TECHNICKÁ UNIVERZITA V LIBERCI
www.tul.cz

Automatická strukturalizace počítačem přepsaných mluvených dokumentů z multimediálních archivů

Autoreferát disertační práce

Studijní program: P2612 – Elektrotechnika a informatika
Studijní obor: 2612V045 – Technická kybernetika

Autor práce: **Ing. Marek Boháč**
Školicí pracoviště: Ústav informačních technologií a elektroniky
Fakulta mechatroniky, informatiky a mezioborových studií
Technická univerzita v Liberci
Studentská 2/1402, Liberec 461 17
Vedoucí práce: prof. Ing. Jan Nouza, CSc.

počet stran: 92
počet ilustrací: 23
počet tabulek: 20
citavaných prací: 73

Liberec 2016

Obsah

1 Úvod a motivace **5**

2 Shrnutí aktuálního stavu problematiky **9**

3 Cíle práce **12**

4 Moduly zapojené do procesu strukturalizace dokumentu **13**
4.1 Strukturalizační elementy a jejich vazby 13
4.2 Nástroje zapojené do strukturalizačních schémat 15
4.2.1 Parametrizace vstupního signálu 15
4.2.2 Použitý LVCSR systém 15
4.2.3 Detekce řečové aktivity v nahrávce a diarizace dokumentu . . 16
4.2.4 Klasifikace řečových segmentů nahrávky 18
4.2.5 Identifikace mluvčího 19
4.2.6 Další parametrizace signálu 20
4.2.7 Dodatečné formátování textu 20
4.2.8 Doplnění interpunkce do přepisu 21

5 Schémata strukturalizace dokumentu **23**
5.1 Strukturalizace s izolovaným rozhodováním 23
5.2 Strukturalizace s kumulovaným rozhodováním 24
5.2.1 Vrstva I 24
5.2.2 Vrstva II 24
5.2.3 Vrstva III 26

6 Vybrané experimety **27**
6.1 Využití časované reference 27
6.2 Vyhodnocovací metriky 28
6.3 Porovnání použitých konfigurací LVCSR 29
6.4 Vyhodnocení doplnění čárkové interpunkce 30
6.5 Vybraná kritéria segmentace nahrávky 31

7 Závěr **32**
7.1 Výzkumné přínosy práce 32
7.2 Praktické přínosy práce 34
7.3 Návrhy budoucí práce 34

1 Úvod a motivace

Tato práce se zaměřuje na řešení komplexního problému jak strukturalizovat (tj. vhodně rozčlenit, textově i foneticky analyzovat a následně upravit) výstup systému pro automatické rozpoznávání řeči (ASR) tak, aby byl co nejčitelnější pro člověka a zároveň připravený pro efektivní strojové zpracování a vyhledávání. Motivací pro řešení tohoto problému byl výzkumný projekt podporovaný Ministerstvem kultury ČR, jehož cílem bylo přepsat mluvené dokumenty z archivu Českého a Československého rozhlasu a zpřístupnit je pro vyhledávání[1]. Vzhledem k rozsahu zpracovávané části archivu (213.000 dokumentů z let 1923 – 2014) bylo nutné navrhnout a zrealizovat takový postup a takové technologie, které by byly schopny zvládnout nejen obrovské množství dat, ale také specifické problémy související s různou kvalitou záznamů, s přítomností českého a slovenského jazyka v dokumentech, se střídajícími se mluvčími, s prokládáním řeči znělkami a písničkami či s hluky na pozadí řeči.

Pro tyto účely byly na Technické univerzitě v Liberci vyvinuty moduly zajišťující automatické rozpoznávání řeči, řečníka a jazyka, dále moduly umožňující segmentaci zvukové nahrávky a následnou klasifikaci těchto úseků do několika tříd, které zohledňují, zda se jedná o řeč (čistou, zašuměnou, telefonní apod.), nebo o neřečový úsek obsahující např. ticho, hluk nebo hudbu. Autor této práce se podílel na vývoji některých těchto modulů a zejména na jejich začleňování do funkčního celku. Řešil optimalizaci jejich činností tak, aby bylo dosaženo co nejvyšší přesnosti zpracování archivních dokumentů a zároveň co nejpřirozenějšího přístupu k vyhledávání v nich.

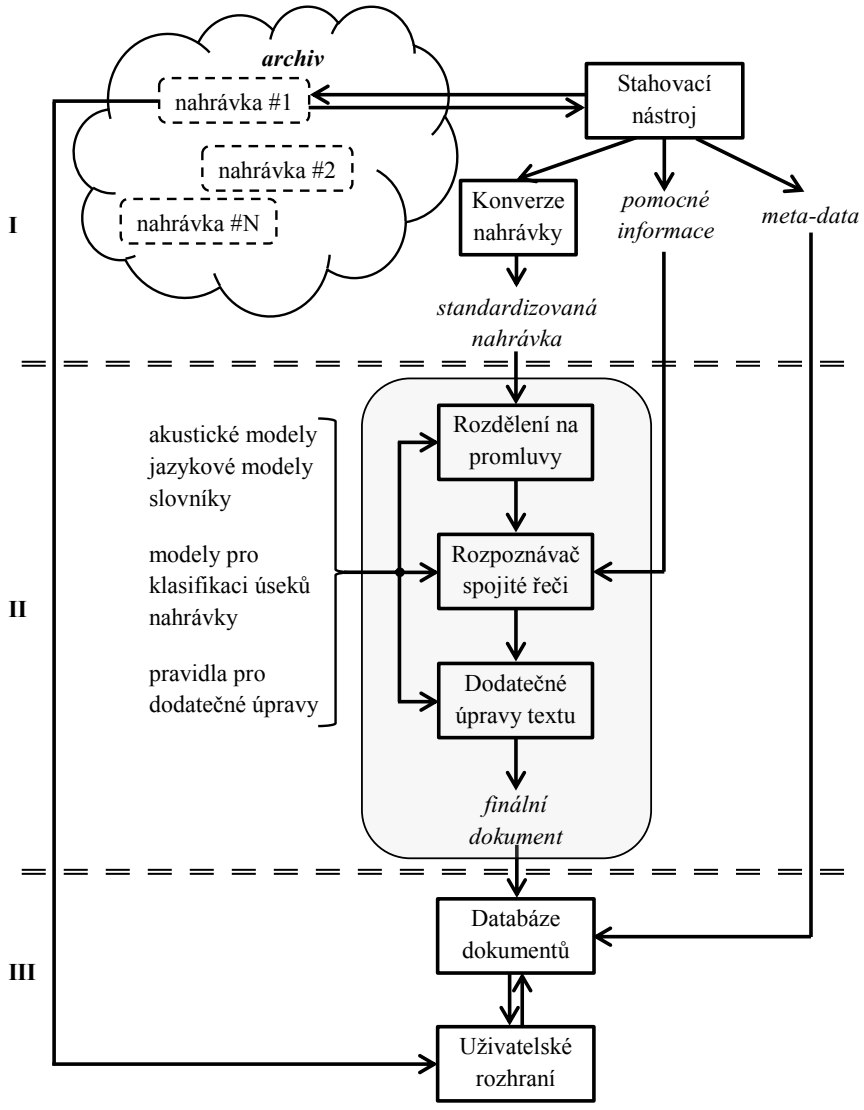
V práci jsou navržena, implementována a experimentálně ověřena dvě schémata (řetězce) zpracování mluveného dokumentu. Schémata jsou koncipována tak, abychom mohli porovnat dva odlišné přístupy k informacím, produkovaným dílčími nástroji. Na schématech také porovnáváme dva navržené moduly pro doplnění interpunkce a hodnotíme možnosti plynoucí z různých konfigurací ASR systému.

První navržené schéma provádí izolované rozhodování (každý krok strukturalizace využívá informaci získanou z jednoho konkrétního nástroje řetězce). Druhé schéma kumuluje rozhodování do vrstev, v nichž využívá všechny dostupné informační zdroje současně (čímž umožňuje vzájemnou verifikaci informačních zdrojů). Druhý zmíněný přístup umožňuje například zpřesnit přiřazení akustických a jazykových modelů rozpoznávače řeči z 87,96% na 91,82% (při použití stejných dílčích modulů). V otázce doplnění interpunkce proti sobě stavíme přístup vycházející z neřečových událostí v nahrávce a statistický popis délek větných celků.

[1] projekt Ministerstva kultury ČR: DF11P01OVV013; Zpřístupnění archivu Českého rozhlasu pro sofistikované vyhledávání

Abychom byli schopni výše zmíněné úlohy vyhodnotit, vytvořili jsme postup, který umožňuje automatické doplnění časových značek do referenčního přepisu. Současně navrhujeme vyhodnocovací nástroje, které vychází z takto časovaného referenčního přepisu, a umožňují tak podrobnější a časově efektivnější vyhodnocení stanovených metrik.

Základní schéma inventarizace (archivní) nahrávky je zachyceno na následující ilustraci (obr. 1.1). První vrstva inventarizace má za úkol standardizovat vstupní data a zajistit veškeré dostupné informace. Druhá vrstva provádí samotné zpracování dokumentu, k čemuž plní čtyři hlavní úkoly: 1) zajistit podmínky pro optimální funkci ASR (automatic speech recognition), 2) doplnit informace potřebné pro indexaci dokumentu a vyhledávání v databázi, 3) zjistit informace využité při zobrazení dokumentu, 4) optimalizovat čitelnost dokumentu a orientaci v něm. Třetí vrstva zpřístupňuje vytvořený archiv uživateli.



Obrázek 1.1: Základní schéma inventarizace archivní nahrávky

Zpracování archivních nahrávek se týká nejen historických archivů (digitalizace archivů Československého rozhlasu - ČRo a Československé televize byly zahájeny od roku 2003). Vznikají i tzv. archivy "paměti" - rozhovory s pamětníky významných událostí (např. projekt MALACH [1, 2] zaměřený na události holocaustu, nebo projekt Paměti národa, který mapuje české dějiny 20. století). Současně lze uvažovat i o zpracování moderních archivů - internet je zdrojem obrovského množství multimediálních dat, stejně jako komerční sféra (call-centra) a oblast bezpečnosti.

Většina zpracovaných pořadů má charakter hlavního zpravodajského pořadu dne. Obsahují proto promluvy řady různých mluvčích, vstupy nejen ze studia, ale i z terénu, telefonní vstupy či ilustrační záznamy projevů. Kromě toho se v pořadech vyskytují různé typy neřečového obsahu (znělky, gongy a různé typy hudby). Kromě zpravodajských pořadů jsou součástí archivu i významné projevy (např. novoroční projevy prezidentů), některé diskuzní pořady a určité množství pořadů populárně naučných. Nahrávky z období před rokem 1993 obsahují i různé množství slovenštiny. Pořady obsahují čtenou, připravenou i zcela spontánní řeč. V nahrávkách se vyskytují promluvy vysoce školených hlasatelů, méně školených řečníků (politici, vědci, umělci) i mluvčích zcela neškolených (účastníci anket, hosté). V datech se prakticky nevyskytuje emocionální řeč (jako např. v MALACHu).

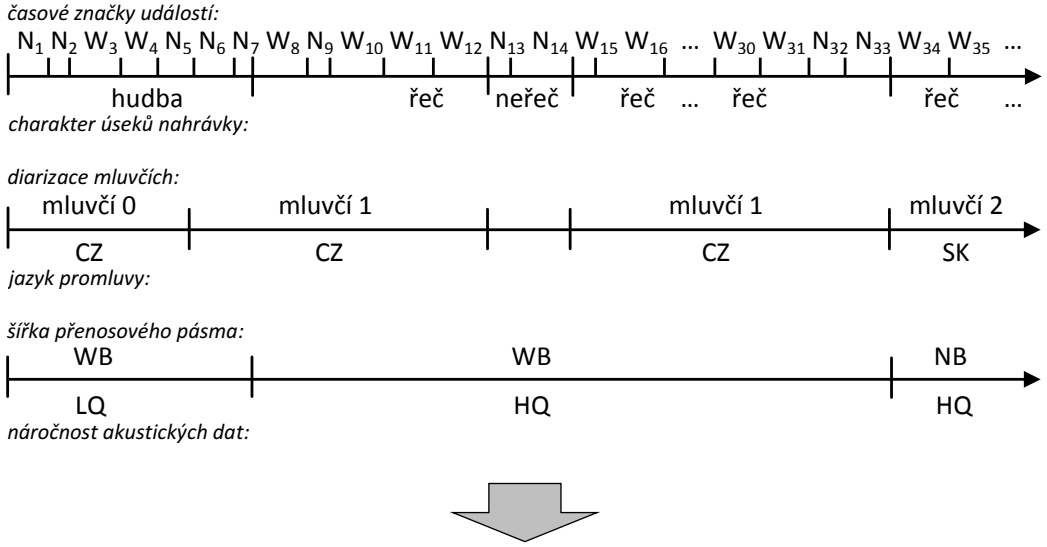
Zpracované nahrávky pochází ze dvou zdrojů. Nejdůležitějším je historický archiv ČRo. Nahrávky v něm obsažené byly před digitalizací uloženy na nejrůznějších analogových médiích (např. fonografové válce, magnetické pásky) a vytvořeny širokou škálou nahrávacích zařízení. Nahrávky byly později digitalizovány – uloženy na kompaktní disky. Digitalizované nahrávky jsou opatřeny popisky (jejichž obsah je velice různorodý). Druhým zdrojem je iRádio – internetový archiv soudobých pořadů. Ze struktury jeho webových stránek lze získat řadu informací, včetně stručných popisů obsahu pořadu. Nahrávky zpřístupněné iRádiem jsou obvykle ve formátu MP3, který není pro zpracování řeči optimální (komprimace zasahuje nevhodným způsobem do přenosového pásma řeči).

