.95	74.44	70.33	71.65	70.74	72.99

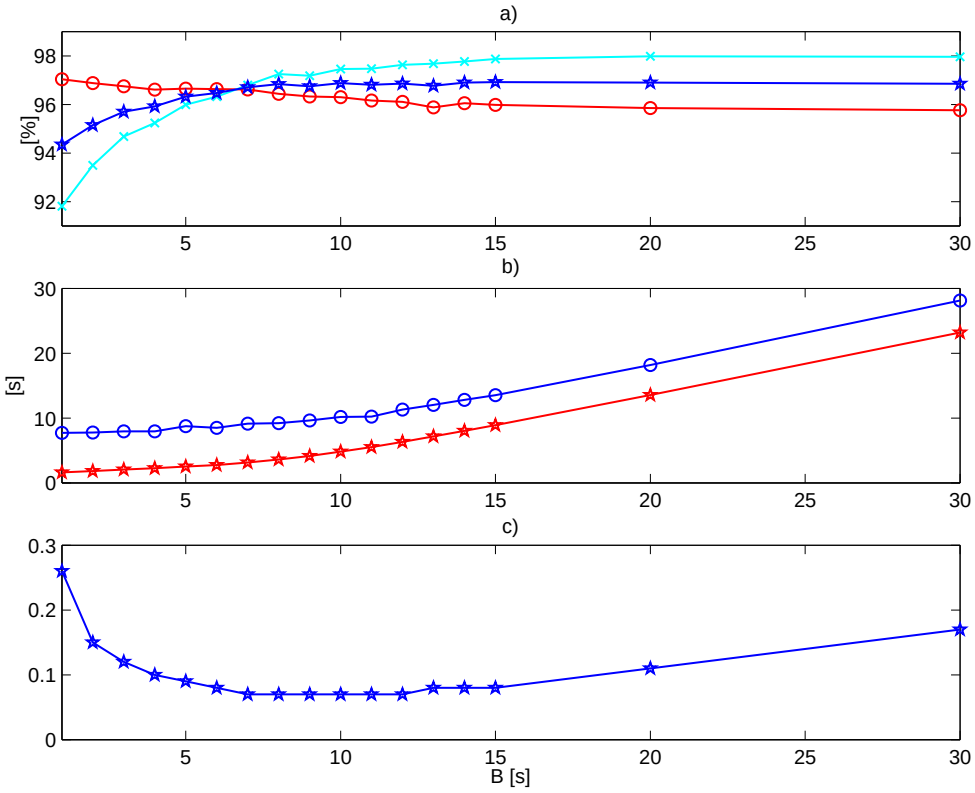
Tabulka 5.1: *Vyhodnocení úspěšnosti detekce změny řečníka metodou MAWIN. Řádek ART-COST označuje test na databázi COST 278 dle scénáře s trénováním na externích datech, COST-COST značí cyklický test.*

Na databázích S-ART, MS-ART a FS-ART lze pozorovat statisticky signifikantní zlepšení oproti metodě MLLR. Při srovnání s metodou GMBIC jsou výsledky metody MAWIN z hlediska statistické signifikance stejné (FS-ART, S-ART) či lepší (MS-ART) na úrovni signifikance $\alpha_0 = 0.1\ \%$. Vyhodnocením metody MAWIN na databázích ART a COST 278 pak zjistíme, že tyto výsledky jsou signifikantně lepší oproti MLLR a naopak signifikantně horší oproti GMBIC. Zajímavé je srovnání přesnosti detekce bodů změny z pohledu jejich časového určení. Z tabulky 5.2 je patrné, že metoda MAWIN v tomto směru poskytuje nejlepší výsledky.

