TECHNICKÁ UNIVERZITA V LIBERCI
www.tul.cz

Automatická strukturalizace počítačem přepsaných mluvených dokumentů z multimediálních archivů

Disertační práce

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

Autor práce: **Ing. Marek Boháč**
Vedoucí práce: prof. Ing. Jan Nouza, CSc.

Liberec 2016

TECHNICAL UNIVERSITY OF LIBEREC
www.tul.cz

Automatic processing of computer-transcribed spoken documents from multimedia archives

Dissertation

Study programme: P2612 – Electrotechnics and informatics
Study branch: 2612V045 – Technical cybernetics

Author: **Ing. Marek Boháč**
Supervisor: prof. Ing. Jan Nouza, CSc.

Liberec 2016

Prohlášení

Disertační práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé disertační práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum: 7. 3. 2016

Podpis:

Abstrakt

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 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. Vzhledem k rozsahu archivu (213.000 dokumentů z období 1923 až 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 i slovenského jazyka v dokumentech, se střídajícími se mluvícími osobami, s prokládáním řeči znělkami, hudebními předěly 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 na více či méně homogenní části a následně 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, hudbu nebo píseň. 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 této práci jsou popsány postupy a metody, které umožňují automaticky rozčlenit jednotlivé audio dokumenty a jejich počítačem pořízené přepisy do kratších úseků na základě identifikace změny charakteru signálu, změny jazyka, změny řečníka, následně též rozdělení textu do vět, doplnění interpunkce a do jisté míry i určité úpravy vzhledu textu (např. převod čísel vyjádřených textem na odpovídající číslice). Zároveň je třeba všechny tyto informace vhodným způsobem popsat a uložit do databáze, a to včetně přesných časových značek, aby bylo možné podle nich vyhledávat a okamžitě přistupovat k nalezeným objektům, kterými mohou být slova, fráze či hovořící osoby. Navržená struktura dat musí také umožňovat hledání podle dalších kritérií, jako jsou např. jazyk promluvy či charakter záznamu.

Pro tyto účely byla navržena, implementována a experimentálně ověřena dvě schémata (řetězce) zpracování mluveného dokumentu. Schémata jsou koncipována tak, abychom mohli porovnat dva různé přístupy k informacím, které produkují jednotlivé nástroje zapojené v řetězci. Současně pomocí schémat porovnáváme vhodnost dvou implementovaných modulů pro doplnění interpunkce a možnosti plynoucí z různých konfigurací systému rozpoznání řeči.

První navržené schéma provádí izolované rozhodování (každý krok strukturalizace využívá informaci získanou z jednoho konkrétního nástroje v řetězci). Druhé schéma kumuluje rozhodování do obsáhlejších vrstev, v nichž využívá všechny dostupné informační zdroje současně. Tento odlišný přístup umožňuje zpřesnit přiřazení správných modelů pro systém rozpoznávače řeči (akustický model a jazykový model) z 87,96% na 91,82% při použití stejných dílčích modulů. V otázce doplnění interpunkce proti sobě stavíme přístup spoléhající na korelaci mezi neřečovými událostmi v nahrávce a přítomností interpunkce v přepisu a přístup spoléhající na statistický popis větných celků.

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.

Navržené metody byly prakticky nasazeny v projektu, v němž se podařilo zpracovat přes 213.000 archivních dokumentů v celkovém trvání přesahujícím 100.000 hodin. V archivu je možné vyhledávat pomocí veřejně přístupné webové aplikace. Vyvinuté technologie a postupy lze využít i pro další typy multimediálních dat obsahujících řeč, např. televizní či filmové archivy.

Klíčová slova: automatická strukturalizace nahrávky, zpřístupnění archivu mluveného slova, rozpoznání řeči.

Abstract

This thesis focuses on solving a complex task how to structure (i.e. appropriately divide, textually and phonetically analyze and subsequently modify) the output of the speech recognition system so it is most readable for human and also prepared for effective machine processing and search. Motivation to solve this task was the research project supported by the Czech Ministry of culture, aimed at transcription of spoken documents contained in the Czech and Czechoslovak radio and to make them available for search. Taking into account the archive size (213,000 documents form the years 1923-2014) it was essential to propose and implement such technologies, that were able to handle not only the waste amount of the data but also some specific issues associated with different acoustic quality of the documents, speaker changes, presence of jingles, music divides and song between the speech segments or with background noise.