Projekt, který budeme označovat *NAKI*[2], si vytyčil poměrně ambiciózní cíle. Samotný rozsah zpracovaných dat (100.000 hodin) patří mezi největší automaticky zpracované archivy. Ambiciózní jsou i požadované vlastnosti výsledných přepisů. Systém musí být schopen detekovat jazyk promluvy (češtinu *CZ*, nebo slovenštinu *SK*), přičemž situaci výrazně komplikují rodilí mluvčí jednoho jazyka hovořící druhým jazykem. Nahrávka má být správně strukturalizována a pro každý segment má být určen vhodný akustický model: plné přenosové pásmo (*WB* – wide band; např. studiové nahrávky), nebo úzké přenosové pásmo (*NB* – narrow band; např. telefonní vstupy, některé typy mikrofonů). Systém dále musí určit totožnost mluvčích (pokud je pro daného mluvčího vytvořen model), nebo alespoň jeho pohlaví (muž *M*, žena *F*, neznámé *X*). Rozpoznaný text je nakonec upraven a strukturován tak, aby byl co nejlépe čitelný (post-processing a doplnění interpunkce). Oba jazyky a jejich historický vývoj kladou poměrně velké nároky na ASR systém, který musí operovat s velkými slovníky a adaptovat jazykové modely podle období vzniku nahrávky.

[2]naki.ite.tul.cz

nezpracovaný výstup systému pro rozpoznání řeči:

[hluk][hluk] rozhlasové noviny [hluk][ticho][nádech] dobrý večer [ticho] vysíláme rozhlasové noviny [nádech][hluk] k dodávce pivovarnického zařízení do sovětského svazu [nádech] tedy hovoří z Moskvy náš stálý zpravodaj [nádech] Ladislav Adamovič [ticho][nádech] druhého marca podpísali v Moskve dohodu o dodávke nášho strojného zariadenia pre desať kompletných pivovarov do sovietskeho zväzu [ticho][nádech] za náš technoexport podpísal túto dohodu námestník generálneho riaditeľa [hluk] súdruh František Samik [ticho][hluk][hluk][hluk] ...



strukturalizovaný dokument:

znělka [CZ,WB,X,LQ,hide] 0:00:00,0 : 0:00:05,7	Rozhlasové noviny
hlasatelka [CZ,WB,F,HQ,show] 0:00:05,7 : 0:00:14,2	Dobrý večer. Vysíláme rozhlasové noviny. K dodávce pivovarnického zařízení do Sovětského svazu tedy hovoří z Moskvy náš stálý zpravodaj Ladislav Adamovič.
Ladislav Adamovič [SK,NB,M,HQ,show] 0:00:14,2 : 0:00:39,3	Druhého marca podpísali v Moskve dohodu o dodávke nášho strojného zariadenia pre 10 kompletných pivovarov do Sovietskeho zväzu. Za náš TechnoExport podpísal túto dohodu námestník generálneho riaditeľa, súdruh František Samik.

Obrázek 1.2: Ilustrace vstupů a výstupu strukturalizace mluveného dokumentu

2 Shrnutí aktuálního stavu problematiky

V této kapitole stručně popíšu existující systémy vytvořené pro rozpoznání a zpřístupnění archivních nahrávek. Velmi stručně zmíním také nástroje počítačového zpracování řeči potřebné pro sestavení strukturalizačních schémat a řetězců, se kterými naše výsledky porovnáváme. V následujícím textu stručně popíšu systémy, které pro zjištění textového obsahu nahrávek používají systém rozpoznání řeči. Existují i systémy, které využívají existující textové přepisy dokumentů, jejich popis není pro tento autoreferát důležitý.

SpeechFind [3, 4] je systém určený ke zpřístupnění National Gallery of the Spoken Word[1] – archivu obsahujícího anglické nahrávky pořízené v průběhu 20. století (politické projevy a debaty, záznamy rozhlasového a televizního vysílání, přenosy NASA). Jedná se o velmi heterogenní směs pořadů a lze u nich předpokládat postupný vývoj jazyka (tím i potřebných jazykových modelů). Systém lze rozdělit do tří vrstev (které nalezneme i u ostatních představených systémů):

- inventarizace nahrávek, získání pomocných dat a meta-dat,
- segmentace nahrávky a rozpoznání ASR systémem,
- uložení dokumentů do databáze a propojení s uživatelským rozhraním.

Inventarizace nahrávek plní tři úlohy: 1) stáhnout nahrávku a standardizovat ji, 2) získat meta-data pro indexaci (původ nahrávky, datum vzniku atd.), 3) získat pomocná data – např. slova mimo slovní zásobu LVCSR (large vocabulary continuous speech recognition) systému, jména mluvčích.

Segmentační nástroj má za úkol detekovat tři typy změn v nahrávce: 1) změnu mluvčího, 2) změnu vlastností přenosového pásma a 3) změnu hlukových podmínek na pozadí řeči. K tomu využili autoři SpeechFindu velmi bohatou směs příznakových vektorů (PMVDR [5], SZCR, logaritomované koeficienty bank filtrů - FBLC). U některých předpokládají schopnost detekovat změny mluvčích, zatímco jiné mají analyzovat spíše pozadí řeči a přenosové pásmo. Jako měřítko podobnosti dvou segmentů používají Bayessovské informační kritérium (BIC).

SpeechFind používá GMM-HMM rozpoznávač řeči Sphinx 3 s akustickým modelem trénovaným na 200 h nahrávek. Systém dosahuje WER 25-40 % při méně než 1,5 % slov mimo slovní zásobu (OOV – out of vocabulary). V dostupných popisech systému je důraz kladen na využití meta-dat při vyhledávání.

[1] http://www.ngsw.org

MALACH je projekt zaměřený na zpřístupnění rozsáhlé sbírky rozhovorů s pamětníky holocaustu (pořízené Shoah Visual History Foundation[2]). Sbírka obsahuje cca 52.000 rozhovorů (celkem 116.000 hodin) ve 32 jazycích. Je důležité zmínit, že pamětníci jsou poměrně staří (a vzhledem k jejich životním osudům hovoří často se silným přízvukem). Řeč bývá emocionální a obsahuje proto nadprůměrné množství různých nespojitostí (váhání, opakování se, pláč). Výhodu představují informace obsažené v protokolu o nahrávce (lze v nich najít např. vlastní jména). Nahrávání bylo prováděno pomocí dvou mikrofonů, což umožňuje oddělit stopu s nahrávkou dotazovaného od promluv tazatele.

Naše schémata lze porovnat s verzí MALACHu určenou pro zpracování anglických a českých nahrávek [2] a s řešením pro maďarštinu [1]. Všechny tyto systémy provedou nejprve detekci řečové aktivity a poté rozpoznání s adaptací na mluvčího. Po rozpoznání je provedeno rozdělení na věty a určení tématu promluvy (viz [6, 7]), na jehož základě je určena konečná segmentace nahrávky. Akustické modely pro angličtinu bylo natrénovány na 200 h nahrávek, pro češtinu na 84 h nahrávek a pro maďarštinu bylo připraveno 26 h nahrávek. Systémy používají různou parametrizaci (různé nastavení GMM a PLP příznaků), dosahují však obdobné přesnosti přepisu – cca 40 % WER (word error rate) při přibližně 8 % slov mimo slovní zásobu.

Systém **InForMedia** [8] je určen k monitoringu anglofonních médií – rádia a televize. Využívá rozpoznávač řeči Sphinx 2 (v GMM-HMM konfiguraci, MFCC parametrizace s 1. a 2. diferencí). Segmentace je zjednodušena na vyhledání regionu s nízkou energií v nahrávce (ticha) a následnou dělbu nahrávky na cca 30 s dlouhé segmenty (tato délka je optimální pro prezentaci výsledků v uživateslkém rozhraní).

SPRACH [9] je systém určený ke zpracování nahrávek anglicky mluvených zpravodajských pořadů. Systém nejprve provádí segmentaci nahrávky, následně rozpozná jednotlivé úseky pomocí několika různých akustických modelů a nakonec přepisy sloučí do jedné výsledné hypotézy. Pro segmentaci je použit komplexní nástroj vyvinutý na Cambridge University [10, 11], který provádí následující kroky: segmentace nahrávky, vyřazení úseků obsahujících hudbu, první průchod ASR systémem a určení pohlaví mluvčího, vyřazení dlouhých úseků ticha a vyhlazení segmentace.

Po provedení segmentace nahrávky jsou řečové úseky v náhrávce rozpoznány ASR systémy ve třech odlišných konfiguracích (CI-RNN-HMM, CD-RNN-HMM a CI-MLP-HMM). Všechny tři systémy mají akustické modely trénovány na stejné sadě 200 hodin nahrávek a produkují výstup ve formě tzv. lattice. Konečný přepis je určen sloučením hypotéz systémem ROVER [12].

Výše zmíněné systémy byly navrženy okolo roku 2000. V té době byl k dispozici dostatečný výpočetní výkon pro zpracování rozsáhlých archivů mluveného slova. Současně dosáhly technologie zpracování řeči potřebné přesnosti. Až na výjimky byly tyto technologie nasazeny na anglická data (jejichž nároky na velikost slovní zásoby jsou podstatně menší než u slovanských jazyků). Nárůst výpočetního výkonu v době mezi vznikem výše zvýšených systémů a prací popsanou v tomto autoreferátu umožnil zpracování velmi rozsáhlého archivu mluveného slova (pro jazyky s velkou slovní zásobou).

[2] https://sfi.usc.edu

Komplexní nástroje, které slouží k počítačovému zpracování řeči, si zasluhují mnohem obsáhlejší a detailnější popis, než jaký je možno vtěstnat do tohoto autoreferátu. Zaměřím se proto na klíčové rozdíly, kterými se odlišují systémy popsané v přehledu současného stavu problematiky od námi navržených řešení.

Klíčovým nástrojem všech představených systémů je rozpoznávač spojité řeči. Statistický přístup k úloze rozpoznání řeči spoléhá na kombinaci akustického procesoru a lingvistického dekodéru [13]. Úkolem Viterbiho dekodéru je pak najít takovou posloupnost slov ($W = \{w_1, w_2, ..., w_N\}$), která s největší aposteriorní pravděpodobností odpovídá akustické informaci ($\mathbf{O} = \{\mathbf{o_1}, \mathbf{o_2}, ..., \mathbf{o_T}\}$, kde $\mathbf{o_i}$ značí příznakový vektor konkrétního framu) a apriorní pravděpodobnosti výskytu konkrétní posloupnosti slov $\hat{W}$ čili $P(W|\mathbf{O})$. Tento vztah akustického procesoru a jazykové složky ASR lze rozepsat pomocí Bayesova vzorce (2.1)

$$\hat{W} = \arg\max_W P(W|\mathbf{O}) = \arg\max_W \frac{P(W)P(\mathbf{O}|W)}{P(\mathbf{O})} \tag{2.1}$$

kde $P(\mathbf{O}|W)$ značí pravděpodobnost, že posloupnost slov W vygeneruje posloupnost příznakových vektorů $\mathbf{O}$, $P(W)$ značí pravděpodobnost, že byla pronese posloupnost slov W a $P(\mathbf{O})$ značí pravděpodobnost výskytu série příznakových vektorů $\mathbf{O}$. Protože $P(\mathbf{O})$ není funkcí W, redukuje se hledání maxima na rovnici (2.2).

$$\hat{W} = \arg\max_W P(W)P(\mathbf{O}|W) \tag{2.2}$$

Viterbiho dekodér tedy hledá maximum součinu dvou členů: $P(W)$, který je dán jazykovým modelem, a $P(\mathbf{O}|W)$, který reprezentuje akustický model.

Zásadní odlišnosti mezi systémy rozpoznání řeči spočívají v jejich akustickém dekodéru. Výkon rozpoznávače je výrazně ovlivněn implementací Viterbiho dekodéru který (na základě výstupu akustického dekodéru, slovníku a jazykového modelu) stanoví nejpravděpodobnější hypotézu o obsahu nahrávky.

Cílem akustického dekodéru je parametrizovat vstupní signál a přiřadit framům (nejkratším analyzovaným úsekům) signálu hypotézu o jejich obsahu. Řeč je modelována jako skrytý markovský proces (HMM – hidden markov model). Akustický dekodér popisuje obsah framu buď jako gaussovskou směs (GMM – gaussian mixture model) příznaků, kterou následně porovnává s modely, nebo může být nahrazeno neuronovou sítí (do níž vstupuje parametrizovaný signál a výstup sítě je hypotézou o obsahu signálu). Obsah nahrávky lze chápat buď jako vzájemně nezávislou posloupnost fonémů (CI – context independent), nebo předpokládáme, že se fonémy vzájemně ovlivňují (CD – context dependent). Mezi nejčastější parametrizace signálu patří mel-frekvenční kepstrální koeficienty (MFCC), banky filtrů [14], nebo tzv. bottle-neck příznaky (parametrizaci provádí neuronová síť). V předchozím textu byly zmíněny dva typy neuronových sítí. RNN (recurrent neural network) je síť obsahující zpětné vazby mezi vrstvami a MLP (multi-layer perceptron) je síť s jednou skrytou vrstvou. Trénování složitějších struktur nebylo v minulosti prakticky možné kvůli nedostupnosti dostatečného výpočetního výkonu. Dnes se nejčastěji využívají DNN (deep neutal network), což jsou neuronové sítě s více skrytými vrstvami (obvykle okolo pěti skrytých vrstev).