Metoda	λ/K	$\Delta_{2/3}$ [ms]	$\Delta_{0.95}$ [ms]	δ_{10} [%]
MAWIN	72.5	30	180	45.96
MLLR	72.5	40	250	42.63
GMBIC	1.10	40	180	9.29

Tabulka 5.2: *Tabulka parametrů histogramů chyb určení časových pozic správně detekovaných bodů změny pro dílčí metody aplikované na testovací část databáze S-ART.*

Z obrázku 5.4 a), kde jsou zobrazeny průběhy měr *F-rate*, *recall* a *precision* v závislosti na délce inicializačního okna B pro databázi S-ART, plyne, že míra *F-rate* se dostává do saturace pro $B \approx 14\ s$. Tato hodnota pravděpodobně souvisí s průměrnou délkou segmentu databáze S-ART, jež je $6.31\ s$. Chceme-li tedy detekovat změny on-line a zároveň s co nejvyšší úspěšnosti, měla by být délka inicializačního okna minimálně dvojnásobkem očekávané délky segmentu. Tento požadavek je pochopitelný, neboť jedině tímto způsobem zajistíme, že rozhodnutí o změně nebude „v průměru“ učiněno dříve, než je k dispozici maximální možné množství relevantních dat. Zároveň z výše uvedeného logicky plyne, že průměrné zpoždění detekce změny AVD pro toto nastavení by mělo nabývat hodnoty zhruba poloviny incializační délky B, což je potvrzeno grafem 5.4 b). Okolo tohoto bodu se také pohybuje nastavení, jež klade nejmenší výpočetní nároky - viz graf 5.4 c).



Obrázek 5.4: *Vyhodnocení on-line vlastností modifikované metody s adaptivním oknem v závislosti na délce inicializačního okna B:* ***a)*** *průběh měr recall (červeně), precision (zeleně) a F-rate (modře);* ***b)*** *průběh maximálního MXD (modře) a průměrného AVD (červeně) zpoždění;* ***c)*** *průběh výpočetní náročnosti NNO.*

5.4 Shrnutí

Výsledkem těchto experimentů popisovaných v této kapitole je tedy zjištění, že metodou MAWIN lze získat nepatrně lepší výsledky než MLLR a stejné či zanedbatelně horší výsledky ve srovnání s metodou GMBIC. Zároveň metoda MAWIN poskytuje nejpřesnější odhad pozic bodů změny. Výpočetní nároky algoritmu jsou nízké, neboť průměrné zatížení PC Pentium IV s 2.4 GHz procesorem se při on-line implementaci pohybovalo kolem 5 % výkonu stroje. Paměťové nároky algoritmu jsou nízké a souvisí s aktuálně analyzovanou délkou signálu. Při parametrizaci, kdy je každý vektor příznaků rozměru 12 získáván každých 10 ms, odpovídá množství spotřebované paměti 40 kB na sekundu analyzujícího okna. Vzhledem k těmto faktům se zdá být metoda MAWIN nepraktičtějším z přístupů ověřovaných v této práci.

ZÁVĚR

Zřejmě nejpopulárnějším přístupem k řešení úlohy detekce změny řečníka je tzv. *metoda fixních oken*. Při její implementaci se však autor setkal s celou řadou problémů, jejichž řešením se odborná literatura příliš nezabývá. Největším z nich je především otázka odhadu parametrů této metody, neboť jich je příliš mnoho a pravděpodobně neexistuje způsob jejich rozumného odhadu. Velké množství parametrů je zapříčiněno zejména nutností implementace detektoru lokálních maxim, což také není právě triviální úloha. Méně užívaným přístupem je tzv. *metoda s adaptivním oknem*, která takový detektor ke své činnosti nepotřebuje. Ani u této metody se však z literatury nelze dozvědět nic o možnostech trénování. Další problematickou částí je její faktická implementace (stejně jako u metody fixních oken), kdy je sice prezentována jako metoda založená na testování hypotéz pomocí Schwarzova (Bayesova) informačního kritéria (FPWSIC), ale ve výsledku není používána zcela v souladu s teoretickými předpoklady. Výsledky publikované v této práci ilustrují, že metodami, jež aplikují teorii testování hypotéz ve správném znění (MLLR, FTSIC), lze dosáhnout lepších výsledků.