For these purposes modules solving automatic speech recognition, speaker and language identification, recording segmentation on more or less homogenous segments, followed with classification of the segments, taking into account the presence of speech (clean, noisy, narrowband, etc.) were developed at the Technical University of Liberec. Author of this thesis participated on development of some of the above mentioned modules and especially contributed to fitting the modules into the processing chain. He solved the module optimizations in order to maximize the accuracy of the processed documents transcriptions while keeping the searching in the archive user friendly.

This thesis describes procedures and methods that allow automatically segment particular audio documents and their computer-produced transcriptions into shorter segments. The segmentation is based on identification of the change of the signal character, change of the language, change of speaker, followed by determination of sentences, punctuation completion and some amount of text formatting (e.g. representation of recognized numbers by corresponding numerals). Simultaneously we need to keep all the information and store it in a database, including the exact time stamps, which are essential for searching and accessing to found objects such as words, phrases or speakers. Proposed data structure must also enable to search by other criteria, e.g. language of utterance or character of the recording.

Two schemes (processing chains) for processing of a spoken document were proposed, implemented and experimentally evaluated for these purposes. The design of schemes allows us to compare two different approaches to the available information produced by individual tools employed in the processing chain. Simultaneously, the two schemes are used to compare two proposed punctuation tools and to investigate possibilities arising from different configurations of the speech recognition system.

First proposed scheme performs isolated decisions (each structuring step employs one type of information gained from one tool in the processing chain). Second scheme accumulates the decision making in larger layers and uses all the available information sources at once. The second approach allows more accurate chooses of models employed by the speech recognition system (acoustic and language model) from 87.96% to 91.82%, while using the same tools in the processing chain. The issue of punctuation completion compares an approach which relies on correlation between non-speech occurrence in the recording and presence of punctuation in the transcription with an approach based on the statistical description of sentences.

In order to evaluate the above mentioned tasks we implemented tools which enable automatic completion of the time stamps into the reference transcription. Concurrently, we propose (and implement) evaluation tools which utilize such timed reference data and provide us more detailed and time-efficient evaluation of established metrics.

Proposed methods were practically applied in the project, which succeeded in processing more than 213,000 archive documents exceeding a total duration of 100,000 hours. The archive can be searched via publically available web application. Developed technologies and methods may be used to process other types of multimedia data containing speech, e.g. television or movie archives.

Key-words: automatic structuralization of recording, making spoken word archive accessible, speech recognition.

Poděkování

Rád bych poděkoval své rodině za trpělivost prokázanou během mých studií a prof. Ing. Janu Nouzovi, CSc. za odborné vedení, řadu inspirativních podnětů a rad při psaní této práce.

Také bych rád poděkoval RNDr. Vojtěchu Kovářovi, Ph.D. z Centra zpracování přirozeného jazyka MU v Brně za vstřícnost při plánování a provádění společných experimentů.

Dík patří také Ing. Karlu Blavkovi za přínosné porady a spolupráci.