3 Cíle práce

Tato práce se zaměřuje na návrh strukturalizačních schémat umožňujících automatické zpracování rozsáhlých archivů mluveného slova. Doposud realizovaná řešení se soustředí (a omezují) na tyto dva kroky:

1. zajistit podmínky pro správnou funkci systému rozpoznání řeči (určit vhodné akustické a jazykové modely a slovníky),
2. extrahovat z přepisu informace nutné pro indexaci a vyhledávání.

Přepisy získané takovými systémy lze přirovnat k automaticky vytvořeným titulkům. Od našich přepisů vyžadujeme bohatší informační obsah – označení identity mluvčího, jazyka promluvy, charakteru přenosového pásma, případně klasifikaci neřečových regionů v nahrávce. Všechny tyto informace chceme ve výsledném dokumentu přehledně zobrazit. Proto jsou do našich schémat zahrnuty další dva úkoly:

3. zajistit informace pro správné zobrazení dokumentu,
4. optimalizovat čitelnost přepisu a orientaci v něm.

Hlavní cíle práce lze shrnout v následujících bodech:

- navrhnout schémata strukturalizace nahrávky,
- definovat elementy pro strukturalizaci přepisu, jejich vzájemnou hierarchii a vazbu na nahrávku,
- navrhnout moduly pro členění textového přepisu, včetně možností automatického doplnění interpunkce,
- připravit dostatečně rozsáhlou a různorodou sadu testovacích dat, která umožní vyhodnotit přesnost získaných přepisů, detekci bodů změny v nahrávce, doplněnou interpunkci, správnost modelů přiřazených systému rozpoznání řeči a vzájemných vlivů jednotlivých nástrojů použitých ke strukturalizaci,
- navrhnout vyhodnocovací metriky a vytvořit nástroje pro jejich vyčíslení,
- porovnat výkonnost dostupných konfigurací rozpoznávače řeči v rámci vytvořených strukturalizačních schémat,
- porovnat výsledky dosažené navrženými schématy,
- připravit navržené postupy a nástroje k reálnému nasazení a vyhodnotit poznatky z reálného provozu.

4 Moduly zapojené do procesu strukturalizace dokumentu

V následující kapitole popišu vytvořená schémata strukturalizace automatického přepisu archivní nahrávky. Do schémat je zakomponována řada dílčích nástrojů, jejichž vzájemné závislosti (požadavky na vstupní data) do určité míry předurčují jejich vzájemnou pozici ve strukturalizačním schématu. V dalším textu popíšu klíčové vlastnosti těchto nástrojů a také strukturalizační elementy, na něž jsou výsledky nástrojů vázány. Řada zmíněných nástrojů je výsledkem činnosti kolektivu Laboratoře počítačového zpracování řeči (která probíhá již cca 15 let a účastnily se jí desítky pracovníků). Vzhledem k velmi omezenému rozsahu tohoto autoreferátu bude popis těchto nástrojů velice stručný (podrobnosti jsou uvedeny v odkazované literatuře). Nástrojům vytvořeným výhradně pro potřeby strukturalizačních schémat bude věnována o něco větší pozornost.

4.1 Strukturalizační elementy a jejich vazby

Proces strukturalizace dokumentu musí vyhledat a zohlednit vazby mezi zvukovou nahrávkou dokumentu, přepisem dokumentu pořízeným LVCSR systémem (systémem rozpoznání řeči pracujícím s rozsáhlou slovní zásobou) a výsledným strukturalizovaným dokumentem. Ve všech třech úrovních lze dokument hierarchicky rozdělit na nižší celky, které si však nejsou vzájemně ekvivalentní. Nejprve popíšu vazby mezi nahrávkou dokumentu (vstupním číslicovým signálem) a výstupem LVCSR systému. Následně popíšu vazby mezi elementy konečného strukturalizovaného dokumentu a výstupem LVCSR systému (s jeho výstupem jsou synchronizovány i ostatní nástroje operující nad nahrávkou dokumentu).

Nejkratším elementem nahrávky (digitalizovaného signálu), se kterým pracujeme, je jeden *frame*. Jeho délka a posun určují časové rozlišení lokalizace událostí v nahrávce (potažmo ve výsledném dokumentu). Hierarchicky výše je postaven *segment* nahrávky. Segmentem rozumíme homogenní úsek nahrávky (respektive vstupního signálu), přičemž homogenita může být určena na více úrovních. Základním kritériem homogenity je dělba segmentů na řečové-neřečové, dále lze rozlišovat jazyk segmentu, charakter přenosového pásma a pohlaví, či identitu mluvčího. Nejvyšším celkem na úrovni vstupního signálu je celá *nahrávka* dokumentu.

Na úrovni výstupu LVCSR systému je nejnižším elementem přepisu *událost*. Rozlišujeme události dvou charakterů: řečová událost (rozpoznaná slova či fráze, ale

i součásti foneticko-akustického inventáře spojené s tvorbou řeči – nádech, váhavý zvuk) a neřečová událost (zbylé položky foneticko-akustického inventáře – hluky modelující hudbu, kašel apod.). Posloupnost událostí detekovaných LVCSR systémem v konkrétním rozsahu časových značek vstupního signálu lze postavit na úroveň segmentu signálu. Celý výstup LVCSR systému odpovídá dokumentu.

Výsledný strukturalizovaný dokument rozlišuje jako nejmenší nedělitelnou jednotku *slovo* (případně jmennou či číselnou entitu). Časové značky slova jsou extrahovány z řečových událostí, kterým odpovídají – mají proto rozlišení jednoho framu. Nadřazeným elementem slova je *věta*. V této práci není věta chápána striktně lingvisticky, spíše ji můžeme popsat jako sérii slov ukončenou interpunkčním znaménkem (v našem případě tečkou nebo čárkou). Obecně nadřazeným (i když potenciálně totožným) elementem je *promluva* – homogenní řečový projev jednoho řečníka (homogenní ve smyslu jazyka a charakteru přenosového pásma). Promluvu lze proto položit na úroveň segmentu vstupní nahrávky (s danou úrovní homogenity). Na promluvu jsou vázány všechny klasifikace nahrávky (jazyk promluvy, identita mluvčího, charakter přenosového pásma). Nejvyšší úrovní je celý přepis nahrávky.

Datový kontejner pro uložení finálního dokumentu získanou informaci částečně redukuje. Jako nejnižší element chápe *frázi* – nejkratší řečovou událost opatřenou časovými značkami (odpovídá slovu, číselné entitě, nebo částem jmenné entity). Sada frází tvoří *paragraf* přepisu (odpovídá promluvě). Proto jsou na paragraf vázány všechny informace o promluvě. *Kapitola* odpovídá celému dokumentu.

Všechny významné jevy v nahrávce (změny řečníka, změny atributů promluv stejně jako přítomnost interpunkce) lze lokalizovat do jedné společné sady časových značek. Tato sada časových značek jsou začátky (a konce) řečových událostí. V dalším textu budou označovány jako *sloty*.

Vzájemný vztah elementů definujících výsledný strukturalizovaný dokument je zachycen na následujícím obrázku (obr. 4.1).

číslicový signál	LVCSR přepis	strukturalizovaný dokument	kontejner dat
frame			
	řečová událost neřečová událost	slovo číselná entita jmenná entita	fráze
		věta	
segment	úsek událostí	promluva neřečový segment	paragraf
nahrávka	všechny události	dokument	kapitola

Obrázek 4.1: Elementy zapojené do tvorby strukturalizovaného dokumentu

4.2 Nástroje zapojené do strukturalizačních schémat

4.2.1 Parametrizace vstupního signálu

Modul parametrizace signálu přiřazuje vstupnímu signálu (nahraný se vzorkovací frekvencí 16 kHz) jeho reprezentaci pomocí zvolené příznakové sady (kterou využívá systém rozpoznání řeči i další moduly). Jako nejměnší (dále nedělitelné) jednotky vstupního signálu jsou parametrizovány framy (úseky dlouhé 20 ms s překryvem 10 ms). Framy jsou popsány 39 MFCC příznaky (mel-frekvenční kepstrální koeficienty) - 13 příznaků a jejich první a druhá diference. Na příznakové vektory je aplikována normalizace odečtením střední hodnoty (cepstral mean subtraction - CMS) buď pro celou nahrávku, nebo v rámci plovoucího okna (volíme okno délky 2 s).

4.2.2 Použitý LVCSR systém

Klíčovým nástrojem pro zpracování archivních nahrávek je systém rozpoznání spojité řeči (v následujícím textu budeme používat zkratku LVCSR - large vocabulary continuous speech recognition). V prezentovaných systémech pracuji s LVCSR systémem vyvinutým na Ústavu informačních technologií a elektroniky FMMIS TUL [15]. LVCSR systém používáme ve dvou konfiguracích akustického dekodéru. První z nich je CD-GMM-HMM (v dalším textu označován *LVCSR-GMM*), druhou je CD-DNN-HMM (*LVCSR-DNN*). Pro konfiguraci LVCSR-GMM bylo vyvinuto rozšíření o adaptaci na mluvčího, pro LVCSR-DNN zatím adaptaci nepoužíváme, ačkoli je principiálně možná [16].

LVCSR-GMM konfigurace využívá 39 MFCC příznaků s CMS normalizací (jak je zmíněna v předchozí pasáži). Volitelnou funkcí dekodéru je adaptace na mluvčího. V našem případě se využívá automatická (unsupervised) adaptace, jejímž vstupem jsou úseky nahrávky, které podle diarizace patří stejnému mluvčímu (nebo mají shodné akustické podmínky, např. hluky z průmyslového závodu), a jejich přepis poskytnutý předchozím průchodem LVCSR systémem. Využíváme metodu Constrained Maximum Likelihood Linear Regression [17], odvozenou z Maximum Likelihood Linear Regression [18]. Její podstata spočívá v nalezení transformační matice, která převádí rozšířený příznakový vektor na adaptovaný příznakový vektor (lépe odpovídající akustickým modelům).

LVCSR-DNN využívá opět 39 MFCC příznaků, bere navíc v úvahu 5 framů před a za zkoumaným framem nahrávky. Síť má 5 skrytých vrstev o šířce 1024 neuronů, aktivační funkcí je sigmoida. Na výstupu sítě je určena přímo věrohodnost jednotlivých senonů (akustických stavů Viterbiho dekodéru). Modul pro adaptaci na mluvčího nebyl pro tuto konfiguraci implementován - předpokládá se dostatečná robustnost samotné sítě.

Pro obě konfigurace LVCSR systému byly připraveny akustické modely (AM) a jazykové (LM) modely a jim odpovídající slovníky (VOC). Konkrétně jsou použity tři jazykové varianty modelů: čeština (CZ), slovenština (SK) a kombinovaný model česko+slovenský (CZ+SK). Podle charakteristiky přenosového pásma rozlišujeme standardní akustický model (WB – wideband) a úzkopásmový akustický

Tabulka 4.1: Přehled velikosti slovníků, jazykových modelů, množství trénovacích dat pro akustické modely a konfigurací akustického dekodéru LVCSR systému

	CZ	SK	CZ+SK
WB	GMM i DNN 300 hod. CZ 550.000 slov	GMM i DNN 100 hod. SK 320.000 slov	GMM 100 hod. CZ + 100 hod. SK 50.000 CZ + 50.000 SK slov
NB	GMM i DNN cca 100 hod. CZ 550.000 slov CZ / 320.000 slov SK		—

model (NB – narrowband). Použité kombinace jsou zobrazeny v tabulce 4.1, spolu s množstvím trénovacích dat a velikostí slovníků. Modely LVCSR-GMM konfigurace jsou označeny GMM a LVCSR-DNN značíme DNN.

4.2.3 Detekce řečové aktivity v nahrávce a diarizace dokumentu

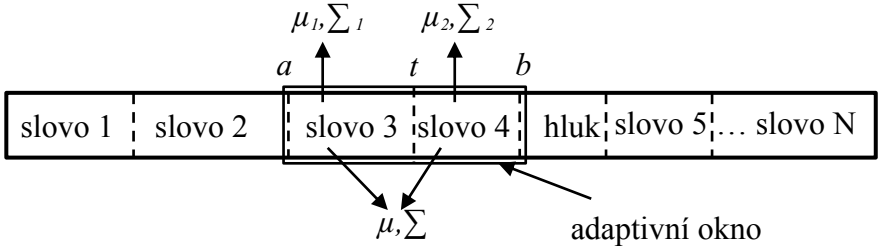
Nástroje určené k segmentaci nahrávky využívají jako vstupní informaci "předběžný" přepis dokumentu pořízený kombinovaným česko+slovenským jazykovým modelem (CZ+SK v předchozí sekci). Výše zmíněný model má limitovanou slovní zásobu, proto proběhne rozpoznání nahrávky rychle. Výstup rozpoznávače je současně dostatečně přesný, aby posloužil k lokalizaci neřečových událostí. Hranice řečových událostí definují sadu možných bodů změny v nahrávce.