V této práci se autor zabýval třemi alternativními způsoby detekce změn řečníka. U všech navrhl jejich efektivní implementaci, algoritmus snadného odhadu volných parametrů a otestoval je na několika typech databází. První z navržených přístupů, *metoda binárního dělení*, není sice novou metodou, ale v oblasti detekce změny řečníka pravděpodobně ještě nebyla nikdy použita. Autor ji implementoval ve třech různých verzích. Verze MLLR je založená na přístupu pomocí testování jednoduchých hypotéz, verze FTSIC na přístupu přes testování kompozitních hypotéz, jež vede na řešení využívající Schwarzova (Bayesova) informačního kritéria. Poslední verze, FPWSIC, byla testována z důvodu její oblíbenosti mezi odbornou veřejností. Jelikož je metoda binárního dělení v jiných oblastech používána již od roku 1981, není překvapující, že poskytuje velmi rozumné výsledky - a to pro všechny její testované verze. Nicméně z provedené analýzy vyšla nejlépe metoda MLLR, když na zidealizované databázi S-ART bylo detekováno takřka 96 % všech existujících změn (*recall*) a téměř 97 % detekovaných změn bylo nalezeno správně (*precision*). Na reálné databázi COST 278 dopadla detekce změn o poznání hůře. Bezchybně bylo detekováno necelých 75 % všech existujících změn a pouze 69 % z nalezených změn bylo určeno správně.

Jelikož metoda binárního dělení je pouze šikovnou aplikací single-change point analýzy na multiple change-point problém, pokusil se autor navrhnout metodu nazvanou *globální maximalizace BIC*, která je definována jako globální jednoprůchodová metoda pro přímé řešení multiple change-point problému. GMBIC sice poskytuje nepatrně lepší výsledky než MLLR, cenou za zlepšení je však obrovský nárůst výpočetní a paměťové náročnosti. Její perspektiva tkví především v možnosti velmi jednoduše zahrnout do procesu detekce apriorní pravděpodobnost délek segmentů, což metody založené na testování hypotéz příliš neumožňují. Snížená funkčnost detektorů bez explicitně užívané apriorní pravděpodobnosti se projevuje především v okamžiku, jestliže je distribuční funkce apriorní pravděpodobnosti délek segmentů vícemodální, tj. jedná-li se např. o segmentaci nahrávky typu rozhovor, kdy se střídá krátká otázka moderátora s dlouhou odpovědí hosta.

Z praktického pohledu se nejvýhodnějším přístupem jeví *metoda s adaptivním oknem*, respektive její *modifikovaná* verze MAWIN. Metoda s adaptivním oknem je dobře známa již několik let, leč není příliš populární, zřejmě díky neobjasněnému způsobu odhadu jejích

tří volných parametrů. S touto nepříjemností se autor vypořádal takovou její modifikací, že dva ze tří parametrů přestávají mít na úspěšnost detekce vliv. Zároveň je v práci experimentálně prokázáno, že třetí volný parametr lze efektivně odhadnout pomocí metody binárního dělení. Velkou výhodou této metody je jak její on-line pojetí, tj. minimální zpoždění detekce bodu změny vůči reálnému času, tak její nízká výpočetní a paměťová náročnost. Úspěšnost metody s adaptivním oknem se pak pohybuje mezi dvěma výše uvedenými přístupy. Zároveň vyniká nejvyšší přesností detekce bodů změny z hlediska jejich časových pozic.

Cílem této práce bylo prozkoumat metody vhodné k detekci změny řečníka z pohledu algoritmického, nikoliv z pohledu vhodných příznaků. Nalezení vhodných příznaků lze považovat za zcela samostatnou problematiku, která je řešena především v úloze identifikace a verifikace řečníka. Zajímavým výsledkem uskutečněných experimentů z hlediska používaných MFCC příznaků je horší detekovatelnost změn žena-žena oproti změnám muž-muž.

Neznalého čtenáře by mohl zarazit prudký pokles úspěšnosti detekce změn při testech na reálné databázi oproti výsledkům z uměle míchaných idealizovaných databází, kdy došlo ke snížení míry *F-rate* o cca 25 %. Velký pokles je způsoben především nárůstem chyb typu inzerce, tedy falešnou detekcí změny řečníka. Tyto chyby jsou způsobeny zejména detekcí bodů, kdy přes sebe mluví několik mluvčích, do řeči hraje hudba či je na pozadí jiný aditivní, často nestacionární, šum. Je zřejmé, že detektory založené na testování změn parametrů stochastického procesu takovou změnu považují za významnější, než je změna mluvčího. Trénováním algoritmu na těchto datech, tj. maximalizací míry *F-rate*, pouze zvyšujeme hodnotu volného parametru, díky čemuž přestane detektor zachycovat některé méně významné změny mluvčích a tím roste zase počet chyb typu delece. Dalším důvodem snížené úspěšnosti detekce změn mluvčího je také skutečnost, že testování proběhlo bez předřadného detektoru řeč/neřeč, tj. algoritmy byly trénovány jako detektory změn mluvčího, testovány jako obecné detektory akustických změn a následně vyhodnoceny opět jako detektory změn mluvčího. V neposlední řadě svou roli sehrála i kvalita databáze COST 278, která byla zvláště v některých jejích národních částech poněkud sporná. S ohledem na praktické nasazení metod detekce změn si je rovněž zapotřebí uvědomit, že cílem této úlohy je rozdělení dlouhých záznamů řeči na kratší úseky, které mají z akustického hlediska homogenní charakter. Tato segmentace je nutná jak pro následnou identifikaci, tak i pro řádnou funkci rozpoznávače. Z tohoto hlediska nejsou navíc vložené body změny, tj. inzerce, takovým problémem, neboť nezpůsobují chybu při určování mluvčího a většinou ani nepoškodí činnost dekodéru spojité řeči. Opomenuté body změny, tj. delece, mají výrazně horší dopad na další zpracování.