Obsah

Seznam zkratek

Acc	Accuracy (vyhodnocovací metrika)
AM	Acoustic Model (akustický model)
ASR	Automatic Speech Recognition (automatické rozpoznání řeči)
BIC	Bayessian Information Criterion (Bayesovské informační kritérium)
CD-DNN	Context Dependent DNN (hluboká neuronová síť; stavy výstupní vrstvy odpovídají stavům akustických modelů, které modelují fonémy s ohledem na jejich okolí - často odpovídají tzv. senonům)
CD-RNN	Context Dependent Recurrent Neural Network (neuronová síť se zpětnou vazbou mezi vrstvami; stavy výstupní vrstvy odpovídají jednotlivým senonům akustického modelu)
CI-DNN	Context Independent Deep Neural Network (DNN; stavy výstupní vrstvy odpovídají monofonům akustického modelu)
CI-MLP	Context Independent Multi-Layer Perceptron (neuronová síť s jednou skrytou vrstvou; stavy výstupní vrstvy odpovídají monofonům akustického modelu)
Corr	Correctness (vyhodnocovací metrika)
ČRo	Český (Československý) rozhlas
DER	Diarization Error Rate (metrika kvality diarizace dokumentu)
DNN	Deep Neural Network (neuronová síť s více skrytými vrstvami)
FMMIS	Fakulta mechatroniky, informatiky a mezioborových studií
F-measure	Harmonický průměr Precision a Recall (vyhodnocovací metrika)
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model (skryté markovské modely)
LM	Language Model (jazykový model)
LPC	Linear Prediction Coefficients (metoda parametrizace nahrávky)
LVCSR-GMM	Označení pro rozpoznávač spojité řeči NanoDictateT[1] v CD-GMM-HMM konfiguraci akustických modelů
LVCSR-DNN	Označení pro rozpoznávač spojité řeči NanoDictateT v CD-DNN-HMM konfiguraci akustických modelů
MED	Minimum Edit Distance (metoda zarovnání textových řetězců)
MFCC	Mel-Frequency Cepstral Coefficients - kepstrální příznaky řeči
OOV	Out of Vocabulary (slovo mimo slovní zásobu)
PCM-Wave	Bezeztrátový formát kódování audia
PLP	Perceptual Linear Prediction (metoda parametrizace nahrávky)
Prec	Precision (vyhodnocovací metrika)
Rec	Recall (vyhodnocovací metrika)
RT	Real Time faktor (poměr délky zpracování a trvání nahrávky)
SER	Slot Error Rate (metrika pro vyhodnocení doplnění interpunkce)
STFT	Short Time Fourier Transform
TUL	Technická Univerzita v Liberci
WER	Word Error Rate (metrika pro vyhodnocení přesnosti ASR)
WFST	Vážené konečné stavové automaty

[1] rozpoznávač řeči vyvinutý v Laboratoři počítačového zpracování řeči, FMMIS, TUL

Seznam obrázků

Seznam tabulek

1 Úvod

Přibližně od 90. let 20. století se postupně daří řešit řadu úloh zpracování mluveného slova pomocí nejrůznějších technologií počítačového zpracování řeči. Počínaje úlohou detekce omezené množiny klíčových slov ve zvukovém záznamu (tzv. keyword spotting) přes nucené zarovnání nahrávky a přepisu (tzv. forced alignment) až po úplné rozpoznání nahrávky systémem počítačového rozpoznání řeči (automatic speech recognition). U všech těchto technologií lze sledovat jak postupný růst přesnosti výsledků, tak schopnost pracovat s většími objemy dat (např. velikosti používaných slovníků), jež přímo souvisí i s růstem dostupného výpočetního výkonu. Nejvyšší metou jsou pak rozpoznávače spojité řeči pracující s velkou slovní zásobou (large vocabulary continuous speech recognition - LVCSR), které umožňují práci s tvaroslovně bohatými jazyky (mezi které patří i všechny jazyky slovanské).

Paralelně se rozvíjí i další skupina technologií zpracovávajících zvukové nahrávky. Jejich společným jmenovatelem je poskytnutí doplňkové informace o obsahu nahrávky, respektive jejích částí. Tato meta-data mohou obsahovat různorodé informace o samotném obsahu nahrávky. Prvně mohou rozlišit mluvené slovo, hudební obsah nebo jiné neřečové události (zvuky dopravních prostředků, výstřely apod.). Druhou důležitou doplňkovou informací je typ přenosového kanálu. S ohledem na podmínky vzniku konkrétní nahrávky mohou hrát roli jak média použitá k uložení záznamu, tak přenosové cesty zapojené během vzniku jednotlivých úseků nahrávky (telefonní vstupy, nahrávky pořízené na různá přenosná zařízení). Další skupina nástrojů se pak zabývá získáním informací o mluvčích v nahrávce. Jedná se o detekci bodů změny mluvčího (speaker-turn detection), případně následovanou nástrojem pro diarizaci mluvčích (určení "kdo mluvil kdy"). Dalšími úkoly jsou určení pohlaví mluvčího, případně jeho identity (jsou-li k dispozici modely důležitých řečníků) a v případech jazykově nehomogenní nahrávky lze detekovat jazyk promluvy, případně nářečí či přízvuk hovořícího.