Nalezení (ne)řečových regionů nahrávky

Extrakci (ne)řečových regionů v nahrávce lze zařadit buď před nebo za detekci bodů změny v nahrávce. Regiony s neřečovým obsahem jsou nalezeny podle podílu řečových a neřečových událostí ve zkoumaném úseku nahrávky. Za řečový obsah považujeme řečové události (slova, krátké fráze) a některé neřečové události související s tvorbou řeči (nádech, váhavý zvuk). Mezi neřečový obsah počítáme všechny zbylé typy hluků a dlouhé úseky ticha.

Podíl řečového a neřečového obsahu je počítán v rámci plovoucího okna fixní délky (čas b je první konec rozpoznané události, který má minimální požadovanou vzdálenost od času a). Je-li nalezen úsek nahrávky s dostatečně malým podílem řeči, jsou jeho hranice upřesněny pomocí druhého (kratšího) plovoucího okna. Délky posuvných oken i hraniční poměry řečového a neřečového obsahu byly určeny na základě experimentů s vývojovými daty [19].

Tento postup může vést k chybnému označení silně zašuměného segmentu za neřečový. Je-li ale odstup řeči od hluku na pozadí natolik malý, že většina úseku je rozpoznána jako neřečové události, případný přepis segmentu by byl pravděpodobně nepřesný a případná škoda je zanedbatelná. Obdobně se klasifikace může zachovat k hudebnímu segmentu, jehož vyloučení z dalšího zpracování je žádoucí.



Obrázek 4.2: Detekce změny mluvčího adaptivním oknem omezeným na hranice událostí detekovaných LVCSR systémem

Nalezení bodů změny v nahrávce

Detekce bodu změny mluvčího (respektive změny charakteru nahrávky) je provedena na základě porovnání podobnosti dvou sousedních úseků nahrávky (*a-t*;*t-b*), které využívá MFCC parametrizaci (již provedenou pro LVCSR systém) a primárně hledá změnu mluvčího. *a,t,b* představují popořadě začátek plovoucího okna, prověřovaný dělicí bod a konec plovoucího okna. Strategie pro postupnou adaptaci délky zkoumaného okna je detailně popsána v [20, 21]. Podobnost prověřovaných intervalů je vyhodnocena pomocí BIC zavedeného vztahy (4.1) a (4.2). Penalizační faktor P a práh pro přijetí hypotézy o přítomnosti dělicího bodu určují chování detektoru (jejich hodnoty jsou určeny experimenty na vývojových datech). N_1 a N_2 značí počet příznakových vektorů před a za prověřovaným dělicím bodem. $\sum$, $\sum_1$ a $\sum_2$ představují kovarianční matice příznakových vektorů v celém zkoumaném okně, před a za dělicím bodem. d je délka příznakového vektoru (39 MFCC) a α představuje penalizační koeficient (v našem případě volíme $\alpha = 1$).

$$BIC = (N_1 + N_2)log(|\Sigma|) - N_1 log(|\Sigma_1|) - N_2 log(|\Sigma_2|) - \alpha P \tag{4.1}$$

$$P = \frac{1}{2}((d + \frac{1}{2}(d(d+1)))log(N_1 + N_2) \tag{4.2}$$

Diarizace dokumentu

Diarizace nahrávky je komplexní proces, do něhož vstupuje nahrávka (respektive segmenty nahrávky získané dvěma předcházejícími nástroji) a na jeho výstupu jsou určeny segmenty pronesené stejnými mluvčími. Segmenty nahrávky jsou postupně shlukovány podle následujícího algoritmu:

1. výpočet podobnosti mezi všemi dvojicemi segmentů,
2. je-li podobnost příliš malá, ukončí se výpočet,
3. sloučení nejpodobnějšího páru segmentů,
4. přepočítání podobnosti v rámci nově definovaných shluků (skupin segmentů),
5. zpět na krok 2.

Shlukování je v našem případě hierarchické. To znamená, že jedna metrika podobnosti je použita k "předshlukování" segmentů a jiná metrika je použita k získání konečné sady shluků. V případě námi použitého systému [22] je k předshlukování segmentů použito *BIC* (stejně jako při hledání bodů změny mluvčího). Finální vrstva parametrizuje porovnávané segmenty pomocí i-vectorů a podobnost měří cosinovou vzdáleností. Výsledek diarizace lze pak využít buď jako kompletní diarizaci (pro rozpoznání s adaptací na mluvčího v rámci celého dokumentu), nebo lze informační hodnotu výstupu redukovat na detekci bodů změny v nahrávce (s redukovaným množstvím falešných bodů změny).

4.2.4 Klasifikace řečových segmentů nahrávky

Nástroje zmíněné v sekci 4.2.3 vyberou z nahrávky regiony obsahující řeč a rozdělí je na úseky pronesené jednotlivými mluvčími. Úkolem klasifikace těchto úseků je nalézt tzv. promluvy, které definujeme jako nejdelší nepřerušené úseky nahrávky pronesené jedním mluvčím, ve kterých se nemění další atributy nahrávky (šířka přenosového pásma a jazyk promluvy). Šířka přenosového pásma a jazyk promluvy jsou klíčové pro správný výběr modelů LVCSR systému (jazykový a akustický model, slovník).

Určení jazyka promluvy

Obvyklým postupem pro rozpoznání jazyka promluvy je natrénování akustického modelu, který pokrývá fonémovou sadu všech rozpoznávaných jazyků. Jazykový model pokrývá výskyt posloupností fonémů v jednotlivých jazycích. Určení jazyka pak optimalizuje pravděpodobnost výskytu posloupnosti fonémů v nahrávce (akustické informace) v pozorovaném kontextu (jazykový model), jak je shrnuto např. v [23]. Systém navržený na našem pracovišti [24] výše zmíněný postup rozšiřuje. Pro všechny rozpoznávané jazyky (v našem případě češtinu – CZ a slovenštinu – SK) je připraven společný akustický model a společný jazykový model. Jazykový model má k dispozici slovníky obou rozpoznávaných jazyků a pro ně vytvoří takový jazykový model, který jednak modeluje každý z dílčích jazyků, současně ale umožňuje přechody mezi nimi. Aby nedošlo ke zvýhodnění některého jazyka, jsou slovní zásoby limitovány. Jednotlivé slovníkové položky na sebe vážou informaci o tom, kterému jazyku přísluší, případně že se foneticky stejná položka vyskytuje v obou jazycích (COM). Modely tedy nehodnotí pouze akustickou informaci v určitém kontextu, ale kontext je delší a zapojuje do modelu vyšší celky jazyka (slova, fráze a jejich n-gramy).

Po rozdělení nahrávky na promluvy jednotlivých mluvčích se určí množství slov každého jazyka ve zkoumaném úseku (jak ilustruje obr. 4.3). Z vyhodnocení jsou vyloučena slova společná pro oba jazyky (značená COM). Jazyk s největším zastoupením v daném úseku je prohlášen za jazyk promluvy (viz řádek *Závěr*). Detaily o kombinovaném československém modelu jsou uvedeny v tabulce 4.1 (klíč CZ+SK). Slovníky použité v této práci operují s 50.000 slov pro každý jazyk a modul určení jazyka pracuje s přesností okolo 99 %.

Přepis:	Dobrý	deň	vitajte	u	správ.	Hlavní	novinou	dnešního	dne	je,	že …
Jazyk:	COM	SK	SK	COM	COM	CZ	COM	CZ	CZ	COM	COM
Závěr:	2xSK ; 3xCOM ; 0xCZ => **SK**					0xSK ; 3xCOM ; 3xCZ => **CZ**					

Obrázek 4.3: Ilustrace určení jazyka promluvy–čeština (CZ), slovenština (SK), slovo společné pro slovníky obou jazyků (COM)

Určení šířky přenosového pásma a pohlaví mluvčího

Obě klasifikace popsané v této pasáži využívají GMM modely s parametrizaci signálu shodnou s LVCSR systémem (sekce 4.2.1). Při určování šířky přenosového pásma chceme odlišit nahrávky vzniklé ve studiu (mají plné přenosové pásmo 16 kHz) od nahrávek, v jejichž nahrávacím řetězci se nachází zařízení s omezeným přenosovým pásmem (typicky okolo 8 kHz) jako například přenosné magnetofony, diktafony, telefonní linky.

Určení pohlaví mluvčího pak užívá GMM modely k rozlišení mužů, žen a "obecného hluku". Hluk je pojistkou pro situaci, kdy by klasifikovaný segment obsahoval hudbu nebo jiné neřečové události. Přítomnost dětských mluvčích nepředpokládáme.

4.2.5 Identifikace mluvčího

Modul identifikace mluvčího přiřazuje promluvě (úseku v němž předpokládáme jediného mluvčího) nejpravděpodobnějšího mluvčího ze sady dostupných modelů. Modely je možné "předtřídit" podle dříve určených kritérií - pohlaví mluvčího, šířky přenosového pásma a jazyka promluvy. Pro každého mluvčího mohou tedy existovat až čtyři modely (kombinace CZ/SK a NB/WB). Pro každý model bylo nalezeno minimálně 10 min trénovacích nahrávek. Detaily sběru dat pro modely jsou popsány v [25, 26].

Identifikace mluvčích je založena na tzv. "joint factor analysis", která je použita jako generátor příznaků popisujících trénovací sadu promluv mluvčích (a přenosových cest). Těmito příznaky je definován tzv. "total variability space" [27], ve kterém jsou promluvy reprezentovány s redukovanými rozměrem příznakového vektoru. Průmět zkoumané promluvy do tohoto prostoru (označovaný jako i-vector) slouží jako reprezentace promluvy, stejně jako reprezentace trénovacích dat. Podobnost zkoumané promluvy s trénovacími daty je určena pomocí cosinové vzdálenosti (4.3), kde x_1 a x_2 značí referenční a zkoumaný i-vector.

$$CDS = \frac{{x_1}' x_2}{\|x_1\| \, \|x_2\|} \tag{4.3}$$

Verifikace mluvčích, která je obvyklou součástí systémů identifikace mluvčího, je v našem případě zjednodušena. Skóre získané nejpravděpodobnějším mluvčím je porovnáno s bezpečnostním prahem a hypotézu o jeho identitě přijmeme, nebo zamítneme.

4.2.6 Další parametrizace signálu

Některé z následujících modulů využívají části prozodické informace, které je možné extrahovat ze zvukové nahrávky. Konkrétně aplikujeme krátkodobou energii signálu a fundamentální frekvenci řeči. Výpočty jsou provedeny nad úseky odpovídajícími řečovým událostem. Úseky obsahující neřečové události jsou označeny a nejsou dále parametrizovány. Krok a překryv je zvolen shodně s parametrizací signálu pro LVCSR (20ms framy s překryvem 10 ms).

U krátkodobé energie je každému slovu přiřazena průměrná hodnota energie $\bar{E}$ a normovaná diference energie E_{nd} (4.4), která vyjadřuje míru "kolísání" krátkodobé energie v průběhu slova.

$$E_{nd} = \frac{max(E) - min(E)}{\bar{E}} \tag{4.4}$$

Při určení fundamentální frekvence řeči ($F0$) využíváme dva hlavní zdroje apriorních informací. Prvním je lokalizace řečových událostí v nahrávce založená na výstupu LVCSR systému. Druhý předpoklad vychází ze známých mezních hodnot fundamentální frekvence řeči (muži 80–160 Hz, ženy 150–300 Hz a děti 200–600 Hz). Pro dospělé mluvčí tedy hledáme fundamentální frekvenci v rozsahu cca 60–400 Hz.

Pro určení fundamentální frekvence řeči existuje několik zavedených metod [28, 29]. První skupina metod využívá autokorelační funkci řečového signálu, druhá analyzuje signál v kepstrální oblasti a třetí pracuje se spektrogramem nahrávky. Metody zpracovávající spektrogram nahrávky jsou obecně považovány za robustnější vůči hlukům na pozadí řeči. Navržená metoda proto vychází z výpočtu STFT nad regiony řečové aktivity (určené výstupem LVCSR systému) po němž je dynamickým dekodérem určena fundamentální frekvence řeči v daném segmentu.

Konkrétně metoda pracuje s framy stejné délky, jako LVCSR systém, které doplňuje nulami. Výstup STFT je omezen na frekvence v rozsahu 60–600 Hz. V každém framu je nalezeno 5 lokálních maxim spektrogramu, mezi nimiž následně dekodér nalezne nejméně penalizovanou "cestu" - fundamentální frekvenci promluvy.

4.2.7 Dodatečné formátování textu

Primárním účelem dodatečného formátování textu (post-processingu) je zvýšení čitelnosti rozpoznaného textu (např. úpravou zápisu zápisu zkratek, titulů a číslovek). Současně může být post-processing zdrojem informace o tom, které rozpoznané řečové události tvoří společně jednu entitu (jmennou entitu, číselný údaj apod.). Textové úpravy jsou implementovány pomocí vážených stavových automatů (WFST – weighted finite state transducers) a jsou rozděleny do série vrstev (jak shrnuje tab. 4.2).