Cesta ke zvýšení úspěšnosti detektoru změn mluvčího na reálných datech pravděpodobně povede přes nalezení robustnějších příznaků a tvorbu komplexnějších rozhodovacích systémů, kde bude součástí detektoru i např. detektor řeč/neřeč či modul identifikace řečníka.

Shrnutí přínosů k rozvoji vědního oboru

V práci je

- podán *jednotný výklad základních přístupů* k úloze detekce změny mluvčího v audio signálu na základě testování hypotéz o změně parametrů gaussovského procesu;
- provedeno *srovnání* běžně používaného přístupu *s alternativními přístupy*, jež plynou z teoretického rozboru;
- ověřena vyšší *spolehlivost autorem navržených alternativních přístupů*;

- potvrzena možnost aplikace *metody binárního dělení* na úlohu detekce vícenásobné změny mluvčího;
- navržena nová *metoda globální maximalizace BIC*;
- navrženo významné vylepšení on-line *metody s adaptivním oknem*;
- u všech metod popsán *způsob trénování* a způsob jejich *efektivní implementace*;
- provedeno *vyhodnocení úspěšnosti* na rozsáhlých databázích jak uměle připravených, tak i reálných vícejazyčných řečových záznamů, přičemž všechny navržené metody jsou buď lepší nebo přinejmenším stejně účinné jako metody, jež užívají partneři spolupracující v rámci projektu COST 278.

Shrnutí přínosů pro praxi

Všechny v práci navržené metody byly postupně testovány v systému pro automatický přepis televizního zpravodajství, vyvíjeného na TUL. Detektor založený na modifikované metodě s adaptivním oknem v současné době tvoří nedílnou součást tohoto transkripčního systému a umožňuje jeho plnou automatizaci. Jeho implementací a zařazením do procesu automatického přepisu odpadla jedna z nejnamáhavějších činností, jíž je ruční segmentace záznamů. Tím, že je segmentace uskutečňována na základě změn mluvčích, mohly být do systému zařazeny také moduly pro automatickou identifikaci řečníka a adaptaci na charakteristiky jeho hlasu.

Literatura

Citovaná literatura

[Chen J., 2000] Chen, J., Gupta, A. K., *Parametric Statistical Change Point Analysis*, Birkäuser, Boston, 2000.

[Chen & kol., 1998] Chen, S. S., Gopalakrishnan, P. S., *Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion*, IBM T.J. Watson Research Center, Yorktown Heights, NY, Technical Report, 1998.

[Chickering & Heckerman, 1996] Chickering, D. M., Heckerman, D., *Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables*, Technical report MSR-TR-96-08, Microsoft Research, 1996.

[Horváth, 1993] Horváth, L., *The Maximum Likelihood Method for Testing Changes in the Parameters of Normal Observations*, Annals of Statistics, Vol. 21, No 2., pp. 671-680,1993.

[Huang & kol., 2001] Huang, X., Acero, A., Hon, H., *Spoken Language Processing*, Prentice Hall PTR, 2001.

[Lehmann, 1986] Lehmann, E. L., *Testing Statistical Hypotheses*, 2^{nd} edition, Wiley & Sons, New York, 1986.

[Schwarz, 1978] Schwarz, G., *Estimating the Dimension of a Model*, Annals of Statistics, Vol. 6, pp. 461–464, 1978.