Druhým pro nás významným fenoménem posledních desetiletí je generování obrovských objemů zvukových i audiovizuálních nahrávek s obsahem řeči. Příkladem lze uvést digitalizaci řady historických archivů (převážně televizních a rozhlasových), v nichž moderní společnost spatřuje součást svého historického a kulturního dědictví. Za součást kulturního dědictví jsou považovány i archivy "živé paměti", které se zaměřují na zachování autentických vzpomínek přímých účastníků významných událostí. Příkladem takových archivů může být rozsáhlý projekt MALACH, který se zaměřuje na události holocaustu [1, 2], či projekt Paměť národa[1], který mapuje

[1] http://archiv.postbellum.cz/cz/pamet-naroda/co-je-pamet-naroda.aspx

významné události 20. století. Současně vznikají nejrůznější archivy užitkové (záznamy bezpečnostních agentur, monitorované obchodní hovory) spolu s nepřeberným množstvím zábavné tvorby.

V případě digitalizovaných historických archivů je jakákoli práce s jejich obsahem odkázána na práci archivářů, kteří na základě kusých poznámek o obsahu nahrávek musí ručně vyhledávat dokumenty, které by mohly být relevantní pro konkrétního uživatele. U archivů nově vzniklých je obvykle k dispozici různé množství meta-dat, podle kterých je možné vyhledávat relevantní dokument a následně získat požadovanou nahrávku. Ani v jednom případě však nelze vyhledávat na základě obsahu zkoumaných dokumentů. Zjistit, jestli daný dokument opravdu obsahuje hledanou informaci, vyžaduje, aby si výzkumník přehrál velkou část daného dokumentu a zjistil tak jeho skutečný obsah. V případě některých existujících multimediálních archivů již byly učiněny úspěšné kroky vedoucí ke zpřístupnění jejich obsahu. U velmi důležitých dokumentů toho bylo dosaženo prostřednictvím pořízení ručních přepisů. V měřítkách celých archivů se však ukazuje, že jediným možným přístupem je využití technologií zpracování řeči. Každý typ archivu má svá specifika, která definují, jakým způsobem chceme nahrávky zpracovat a také jak náročný tento proces bude. Úlohy spojené se zpřístupňováním archivů lze rozdělit do následujících podskupin:

1. rozdělení nahrávky či detekce sub-dokumentů (např. rozdělení nahrávky ve smyslu zákazník–operátor, lokalizace jednotlivých účastníků debaty),

2. generování meta-dat relevantních pro vyhledávání (detekce hudby a znělky, klasifikace přenosového pásma, identifikace mluvčích, určení jazyka promluvy),

3. pořízení časovaného slovního přepisu nahrávky, který umožní vyhledávání a navigaci v dokumentu, a

4. pořízení kompletního informačně bohatého přepisu, který umožňuje dobrou orientaci v dokumentu a zprostředkovává maximum informací uživateli.

Posledně jmenovaná varianta zahrnuje prakticky všechny předešlé úlohy zpracování nahrávky. Lze také definovat určitý hybridní přístup, kdy je nahrávka automaticky zpracována a vyškoleným pracovníkům je následně umožněno doopravit výsledný dokument. Takový přístup je možné zvolit buď u významných historických dokumentů, nebo u moderních nahrávek, kdy například moderátor pořadu může zdokonalit přepis vlastního pořadu krátce po jeho skončení. Hlavní výhodou hybridního zpracování je vysoká přesnost přepisů a přijatelné množství lidského úsilí.

2 Motivace

V České republice, stejně jako v řadě dalších vyspělých zemí, došlo v nedávné době k digitalizaci multimediálních archivů, jejichž obsah je považován za kulturní dědictví. Například od roku 2003 probíhala digitalizace archivu Československého rozhlasu (2. nejstarší rozhlasové stanice v Evropě), na kterou navazuje průběžná digitalizace současného vysílání ČRo. Obdobně probíhá zpracování archivů a vysílání České (Československé) televize. Nejnovějším trendem jsou archivy "pamětí", mezi nimiž lze jmenovat MALACH (obsahující rozhovory s pamětníky holocaustu [1, 2]) či výše zmíněné projekty Paměti národa (zaměřené na události 20. století).