Vrstva odstranění hluků odstraňuje z přepisu tagy neřečových událostí. Vrstva číslovek detekuje slova, která dohromady tvoří číselnou entitu (včetně fyzikálních jednotek). Vrstva velkých písmen zvětšuje velká písmena u slov, jejichž zvětšování je závislé na kontextu (např. most, nebo ústí). Zkratky a tituly převádí slova (posloupnosti slov) na standardní zápis titulů a zkratek, stejně jako formátování speciálních znaků (zavináč, paragraf). Vrstva doplnění čárkové interpunkce vkládá

do textu interpunkční znaménka (jak bude popsáno v následující sekci). Volitelnými vrstvami jsou specifická pravidla pro formátování textů z některých oborů, případně opravy speciálních chyb (např. jména komentátorů jako Chudoba, nebo Chalupa).

4.2.8 Doplnění interpunkce do přepisu

V této práci jsou navržena dvě schémata pro doplnění interpunkčních znamének (teček a čárek) do přepisu dokumentu. Obě navržená schémata používají stejný nástroj pro doplnění čárek. Vloženou interpunkcí jsou definovány úseky, které označujeme za věty, nejedná se však o věty v jazykovědném významu. Jelikož má navrhované řešení za úkol zpracovat přepisy češtiny i slovenštiny, jsou navržená řešení jazykově nezávislá (což současně vylučuje využití lingvistické analýzy přepisu).

Nástroj pro doplnění čárkové interpunkce vychází ze statistické analýzy jazykových korpusů. Výsledkem analýzy jsou jedno a dvou-slovné sekvence, před a za kterými se nacházejí čárky. Tyto sekvence jsou upřesněny o kontext dalšího slova který může vložení čárky zabránit. Tato pravidla je možné snadno implementovat pomocí vážených stavových automatů, jak je naznačeno na obr. 4.4.

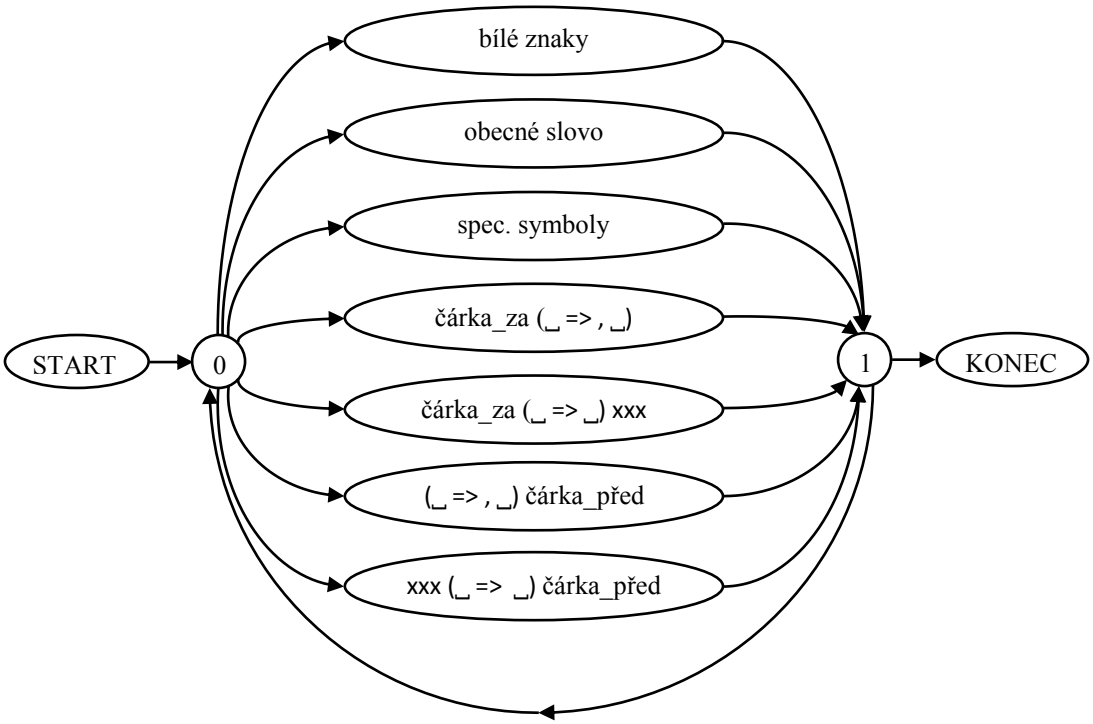
Pro češtinu bylo určeno 1.243 jednoslovných a dvouslovných pravidel, která před sebou generují čárku, a 1.883 prodloužených verzí těchto pravidel, které zakazují umístění čárky. Obdobně je definováno 130 jednoslovných a dvouslovných pravidel, která generují čárku za sebou, nejedná-li se o některou z 518 konkrétnějších frází. Pro slovenštinu bylo nalezeno 2.518 frází generujících čárku před sebou (s 5.071 negativními rozšířeními) a 333 frází generujících čárku za sebou (s 5.752 negativními rozšířeními).

V sekci 6.4 jsou výsledky našeho systému pro doplnění čárkové interpunkce [19] porovnány s českým systémem SET [30], který vychází z lingvistické analýzy textů, a se slovenským nástrojem, který používá statistický přístup [31].

Doplnění tečkové interpunkce vychází ze dvou odlišných principů. Modul, který označujeme **Interpunčkní schéma A** vychází z principů pauzové interpunkce - dělí přepis na celky dlouhé cca 8–14 slov. Pozice dělicích bodů jsou určovány přítomností neřečových událostí v přepisu (ticha, hluky, nádechy) a pomocí trendu fundamentální frekvence promluvy okolo potenciálního dělicícho bodu. Postup je

Tabulka 4.2: Vrstvy textového post-processingu

Odstranění hluků (povinná)
Číslovky, řadové číslovky (experimentální)
Velká písmena
Zkratky
Tituly
Speciální symboly
Doplnění čárkové interpunkce
Oborově-specifická pravidla a formátování (volitelné, více variant)
Oprava specifických chyb v dokumentu (volitelné)



Obrázek 4.4: Struktura WFST automatu pro doplnění čárek do přepisu

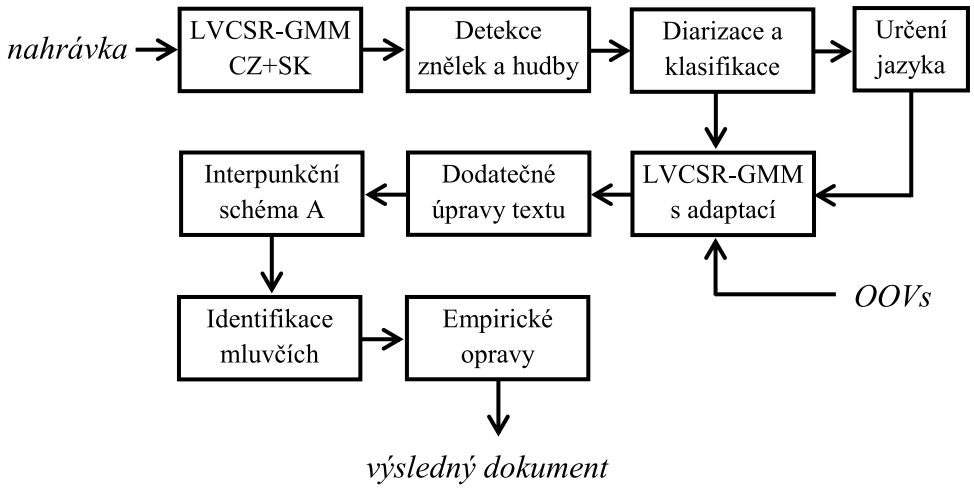
obohacen o seznamy slov, v jejichž okolí se nemají tečky umísťovat (např. ”od”, ”pro”, ”zhruba”).

Interpunčkní schéma B je hlouběji provázáno s celkovým strukturalizačním schématem. Doplnění interpunkce je provedeno po jednotlivých promluvách, ve kterých je každému slotu určena pravděpodobnost přítomnosti interpunkce (na základě prozodických příznaků, přítomnosti neřečových událostí a seznamů zakázaných slov v okolí). Některým slotům je umístění interpunkce zakázáno na základě faktu, že slova dohromady tvoří jmennou (či jinou) entitu. Druhou zapojenou informací je statistika výskytu větných celků určitých délek. Větné celky dělime na čtyři typy: začátek věty - konec věty / začátek věty - čárka / čárka - čárka / čárka - konec věty. Kriteriální funkce pak kombinuje oba zdroje informací - pravděpodobnost přítomnosti interpunkce ve slotu a statistiku délek větných celků. Obsazení slotů všech slotů tečkou / čárkou / žádnou interpunkcí tak vede na úlohu s komplexitou $O = N^3$. Pro vyhodnocení kriteriální funkce byl proto vytvořen generátor variant vložení interpunkce (postupující od začátku rozpoznaného textu ke konci) s možností prořezávání navržených variant. Právě kombinace generátoru variant s prořezáváním variant, jejichž umístění interpunkce dosahuje velmi nízkého skóre, umožňuje vyhodnotit ”všechna” rozmístění interpunkce i u velmi dlouhých promluv.

5 Schémata strukturalizace dokumentu

Mezi hlavní cíle této práce patří porovnání dvou navržených strukturalizačních schémat, které umožní popsat výhody a nevýhody nasazení odlišných konfigurací akustického dekodéru LVCSR systému a dva odlišné přístupy k využití dostupných informačních zdrojů. První popsané schéma (sekce 5.1) je vystavěno na stejném principu, jako systémy popsané v kapitole 2. Druhé schéma (sekce 5.2) kumuluje dostupné informace v rámci rozhodovacích vrstev a pak teprve provádí požadované kroky strukturalizace, čímž umožňuje kombinovat aktuálně dostupné informační zdroje.

5.1 Strukturalizace s izolovaným rozhodováním



Obrázek 5.1: Strukturalizační schéma s izolovaným rozhodováním

Prvním krokem schématu je rozpoznání dokumentu LVCSR systém s kombinovaným (CZ+SK) jazykovým modelem. Následně je provedena detekce (ne)řečových segmentů nahrávky, na kterou navazuje diarizace dokumentu. Pro každý segment je následně klasifikována šířka přenosového pásma a určen jazyk promluvy. Na základě diarizace a provedených segmentací je nahrávka rozdělena na jednotlivé promluvy, podle kterých jsou segmenty nahrávky rozpoznány s adaptací na mluvčího. Dalšími provedenými kroky jsou dodatečné formátování textu a doplnění interpunkce (interpunkčním schématem A - sekce 4.2.8). Nakonec je provedena identifikace mluvčích jednotlivých promluv.

5.2 Strukturalizace s kumulovaným rozhodováním

Strukturalizační schéma s kumulovaným rozhodováním je zachyceno na obr. 5.2. Jeho činnost lze rozdělit do tří vrstev (**I-III**). První vrstva má za úkol připravit takovou segmentaci nahrávky, která umožní rozpoznat nahrávku odpovídajícími akustickými a jazykovými modely a slovníkem. Úloha druhé vrstvy spočívá ve vygenerování finálního textového přepisu nahrávky. Třetí vrstva určuje finální segmentaci dokumentu, doplňuje promluvám požadované informace a provádí formátování textu.

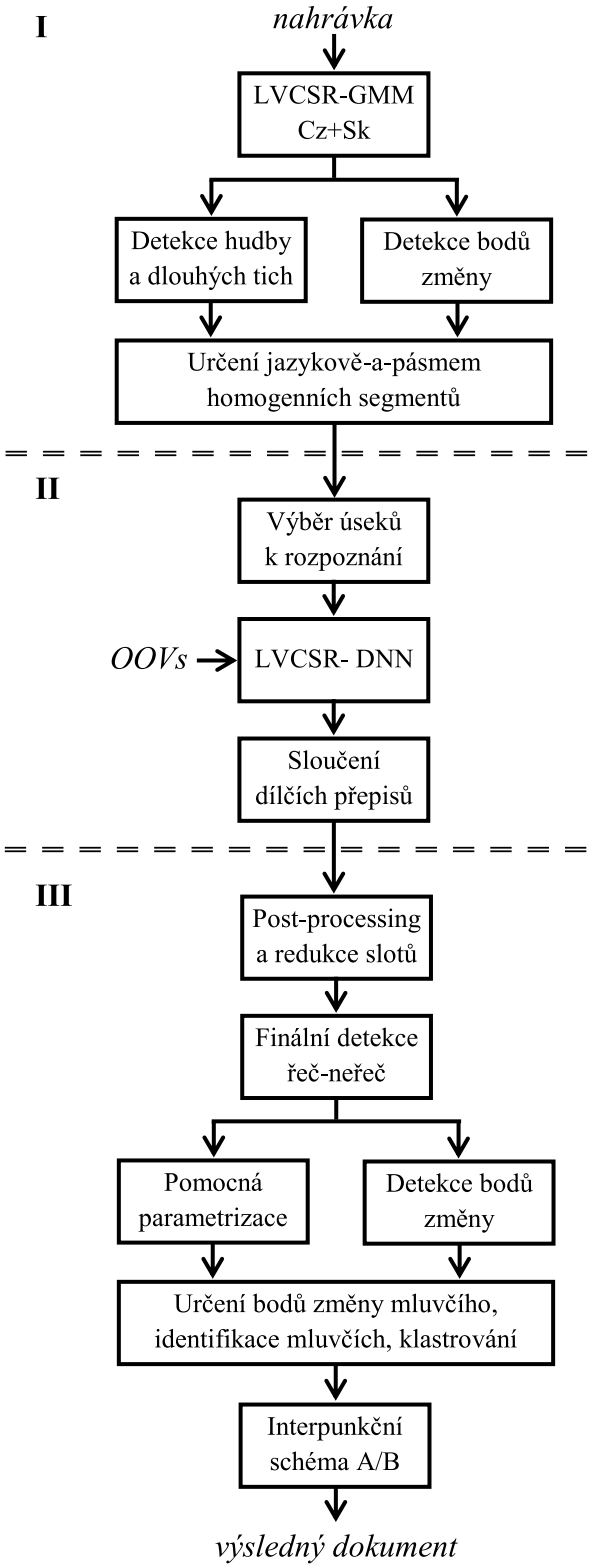
5.2.1 Vrstva I

Úvodním krokem první vrstvy je, stejně jako u předchozího schématu, rozpoznání kombinovaným česko+slovenským LVCSR systémem. Jeho výstup je podkladem pro provedení dvou vzájemně nezávislých analýz. První z nich je detekce bodů změny v nahrávce (sekce 4.2.3). Paralelně jsou detekovány úseky v nahrávce, jejichž obsah není považován za mluvenou řeč. Za neřečové úseky lze považovat dlouhé úseky neřečových událostí (min. 1 s) a veškerý hudební obsah (znělky, písničky apod.). Okraje neřečových regionů jsou interpretovány jako možné body změny v nahrávce.