[Vostrikova, 1981] Vostrikova, L. Ju., *Detecting Disorder in Multidimensional Random Processes*, Soviet Mathematics Doklady, 1981.

Seznam vlastních prací

[1] Zdansky, J., Nouza, J., *Detection of Acoustic Change-Points in Audio Records via Global BIC Maximization and Dynamic Programming*, In Proceedings of 9^{th} International Conference on Speech Communications and Technology Interspeech 2005, pp. 669–672, Lisboa (Portugal), 2005.

[2] Nouza, J., Zdansky, J., David, P., Cerva, P., Kolorenc, J., Nejedlova, D., *Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon*, In Proceedings of 9^{th} International Conference on Speech Communications and Technology Interspeech 2005, pp. 1681–1684, Lisboa (Portugal), 2005.

[3] Žibert J., Mihelič F., Martens J. P., Meinedo H., Neto J., Docio L., Garcia-Mateo C., David P., Zdansky J., Pleva M., Cizmar A., Žgank A., Kačič Z., Teleki C., Vicsi K., *The COST 278 Broadcast News Segmentation and Speaker Clustering Evaluation* -

Overview, Methodology, Systems, Results, In proceedings of 9^{th} International Conference on Speech Communications and Technology Interspeech 2005, pp. 629–632, Lisboa (Portugal), 2005.

[4] Nouza, J., Červa, P., Žďánský, J., Kolorenč, J., David, P., *Towards Automatic Transcription of Parliament Speech*, In proc. of 16^{th} Conference on Electronic Speech Signal Processing, pp. 237–244, Prague, September 2005.

[5] Zdansky, J., *Detection of Acoustic Change-Points in Audio Streams and Signal Segmentation*, Radioengineering, Vol. 14, No. 1, pp. 37–40, April 2005.

[6] Zdansky, J., *Novel Algorithm for Speaker Segmentation of TV Broadcast News*, In Proc. of Radioelektronika 2005, April 2005, Brno, Czech Republic.

[7] Zdansky, J., David, P., Nouza, J., *An Improved Preprocessor for the Automatic Transcription of Broadcast News Audio Stream*, In Proc. of ICSLP 2004, pp. 1065–1068, Jeju (South Korea), October 2004.

[8] Nouza, J., Nejedlova, D., Zdansky, J., Kolorenc, J., *Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast*, In Proc. of ICSLP 2004, pp. 409–412, Jeju (South Korea), October 2004.

[9] Nouza, J., Zdansky, J., David, P., *Fully Automated Approach to Broadcast News Transcription in Czech Language*, In Proc. of Text, Speech and Dialogue, pp. 401–408, Springer-Verlag, Berlin, 2004.

[10] Zdansky, J., Kroul, M., *Semi-Automatic Non-speech Events Database Formation*, In 8th International Student Conference on Electrical Engineering - POSTER 2004 [CD-ROM], Prague, May 2004.

[11] Zdansky, J., David, P., *Automatic Audio Segmentation of Tv Broadcast News*, In Proc. of Radioelektronika 2004, pp. 358-361, Bratislava (Slovak Republic), April 2004.

[12] Zdansky, J., Nouza, J., *Experimental Optimization of the Continuous Speech Recognition System*, In Proc. of 13^{th} Czech-German Workshop „Speech Processing“, pp. 129-134, Prague, September 2003.

[13] Zdansky, J., *Přínos apriorní informace o dialektu pro HMM systémy rozpoznávání češtiny*, Analýza a zpracování signálů IV, Sborník seminářů katedry teorie obvodů, Praha, březen 2003.

[14] Zdansky, J., *Gender dependency of mel-frequency cepstral coefficients*, In 7^{th} International Student Conference on Electrical Engineering - POSTER 2003 [CD-ROM], Prague, May 2003.

[15] Zdansky, J., *Optimalizace struktury HMM*, Analýza a zpracování signálů III, Sborník seminářů katedry teorie obvodů, ISBN 80-01-02726, Praha, 2003.

Ing. Jindřich Žďánský

Metody detekce změny mluvčího v akustickém signálu
Autoreferát disertační práce

Technická univerzita v Liberci
Fakulta mechatroniky a mezioborových inženýrských studií

Náklad 20 výtisků

listopad 2005