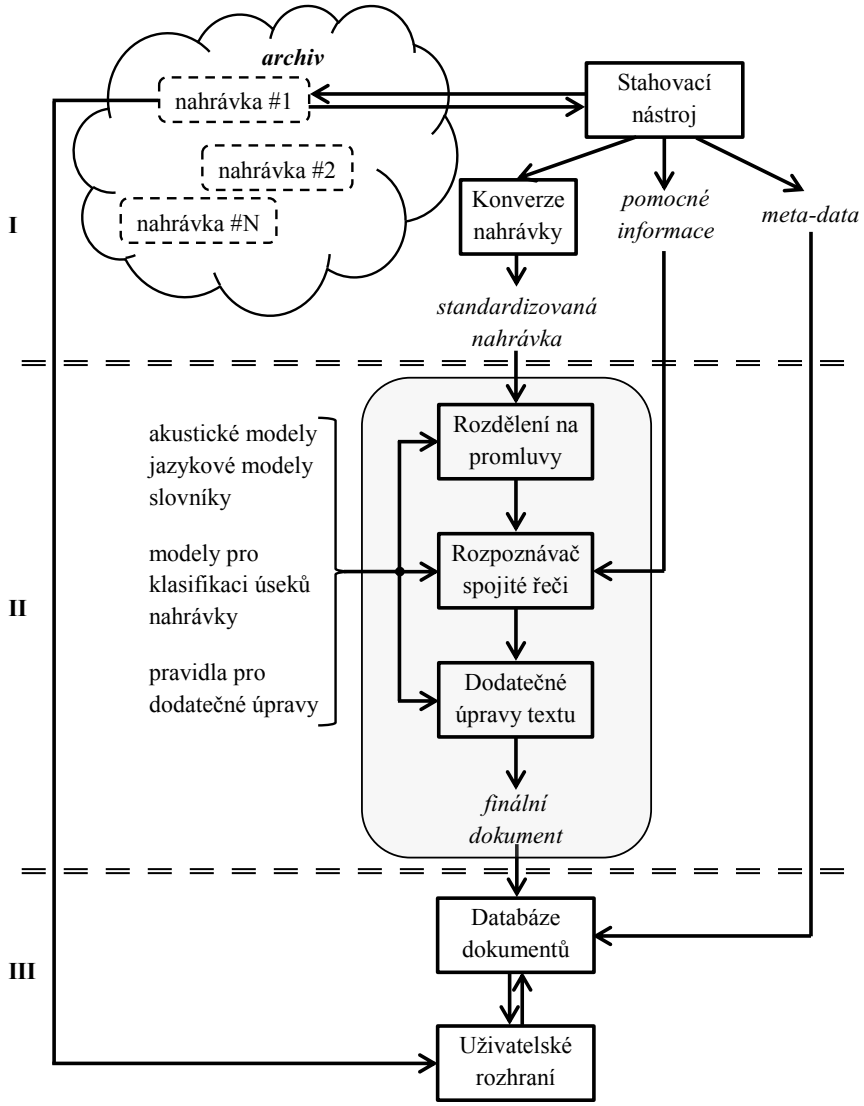
Na rozdíl od moderních multimediálních archivů, jejichž součástí bývají rozsáhlá doplňková data (výtahy ze zpravodajství, odkazy na související články, seznamy klíčových slov, anotace témat, částečné přepisy a titulky), historické archivy neumožňují efektivní vyhledávání ve svém obsahu. Tím se silně komplikuje práce badatelům (historikům, sociologům, lingvistům) a prakticky znemožňuje využití těchto cenných archivů jako doplňkových materiálů při výuce či odborných diskuzích. Je proto společensky žádoucí historické archivy zpřístupnit (nejen odborné) veřejnosti.