Kombinace obou analýz nám umožní nejprve definovat sloty podezřelé ze změny mluvčího a vzájemně je verifikovat. Prvním krokem je zarovnání bodů změny detekovaných uvnitř a okolo neřečových regionů. Body změny detekované uvnitř neřečových regionů jsou zarovnány na okraje těchto regionů. Body změny, nacházející se mimo neřečové regiony, ale v jejich těsném okolí jsou zakázány. Nad redukovanou sadou bodů změny jsou provedeny klasifikace přenosového pásma a jazyka (sekce 4.2.4). Na základě označení segmentů jsou nalezeny co nejdelší nepřerušené úseky CZ-NB/CZ-WB/SK-NB/SK-WB/neřeč.

5.2.2 Vrstva II

Podle označení úseků získaných předešlou vrstvou jsou vystříhány spojité úseky (které jsou rozšířeny o případné sousední neřečové úseky). Následuje rozpoznání vystříhaných úseků s pomocí LVCSR-DNN a sloučení získaných přepisů. Slučování přepisů umožňuje vybrat nejvhodnější přepis pro původně sporné hraniční úseky nahrávky (”zleva” a ”zprava” jsou rozpoznány odlišnými modely). To umožňuje vybrat vhodný přepis i pro úseky původně považované za neřečové (např. úsek telefonní nahrávky původně označený za neřečový). Nakonec je provedeno nové klastrování nahrávky. V tomto okamžiku již disponuje vícezdrojové schéma zpracování nahrávky konečným textovým přepisem celé nahrávky, tzn. disponuje konečnou sadou (řečových i neřečových) událostí detekovaných systémem rozpoznání řeči. To by mělo dávat následující vrstvě možnost vycházet z nejpřesnějšího dostupného přepisu a překonat tak přesnost segmentace předchozího schématu.



Obrázek 5.2: Strukturalizační schéma s kumulovaným rozhodováním

Tabulka 5.1: Velikost slovníků pro detekci jmenných entit

	čeština	slovenština
příjmení	285.300	15.370
mužská jména	27.400	2.940
ženská jména	21.830	1.290

Tabulka 5.2: Váhy informačních zdrojů pro detekci změny mluvčího

	změna mluvčího	neutrální závěr	beze změny
F0 řeči	1,0	0,0	-1,5
krátkodobá energie	1,0	0,0	-1,5
hluk za slotem	0,5/1,0	–	–
zakázaná slova	–	–	-1,0/-2,0
čárka ve slotu	–	–	-1,0

5.2.3 Vrstva III

Úvodním krokem třetí vrstvy je provedení formátování textu, které je pro následující moduly důležité ze dvou důvodů. Prvním je označení slov, která patří k sobě tím, že jsou mezi nimi odstraněny mezery (např. "s.r.o.", "n.L."). Druhým je "standardizace" vstupů pro následující zpracování - správná velikost písmen, zkratky a číslovky.

Pro **redukci slotů** je využito vyhledání jmenných entit, oslovení a frází. Mezi jmenné entity řadíme jména osob (včetně případných hodností, titulů, oslovení a funkcí), jména obcí (nalezená podle sekvencí jako "pod Sněžkou", "nad Labem") a názvy společností (detekované podle sekvencí typu "a.s.", "s.r.o." atd.). Jména byla získána z naší databáze mluvčích a ze seznamů českých[1] a slovenských[2] státních institucí (viz tab. 5.1). Výše zmíněné jmenné entity jsou podobné jako v [32] s tím rozdílem, že naše implementace se omezuje na použití slovníků a pravidel.

Následujícím krokem zpracování nahrávky je konečná **klasifikace řeč-neřeč**. Díky předchozím krokům je již provedena nad finálním přepisem nahrávky (což by mělo snížit riziko záměny telefonních nahrávek s hudbou) a možné začátky/konce úseků jsou limitovány na redukovanou sadu slotů.

Úseky nahrávky již mají přiřazený jazyk a šířku přenosového pásma (a redukovaný počet slotů, ve kterých může dojít ke změně mluvčího). Na základě diarizace a dostupné prozodické informace jsou nalezeny body změny mluvčího a promluvám je přiřazena identita (nebo pohlaví) mluvčího. Jednotlivé informační zdroje představují slabé klasifikátory, jsou proto sloučeny v jedno kumulativní skóre (s váhami shrnutými v tabulce 5.2).

Posledním krokem je doplnění interpunkce do přepisu jednotlivých promluv. Provedená redukce slotů umožňuje nasazení obou vytvořených interpunkčních schémat.

[1] http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx
[2] http://www.minv.sk/

6 Vybrané experimety

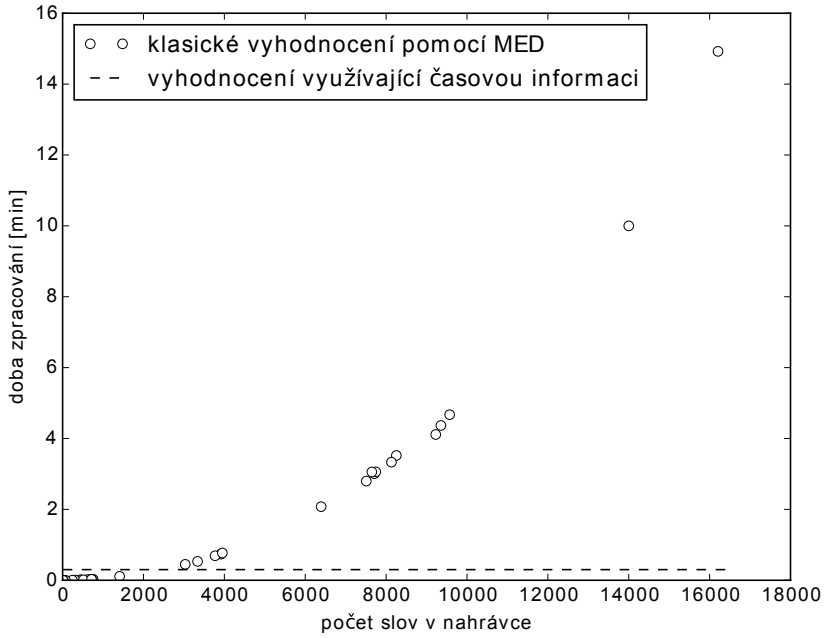
V této kapitole budou uvedeny vybrané experimenty provedené v rámci disertační práce. Strukturalizované dokumenty jsou hodnoceny z řady hledisek a veškeré reference pro tato vyhodnocení jsou obsažena v jedné univerzální struktuře referenčních dat. Reference obsahuje časovaný přepis dokumentu, anotaci neřečových událostí, hranice promluv a tagy atributů promluv (jazyk, šířku přenosového pásma), stejně jako informaci o identitě mluvčího. Tyto reference byly vytvořeny robustním automatickým časováním [33, 34] ručně vytvořených přepisů nahrávek (které byly vytvořeny v rámci projektu NAKI a na projektu spolupracující společností Newton[1]).

Testovací data jsou složena ze 6 skupin nahrávek, kterými se snažíme zaručit maximální různorodost testovacích množin. První jsou nejstarší nahrávky z archivu ČRo (1926–1953) v rozsahu 73 hodin, jejichž kvalita je silně ovlivněna záznamovými zařízeními. Druhou a třetí skupinou jsou nahrávky z let 1971–1999, kdy jedna skupina obsahuje výhradně české promluvy, druhá má zaručen výskyt slovenštiny (celkem 480 hodin). Čtvrtou skupinou jsou moderní zpravodajské pořady (2001–2010) v rozsahu 52 hodin. Pátou skupinu tvoří diskuzní pořady z období 2009–2014, celkem 101 hodin. Poslední testovací skupinu tvoří pořady z internetového portálu Stream.cz[2] (83 hodin nahrávek z období 2012–2014), které obsahují velké množství hluků a hudby na pozadí.

6.1 Využití časované reference

Využití časované reference má řadu výhod. Zaprvé výrazně snižuje časové nároky vyhodnocení experimentů, zejména určení přesnosti přepisu oproti klasické implementaci algoritmu Minimum Edit Distance [35], jak ukazuje obr. 6.1. Zadruhé, umožňuje provést "M-na-N" zarovnání vyhodnocovaného a referenčního přepisu [36]. Toto pokročilejší zarovnání umožňuje správné vyhodnocení chyb způsobených bílými znaky (např. "protože" vs. "proto že", které by MED vyhodnotilo jako substituci a inzerci), stejně jako správné vyhodnocení delších sekvencí záměn slov.

[1] http://www.newtonmedia.cz/cs
[2] https://www.stream.cz



Obrázek 6.1: Porovnání výpočetních nároků výpočtu WER metodou MED a při využití časované reference

6.2 Vyhodnocovací metriky

Pro vyhodnocení vybraných experimentů jsou použity následující metriky: *správnost*, *přesnost* (angl. precision), *úplnost* (angl. recall) a harmonický průměr přesnosti a úplnosti (obvykle značený *F-measure*). V následujícím textu budu pro značení metrik používat zkratky vycházející z jejich anglického označení, které bývá využíváno i v české literatuře. Zavedeme-li *správnost* jako podíl správně klasifikovaných příkladů ku celkovému počtu příkladů, odpovídají definici dvě metriky. První z nich je *accuracy* (6.1), druhou je *correctness* (6.2). Obě metriky nabývají stejných hodnot, pokud je počet hodnocených jevů v referenci a výsledku shodný (může dojít pouze ke správné nebo chybné klasifikaci). Pokud není tento předpoklad dodržen (např. v úloze rozpoznání řeči, kdy rozpoznaný text může obsahovat nejen shody a substituce, ale i inzerce a delece slov), definice *accuracy* se mění (budeme ji dále značit jako *word accuracy* – *WAcc*). Řada autorů citovaných v kapitole 2 využívá pro vyčíslení přesnosti systémů rozpoznání řeči metriku *word error rate* – *WER*, která je komplementární k *WAcc*. *WAcc* a *WER* jsou zavedeny vztahy (6.3) a (6.4).

V předcházejících (i následujících) vztazích zastupuje TP (true positive) počet správně detekovaných výskytů hledaného jevu, TN (true negative) značí počet správně nedetekovaných výskytů hledaného jevu, FP (false positive) představuje počet falešných detekcí hledaného jevu a FN (false negative) je počet chybějících detekcí. N zastupuje celkový počet pozorovaných jevů (podle reference), H (hit) značí počet shod mezi výsledkem a referencí, S (substituce) je počet záměn mezi

referencí a výsledkem, D (delece) je počet jevů, které ve výsledku chybí, a I (inzerce) je počet jevů, které ve výsledku přebývají.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\ [\%] \tag{6.1}$$

$$correctness = \frac{TP}{N} = \frac{H}{N} \times 100\ [\%] \tag{6.2}$$

$$WER = \frac{S + D + I}{S + D + H} = \frac{S + D + I}{N} \times 100\ [\%] \tag{6.3}$$

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N} \times 100\ [\%] \tag{6.4}$$

Přesnost (angl. precision), zavedená vztahem (6.5), udává podíl počtu příkladů správně zařazených do dané třídy a počtu všech příkladů zařazených do třídy (zachycuje četnost výskytu falešných detekcí). *Úplnost* (angl. recall), daná vztahem (6.6), udává podíl počtu všech příkladů správně zařazených do dané třídy a celkového počtu příkladů dané třídy (zachycuje schopnost detekovat danou třídu). *F-measure*, daná vztahem (6.7), je harmonickým průměrem přesnosti a úplnosti (precision a recall). Umožňuje vyčíslit (případně najít) kompromis mezi schopností nástroje detekovat hledaný jev a množstvím falešných detekcí, které nástroj vyprodukuje.

$$precision = \frac{TP}{TP + FP} \times 100\ [\%] \tag{6.5}$$

$$recall = \frac{TP}{TP + FN} \times 100\ [\%] \tag{6.6}$$

$$F\text{–}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \times 100\ [\%] \tag{6.7}$$

Při M-na-N zarovnání (proto značíme metriku $accuracy_{M2N}$), definované vztahem (6.8), považujeme za správně rozpoznaná i slova, která jsou (nebo naopak nejsou) oddělena bílými znaky tak jako v referenci (jejich počet značíme ws). Jako správně rozpoznaná vyhodnocujeme i slova s chybnou koncovkou (jejich počet značíme $ends$).

$$accuracy_{M2N} = \frac{H - I + ends + ws}{N} \times 100\ [\%] \tag{6.8}$$

6.3 Porovnání použitých konfigurací LVCSR

Protože v práci nasazujeme dvě různé konfigurace LVCSR, je logické porovnat přesnost jejich výstupů. LVCSR-GMM rozpoznávač testujeme s adaptací na mluvčího (dle referenční segmentace nahrávek), LVCSR-DNN bez adaptace. Obě konfigurace jsou testovány na 229 hodinách českého zpravodajství. V následující tabulce (tab. 6.1) je vyčíslena *word accuracy* – *WAcc* daná vztahem (6.4) a *correctness* – Corr

dle vztahu (6.2). Pro porovnání je uvedena i hodnota ACC_{M2N} daná vztahem (6.8) - ta nám ukazuje, že vyloučíme-li ze seznamu chyb bílé znaky a chyby v koncovce slov (obvykle sloves, která jsou v řadě případů způsobené špatnou výslovností mluvčího), systém má poměrně malý prostor pro další zlepšení. Ve druhé části tabulky je vyčíslena přesnost detekce neřečových událostí v nahrávce. K popisu kvality detekce neřečových událostí jsou využity metriky *precision* (Prec) - vztah (6.5), *recall* (Rec) – vztah (6.6) a *F-measure* – vztah (6.7).

Tabulka 6.1: Porovnání přesnosti použitých konfigurací LVCSR

řečové události	*WAcc* [%]	Corr [%]	ACC_{M2N} [%]
LVCSR-GMM	82,8	91,3	94,4
LVCSR-DNN	85,5	93,8	96,7
neřečové události	Prec [%]	Rec [%]	F-measure [%]
LVCSR-GMM	92,6	94,6	93,6
LVCSR-DNN	89,0	83,5	86,1

6.4 Vyhodnocení doplnění čárkové interpunkce

V této sekci porovnáme náš nástroj pro dolplnění čárkové interpunkce (sekce 4.2.8) se dvěma současnými systémy. Náš systém bude značen jako *FST*, porovnávaný český systém [30] bude značen *SET* a slovenský nástroj [31] bude značen *SVK*.

Porovnání českých nástrojů pro doplnění čárkové interpunkce proběhlo ve spolupráci s RNDr. Vojtěchem Kovářem, Ph.D.[3] (autorem SETu). Díky tomu byly oba systémy porovnány na stejných testovacích datech v několika odlišných scénářích. Pro testy bylo vybráno 500 úryvků moderního českého zpravodajství (promluvy s délkou 30–45 s). K těmto úryvkům byly k dispozici profesionální přepisy, které posloužily jako reference. Nástrojům byla tato data předána ve třech scénářích, jejichž výsledky jsou porovnány v tabulce 6.2:

1. ruční přepis s ručně určenými konci vět (*sent_manual*)
2. ruční přepis bez určených konců vět (*par_manual*)
3. automaticky rozpoznaný text (*par_asr*)

Pro porovnání našeho nástroje se slovenským nástrojem, vycházejícím také z N–gramových jazykových modelů, jsme nemohli použít stejná testovací a trénovací data. SVK je totiž určen pro zpracování právních textů, zatímco náš systém je trénován na zpravodajských datech. Přesto se domníváme, že je možné porovnat oba nástroje – každý ve své doméně. Oba systémy jsou porovnány na manuálních přepisech v režimu, kdy jsou známy konce vět (zpracování je provedeno po větách). To odpovídá schématu *sent_manual* v předchozím experimentu.

[3] https://nlp.fi.muni.cz/web3/

Tabulka 6.2: Porovnání nástrojů pro doplnění čárkové interpunkce

	čeština						slovenština	
	sent_manual		par_manual		par_asr		sent_manual	
	SET	FST	SET	FST	SET	FST	SVK	FST
precision [%]	93,7	90,2	86,7	82,3	80,1	75,6	95,3	96,3
recall [%]	46,1	54,2	46,1	54,0	42,5	48,3	49,6	49,0
F-measure [%]	61,8	67,7	60,2	65,2	55,5	58,9	65,3	65,0

6.5 Vybraná kritéria segmentace nahrávky

V experimentech používáme následující značení: SCh_{IR} označuje strukturalizační schéma s izolovaným rozhodováním, SCh_{KR} značí schéma s kumulovaným rozhodováním. Sady testovacích dat jsou označeny $\_LQ$ pro náročná data (nejstarší archivní nahrávky a pořady z internetu) a $\_BC$ pro "standardní" rozhlasové vysílání. Pro vyhodnocení schopnosti schémat detekovat body změny v nahrávce jsou použity metriky *precision* (Prec), *recall* (Rec), *F-measure* a *accuracy* (Acc).

Tabulka 6.3: Detekce bodů změny v nahrávce

	Prec [%]	Rec [%]	F-measure [%]	Acc [%]
SCh_{IR}_LQ	28,74	19,96	23,56	97,67
SCh_{KR}_LQ	30,56	56,32	39,62	97,07
SCh_{IR}_BC	73,08	72,13	72,60	99,20
SCh_{KR}_BC	62,74	74,09	67,95	98,97

V tabulce 6.4 sloupec *modely* určuje, pro jakou část řečových segmentů v nahrávce jsou zvoleny správné modely LVCSR (akustický a jazykový model a slovník). Sloupec *VAD* určuje správnost nalezení (ne)řečových regionů v nahrávce. Položky *neřeč⇒řeč* a *řeč⇒neřeč* vyčíslují chybné klasifikace neřečových úseků jako řečových a naopak. Sloupec *mluvčí* pak určuje, jaké části nahrávek bylo přiřazeno správné pohlaví mluvčího.

Tabulka 6.4: Porovnání přesnosti klastrování nahrávky

	modely [%] řeči	VAD [%] pořadu	neřeč⇒řeč [%] pořadu	řeč⇒neřeč [%] pořadu	mluvčí [%] řeči
SCh_{IR}_LQ	30,15	91,48	7,25	1,26	53,43
SCh_{KR}_LQ	41,74	92,85	2,71	3,88	53,02
SCh_{IR}_BC	87,96	98,14	1,68	0,16	97,45
SCh_{KR}_BC	91,82	98,49	1,14	0,26	97,34

7 Závěr

7.1 Výzkumné přínosy práce

V této práci jsou navržena dvě komplexní víceprůchodová schémata strukturalizace archivních nahrávek. Obě schémata produkují informačně bohaté dokumenty, k čemuž plní následující požadavky:

- zajišťují zpracování nahrávky systémem rozpoznání řeči (za použití vhodných akustických a jazykových modelů a slovníku) a získávají informace potřebné pro indexaci rozpoznaného obsahu nahrávky
- umožňují přehledné zobrazení výsledného dokumentu a zlepšují čitelnost a orientaci v přepisu

Existující systémy inventarizace archivních nahrávek (viz kapitola 2) implementují první dva úkony: segmentaci nahrávky na homogenní úseky (ne nutně odpovídající promluvám mluvčích) a klasifikaci vlastností těchto segmentů. Cílem je zjistit, které akustické modely, jazykové modely a slovníky jsou optimální pro rozpoznání daných úseků nahrávky. Systémy tak dosahují maximální možné přesnosti přepisu, který svou strukturou i zobrazením připomíná spíše titulky než přehledný dokument.

Ze stejné logiky zpracování nahrávky vychází i první schéma navržené v této práci (SCh_{IR}). Na rozpoznání nahrávky a klasifikace obsahu nutné pro optimální funkci LVCSR systému navazují další klasifikace (zejména určení identity mluvčího). Podle atributů přiřazených jednotlivým úsekům nahrávky je přepis strukturalizován a je optimalizována jeho čitelnost.

Strukturalizační schéma SCh_{KR} na rozdíl od předchozích systémů vychází z poznatku, že většina sledovaných jevů v nahrávce (např. body změny v nahrávce, interpunkce) jsou navázány na konkrétní sadu pozic v nahrávce - slotů. Snažíme se proto využít informace dostupné z různých modulů pro provedení redefinovaných kroků strukturalizace. Ilustrací může být využití textového obsahu přepisu k redukci slotů v nahrávce. Ve schématu SCh_{KR} jsme dosáhli cca 1,75% míry redukce slotů (zejména svázáním číslovek a víceslovných jmenných entit). Jak ukazuje nedávný výzkum v oblasti víceslovných výrazů [37], míra redukce slotů by mohla dosáhnout až 40 %, kdy autoři detekují ustálená přísloví, aforizmy apod. Výsledky SCh_{KR} překonávají SCh_{IR} ve většině sledovaných metrik, výjimkou jsou chyby způsobené chybějící detekcí některých neřečových událostí (viz tab. 6.1).

Porovnáme-li přesnost rozpoznání nahrávek mezi systémy MALACH, SpeechFind a systémy navrženými v této práci, musíme nejprve zmínit několik zásadních

faktů. První skutečností je časový odstup (přibližně 10 let), který dává našim nástrojům určitý technologický náskok. Druhý důležitý faktor spočívá v charakteru zpracovaných dat. Zatímco SpeechFind pracuje s obdobným typem dat jako my, MALACH zpracovává emocionální nahrávky mluvčích s obtížemi výslovnosti a možným silným přízvukem. MALACH je navíc zatížen přibližně 8 % slov nepokrytých slovníkem (OOV). Navzdory těmto obtížím dosahuje MALACH přibližně 40 % WER. Autoři systému SpeechFind pracují s 1,5 % OOV a dosahují WER 25–40 %. Naše slovníky mají také přibližně 1,5 % OOV, přepisy pak mají méně než 20 % WER. Zajímavé pak je, že ignorujeme-li bílé znaky a chyby v koncovce slov (obvykle sloves), dostává se WER pod 10 %.

Navržená strukturalizační schémata umožnila porovnat výhody a nevýhody použití dvou konfigurací akustického dekodéru systému rozpoznání řeči. LVCSR-GMM provádí druhý průchod s adaptací na mluvčího, zatímco LVCSR-DNN pracuje v jednom průchodu. Tento rozdíl umožňuje výrazné změny v pořadí jednotlivých kroků strukturalizace a vede i ke zrychlení celého procesu (z 3,85 RT na 3,20 RT, RT značí real-time faktor). LVCSR-DNN také prokázal vyšší robustnost a přesnost rozpoznání řečových událostí v nahrávce než LVCSR-GMM.

Nasazená interpunkční schémata ukázala, že pro doplnění čárkové interpunkce lze použít statistické modely vycházející z jazykových korpusů. V otázce detekce hranic větných celků se statistický přístup neosvědčil. Oproti tomu, přístup vycházející z prozodické informace dosahuje přesnějších výsledků.

Hlavní výsledky práce (a posun vůči současném stavu problematiky) lze shrnout v následujících bodech:

- Narozdíl od dřívějších řešení, naše systémy produkují strukturalizované dokumenty, které umožňují snadnou orientaci v rozpoznaném dokumentu. Tyto dokumenty jsou obohaceny o doplňkové informace (jazyk promluvy, dientita mluvčího, šířka přenosového pásma) a přepis je upraven pro lepší čitelnost (post-processing, rozdělení na věty).

- Doposud vytvořené systémy aplikovaly jednotlivé nástroje jako posloupnost vzájemně nezávislých úloh. Navržené strukturalizační schéma s kumulovaným rozhodováním zavádí pojem slot (sadu časových značek, vůči kterým jsou nástroje synchronizovány) a umožňuje kombinaci dílčích informačních zdrojů.

- Využití časované reference v kombinaci s nástroji pro automatické časování referenčních dat umožňuje časově efektivní tvorbu referenčních dat a podrobnější vyhodnocení jednotlivých metrik.

- Pomocí navržených nástrojů byla zpřístupněna část archivu Českého rozhlasu v rozsahu větším než 100.000 hodin archivních dokumentů.

7.2 Praktické přínosy práce

Strukturalizační schémata popsaná v této práci byla postupně nasazena ke zpracování části archivu Českého rozhlasu (popsaný v tab. 7.1). K provozu celého archivu bylo zapotřebí (kromě inventarizace a strukturalizace nahrávek popsané v této práci) navrhnout uživatelské rozhraní a propojit ho s databází přepisů a stream-serverem. Uživatelské rozhraní lze rozdělit na vyhledávací rozhraní a přehrávací rozhraní.

Filtry vyhledávacího rozhraní umožňují provádět řadu analýz obsahu archivu (nalezené výsledky lze třídit např. podle období vzniku nebo jazyka promluvy). Již během vytváření archivu byly vypracovány první studie pozorující vývoj jazyka, slovní zásoby a výslovnosti v čase [38, 39, 40]. Po zveřejnění archivu ho začali hojně využívat pracovníci Českého rozhlasu i různá výzkumná pracoviště (např. oddělení současné lexikologie a lexikografie ÚJČ AV ČR).

Tabulka 7.1: Rozsah zpracované části archivu ČRo

Celkový objem zpracovaných nahrávek (hodiny)	102.953
Počet rozhlasových stanic	20
Počet pořadů	326
Počet zpracovaných dokumentů	213.453
Počet zaindexovaných slov	469.976.314
Odhad celkového objemu výpočetního času	1.500.000

7.3 Návrhy budoucí práce

Jak již bylo řečeno v předchozím textu, při úloze strukturalizace dokumentu jsou důležité nejen řečové, ale i neřečové události v nahrávce. Proto by bylo vhodné zapojit do procesu trénování DNN akustického dekodéru (s jehož topologií, probíhá řada experimentů) navržené vyhodnocovací schéma využívající časovanou referenci.