Vzhledem k tomu, že výše zmíněné archivy obsahují řádově statisíce hodin mluvené řeči, jediným myslitelným způsobem jejich zpřístupnění je použití technologií počítačového rozpoznání řeči. Obecné schéma inventarizace archivní nahrávky a jeho hlavní komponenty jsou zachyceny na obr. 2.1. Hlavním úkolem při inventarizaci nahrávky je pořízení jejího přesného textového přepisu. Časové značky, které jsou součástí přepisu, umožňují obousměrné provázání nahrávky s přepisem. Přepis je zaindexován do databáze, čímž umožňuje vyhledávat v obsahu celého archivu. Chcemeli zajistit komfort práce s dokumentem a potenciálně zvýšit přesnost přepisu, provádí se řada kroků, které souhrnně označujeme jako strukturalizaci dokumentů.

První vrstva inventarizace nahrávky (obr. 2.1/vrstva I) má za cíl normalizovat vstupy všem následujícím nástrojům (konverze nahrávky) a vytvořit co nejširší informační základnu pro následující kroky. Informační základna obsahuje informace nutné pro indexaci dokumentu (např. datum vzniku nahrávky, název pořadu, vysílací stanici, popis pořadu, jména hostů) a informace využitelné pro lepší funkci rozpoznávacího systému (specifickou slovní zásobu, vlastní jména).

Druhá vrstva inventarizace dokumentu (obr. 2.1/vrstva II) provádí rozpoznání nahrávky a strukturalizaci dokumentu. Strukturalizace si klade čtyři hlavní cíle: 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.

Úkolem třetí vrstvy inventarizace nahrávky (obr. 2.1/vrstva III) je zpřístupnění archivu uživateli. Požadavky jsou kladeny především na rychlost odezvy uživatelského rozhraní a na ergonomii zobrazení nalezených výsledků (intuitivní a rychlé zjištění, jestli nalezený dokument je dokumentem hledaným). Sekundárním požadavkem může být možnost editovat dokumenty (a výsledky editací promítat do databáze).



Obrázek 2.1: Základní schéma inventarizace archivní nahrávky

Většina práce představené v následujících kapitolách vznikla v rámci projektu Ministerstva kultury NAKI[1] [3]. Ten měl za cíl zpřístupnit převážně zpravodajské a publicistické pořady shromážděné v archivu Českého (Československého) rozhlasu (ČRo) a probíhal v letech 2011–2014. Tato část archivu ČRo obsahuje 100.000 hodin nahrávek od 30. let 20. století až do současnosti. V následujících odstavcích uvedu základní charakteristiky těchto pořadů a nároky kladené na jejich zpracování.

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

Většina pořadů, které byly zpracovány, mají charakter hlavního zpravodajského pořadu dne. Obsahují tudíž promluvy řady různých mluvčích, vstupy nejen ze studia, ale i z terénu, telefonní vstupy, 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 různě velké 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 je tomu v dříve zmíněných projektech MALACH a Paměť národa).

Zpracované nahrávky pochází ze dvou zdrojů. Prvním 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. Tyto 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 mnohem více informací, včetně stručných popisů obsahu pořadu. Nahrávky zpřístupněné iRádiem jsou obvykle ve formátu MP3, který není pro zpracování řeči optimální (komprimace zasahuje od přenosového pásma řeči). Část pořadů (cca 2.000 hodin) byla zpracována za použití ručních přepisů vytvořených spoluřešitelskou společností, která se zabývá monitoringem médií. Tato podskupina přepisů se vyznačuje jak vysokou kvalitou přepisů, tak informací o identitě mluvčích. Jejich jedinou nevýhodou je, že některé pasáže, nezajímavé pro klienty, nejsou přepsány.

Projekt, který budeme označovat *NAKI*, 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.

Za součást systému je třeba považovat i uživatelské rozhraní. Zanesení zpracovaného dokumentu do vyhledávacího indexu využívá komponenty jako databáze mluvčích, propojení s původním zdrojem nahrávky či konverze nahrávky do formátu vhodného pro streamování. Na základě akustické kvality nahrávek (množství hluku na pozadí řeči, nahrávacím řetězci) je účelné odlišit segmenty s kvalitním zázna-