Druhá důležitá změna spočívá v extrakci prozodických příznaků. V této práci jsou prozodické příznaky vázány na jednotlivé řečové a neřečové události v nahrávce. Jako vhodnější se jeví navázat prozodickou informaci na menší jednotky – slabiky. K tomu by bylo zapotřebí extrahovat časové značky na úrovní fonémů a vytvořit nástroj pro slabikování slov.

Třetím krokem, který by mohl snížit počet chyb v segmentaci, je větší míra redukce počtu slotů v nahrávce. Lingvistický výzkum ukazuje, že na základě analýzy textového obsahu přepisu lze "svázat" až 40 % slov přítomných v přepisu [37].

Pro optimální zpracování jednotlivých nahrávek by bylo vhodné navrhnout strategii pro rozlišení typů pořadů (diskuzní pořad/zpravodajské vysílání/projev). K tomu by mohla posloužit informace o četnosti změn mluvčích, počtu mluvčích v nahrávce a délce trvání konkrétních promluv (podobně jako v [41]). Klasifikace typu pořadu by umožnila specifikovat nároky na výstupní dokument, odhadnout množství interpunkce v pořadu apod.

Literatura

[1] P. Mihajlik, T. Fegyó, B. Németh, Z. Tüske, and V. Trón, "Towards automatic transcription of large spoken archives in agglutinating languages–hungarian asr for the malach project," in *TSD 2007*, pp. 342–349, Springer, 2007.

[2] W. Byrne, D. Doermann, M. Franz, S. Gustman, J. Hajič, D. Oard, M. Picheny, J. Psutka, B. Ramabhadran, D. Soergel, *et al.*, "Automatic recognition of spontaneous speech for access to multilingual oral history archives," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 4, pp. 420–435, 2004.

[3] J. Hansen, R. Huang, P. Mangalath, B. Zhou, M. Seadle, and J. R. Deller Jr, "Speechfind: spoken document retrieval for a national gallery of the spoken word," in *les actes de Nordic Signal Processing Symposium (NORSIG)*, 2004.

[4] J. Hansen, R. Huang, B. Zhou, M. Seadle, J. Deller, J.R., A. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, pp. 712–730, Sept 2005.

[5] U. H. Yapanel and J. H. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition.," in *INTERSPEECH*, 2003.

[6] M. Franz, J. McCarley, T. Ward, and W. Zhu, "Segmentation and detection at ibm: Hybrid statistical models and two-tiered clustering," *1999 TDT Evaluation System Summary Papers*, 1999.

[7] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[8] H. D. Wactlar, A. G. Hauptmann, and M. J. Witbrock, "Informedia tm: News-on-demand experiments in speech recognition," in *Proc. of ARPA Speech Recognition Workshop*, pp. 18–21, 1996.

[9] G. Cook, J. Christie, D. P. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams, "An overview of the sprach system for the transcription of broadcast news," in *Proceedings of the DARPA Broadcast News Workshop, February 28-March 3, 1999, Hilton at Washington Dulles Airport, Herndon, Virginia*, Information Technology Laboratory, National Institute of Standards and Technology, 1999.

[10] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA speech recognition workshop*, vol. 1997, 1997.

[11] T. Hain, S. Johnson, A. Tuerk, P. Woodland, and S. Young, "Segment generation and clustering in the htk broadcast news transcription system," in *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*, pp. 133–137, 1998.

[12] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pp. 347–354, IEEE, 1997.

[13] J. Psutka, L. Müller, J. Matoušek, and V. Radová, *Mluvíme s počítačem česky*. Prague: Academia, 2006.

[14] M. D. Skowronski and J. G. Harris, "Improving the filter bank of a classic speech feature extraction algorithm," in *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on*, vol. 4, pp. IV–281, IEEE, 2003.

[15] J. Nouza, K. Blavka, P. Červa, J. Žďánsky, J. Silovský, M. Boháč, and J. Pražák, "Making czech historical radio archive accessible and searchable for wide public," *Journal of Multimedia*, vol. 7, no. 2, pp. 159–169, 2012.

[16] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks.," in *Interspeech*, pp. 437–440, 2011.

[17] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *COMPUTER SPEECH AND LANGUAGE*, vol. 12, pp. 75–98, APR 1998.

[18] M. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *COMPUTER SPEECH AND LANGUAGE*, vol. 10, pp. 249–264, OCT 1996.

[19] M. Boháč, K. Blavka, M. Kuchařová, and S. Škodová, "Post-processing of the recognized speech for web presentation of large audio archive.," in *International Conference on Telecommunications and Signal Processing - TSP*, pp. 441–445, IEEE, 2012.

[20] J. Pražák and J. Silovský, "Comparison of segmentation and clustering methods for speaker diarization of broadcast stream audio," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues* (A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, eds.), vol. 6800 of *Lecture Notes in Computer Science*, pp. 214–222, Springer Berlin Heidelberg, 2011.

[21] J. Pražák and J. Silovský, "Speaker diarization using plda-based speaker clustering," in *Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2011 IEEE 6th International Conference on*, vol. 1, pp. 347–350, Sept 2011.

[22] J. Silovský and J. Pražák, "Speaker diarization of broadcast streams using two-stage clustering based on i-vectors and cosine distance scoring," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4193–4196, March 2012.

[23] Y. Yan, E. Barnard, and R. A. Cole, "Development of an approach to automatic language identification based on phone recognition," *Computer Speech and Language*, vol. 10, no. 1, pp. 37–54, 1996.

[24] J. Nouza, P. Cerva, and J. Silovsky, "Dealing with bilingualism in automatic transcription of historical archive of czech radio," in *New Trends in Image Analysis and Processing–ICIAP 2013*, pp. 238–246, Springer, 2013.

[25] J. Silovsky, J. Nouza, and M. Kucharova, "Search for speaker identity in historical oral archives," *Multimedia Tools and Applications*, pp. 1–20, 2014.

[26] M. Kuchařová, S. Škodová, L. Šeps, and M. Boháč, "Study on phrases used for semi-automatic text-based speakers names extraction in the czech radio broadcasts news," in *Text, Speech and Dialogue* (P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds.), vol. 8655 of *Lecture Notes in Computer Science*, pp. 416–423, Springer International Publishing, 2014.

[27] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[28] H. Atassi, "Metody detekce základního tónu řeči," *Elektrorevue*, no. 4, pp. 4–1 – 4–17, 2008.

[29] X. Sun, "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–333–I–336, May 2002.

[30] V. Kovář, "Partial grammar checking for czech using the set parser," in *17th International Conference, TSD 2014*, (Berlin Heidelberg), pp. 308–314, 2014.

[31] R. Sabo and t. Beňuš, "Detecting commas in slovak legal texts," in *Text, Speech and Dialogue* (P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds.), vol. 8655 of *Lecture Notes in Computer Science*, pp. 62–67, Springer, 2014.

[32] P. Král, "Features for named entity recognition in czech language," in *KEOD 2011 - Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pp. 437–441, 2011.

[33] M. Boháč and K. Blavka, "Automatic segmentation and annotation of audio archive documents," in *International Workshop on Electronics, Control, Measurement and Signals*, 2011.

[34] M. Boháč and K. Blavka, "Text-to-speech alignment for imperfect transcriptions," in *Text, Speech, and Dialogue* (I. Habernal and V. Matoušek, eds.), vol. 8082 of *Lecture Notes in Computer Science*, pp. 536–543, Springer, 2013.

[35] R. A. Wagner and M. J. Fischer, "The string-to-string correction problem," *Journal of the ACM*, vol. 21, no. 1, pp. 168–173, 1974.

[36] M. Boháč, J. Nouza, and K. Blavka, "Investigation on most frequent errors in large-scale speech recognition applications," in *Text, Speech and Dialogue* (P. Sojka, A. Horák, I. Kopeček, and K. Pala, eds.), vol. 7499 of *Lecture Notes in Computer Science*, pp. 520–527, Springer Berlin Heidelberg, 2012.

[37] A. Savary, M. Sailer, Y. Parmentier, M. Rosner, V. Rosén, A. Przepiórkowski, C. Krstev, V. Vincze, B. Wójtowicz, G. S. Losnegaard, *et al.*, "Parseme–parsing and multiword expressions within a european multilingual network," in *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, 2015.

[38] V. Lábus, "Atyp v cihle aneb o jednom progresivním způsobu neologizace," *Naše řeč*, no. 4, pp. 187–197, 2012.

[39] M. Kuchařová, S. Škodová, L. Šeps, V. Lábus, J. Nouza, and M. Boháč, "On the quantitative and qualitative speech changes of the czech radio broadcasts news within years 1969–2005," in *TSD 2013*, pp. 360–368, Springer, 2013.

[40] S. Škodová, M. Kuchařová, and L. Šeps, "Discretion of speech units for the text post-processing phase of automatic transcription (in the czech language)," in *Text, Speech and Dialogue*, pp. 446–455, Springer, 2012.

[41] D.-C. Lyu, R.-Y. Lyu, Y.-C. Chiang, and C.-N. Hsu, "Cross-lingual audio-to-text alignment for multimedia content management," *Decision Support Systems*, vol. 45, no. 3, pp. 554–566, 2008.

Seznam autorových publikací

Mezinárodní konference

1. M. Boháč a K. Blavka, "Automatic segmentation and annotation of audio archive documents," in *Electronics, Control, Measurement and Signals (ECMS), 2011 10th International Workshop on*, pp. 1–6, June 2011.

2. J. Nouza a M. Boháč, "Using TTS for fast prototyping of cross-lingual ASR applications," in Analysis of Verbal and Nonverbal Communication and Enactment, pp. 154–162, 2011.

3. M. Boháč, J. Nouza, a K. Blavka, "Investigation on most frequent errors in large-scale speech recognition applications.," in Text, Speech and Dialogue *TSD*, pp. 520–527, 2012.

4. M. Boháč, K. Blavka, M. Kuchařová, a S. Škodová, "Post-processing of the recognized speech for web presentation of large audio archive," in *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on*, pp. 441–445, July 2012.

5. M. Boháč, "Performance comparison of several techniques to detect keywords in audio streams and audio scene," in *ELMAR, 2012 Proceedings*, pp. 215–218, Sept 2012.

6. J. Nouza, K. Blavka, M. Boháč, P. Červa, J. Žďánský, J. Silovský a J. Pražák, "Voice technology to enable sophisticated access to historical audio archive of the Czech radio," in 14th International Workshop on Multimedia Signal Processing *MMSP*, pp. 337–342, 2012.

7. J. Nouza, K. Blavka, J. Žďánský, P. Červa, J. Silovský, M. Boháč, J. Chaloupka, M. Kuchařová a L. Šeps, "Large-scale processing, indexing and search system for Czech audio-visual cultural heritage archives," in Multimedia for Cultural Heritage, pp. 27–38, 2012.

8. J. Pražák a M. Boháč, "Speaker diarization of broadcast audio using automatic transcription, iVectors and cosine distance scoring," in Proceedings of *ELMAR*, pp. 211–214, 2012.

9. M. Boháč, J. Málek, a K. Blavka, "Iterative grapheme-to-phoneme alignment for the training of wfst-based phonetic conversion," in *TSP*, pp. 474–478, 2013.

10. M. Boháč a K. Blavka, "Text-to-speech alignment for imperfect transcriptions," in *Text, Speech, and Dialogue* (I. Habernal and V. Matoušek, eds.), vol. 8082 of *Lecture Notes in Computer Science*, pp. 536–543, Springer Berlin Heidelberg, 2013.

11. M. Boháč a L. Šeps, "Comparison of several techniques for detection of key slides in lecture support materials," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, pp. 783–787, July 2013.

12. M. Kuchařová, S. Škodová, V. Lábus, L. Šeps, M. Boháč, J. Nouza, "On the Quantitative and Qualitative Speech Changes of the Czech Radio Broadcasts News within Years 1969–2005," in Proceedings of Text, Speech and Dialogue *TSD*, pp. 360–368, 2013.

13. M. Boháč a K. Blavka, "Using suprasegmental information in recognized speech punctuation completion," in Text, Speech and Dialogue *TSD*, pp. 555–562, 2014.

14. M. Boháč, M. Rott a K. Blavka, "On Automatic Cross-Lingual Subtitle Timing," in Electronics, Control, Measurement, Signals and their Application to Mechatronics (*ECMSM*), pp. 1–6, 2015.

15. M. Boháč a M. Rott, "Exploiting of the timing information in subtitle-like parallel multilingual data," in 7th Language & Technology Conference (LTC'15), Poznań, pp. 208-212, 2015.

Časopisecké publikace

1. M. Boháč, M. Kuchařová, Z. Cajellas, J. Nouza, a P. Červa, "A cross-lingual adaptation approach for rapid development of speech recognizers for learning disabled users," *EURASIP Journal on Audio, Speech, and Music Processing*, vol.1, 2014.

2. J. Nouza, K. Blavka, P. Červa, J. Žďánský, J. Silovský, M. Boháč a J. Pražák, "Making czech historical radio archive accessible and searchable for wide public," in Journal of Multimedia, vol. 7, pp. 159–169, 2012.