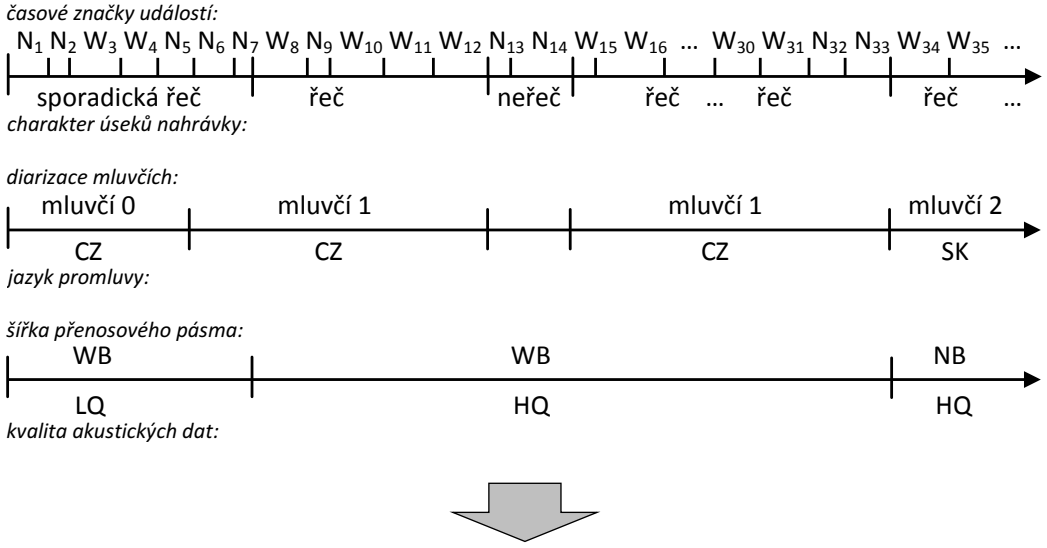
mem řečového obsahu (*HQ* - high quality) a na úseky s nízkou kvalitou (*LQ* - low quality), jejichž obsah můžeme indexovat s menší vahou. Pro rozhraní je důležitá i informace o typu neřečových úseků, protože řadu z nich je zbytečné zobrazovat (např. přepis písně uprostřed zpravodajské relace je irelevantní pro vyhledávání). Tuto informaci můžeme označit např. *show/hide*. Ilustrace vstupů a požadovaného výstupu strukturalizačního systému je zobrazena na obr. 2.2.

Uživatelské rozhraní navázané na strukturalizovaný dokument má dvě hlavní funkce. První funkcí je rychlé vyhledávání v zaindexovaných datech, umožňující co nejširší výběr omezujících podmínek (časové rozmezí, stanice, jazyk, mluvčí, kritérium relevance pro třídění nalezených dokumentů atd.). Vyhledávací rozhraní má i doplňkové funkce (např. zobrazení počtu nalezených výskytů hledané fráze), které umožní zjistit, jestli byl zadaný dotaz dostatečně konkrétní. Užitečné je i zobrazení "náhledů" nalezených dokumentů a některých informací o nich (viz obr. 8.1).

Druhou funkcí je zobrazení nalezených výsledků takovým způsobem, který umožní uživateli efektivně pracovat s nalezeným dokumentem. Rozhraní pro práci s konkrétním nalezeným dokumentem je ukázáno na obr. 8.2. V horní části rozhraní jsou zobrazeny dostupné doplňkové informace. Mezi ty patří zařazení dokumentu ve struktuře archivu (stanice, pořad, čas vysílání) a stručný popis pořadu (v tomto případě pocházející z webu ČRo). Orientaci v přepisu usnadňuje časová osa. Ta zobrazuje střídání jednotlivých mluvčích v pořadu (v našem případě muže a ženy), výskyty hledané fráze a aktuální pozici v dokumentu. Klíčovou komponentou je pak zobrazení samotného textového obsahu dokumentu. Přepis je strukturován do odstavců podle promluv jednotlivých mluvčích. Textový přepis byl formátován pro zvýšení čitelnosti. Výskyty hledané fráze jsou zvýrazněny a aktuálně přehrávaný text je zobrazen červeně. Navigace v dokumentu je možná jak skrz přepis, tak přes časovou osu (což uživateli umožňuje např. přeskakovat promluvy některých mluvčích apod.).

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 2.2: Ilustrace vstupů a výstupu strukturalizace mluveného dokumentu

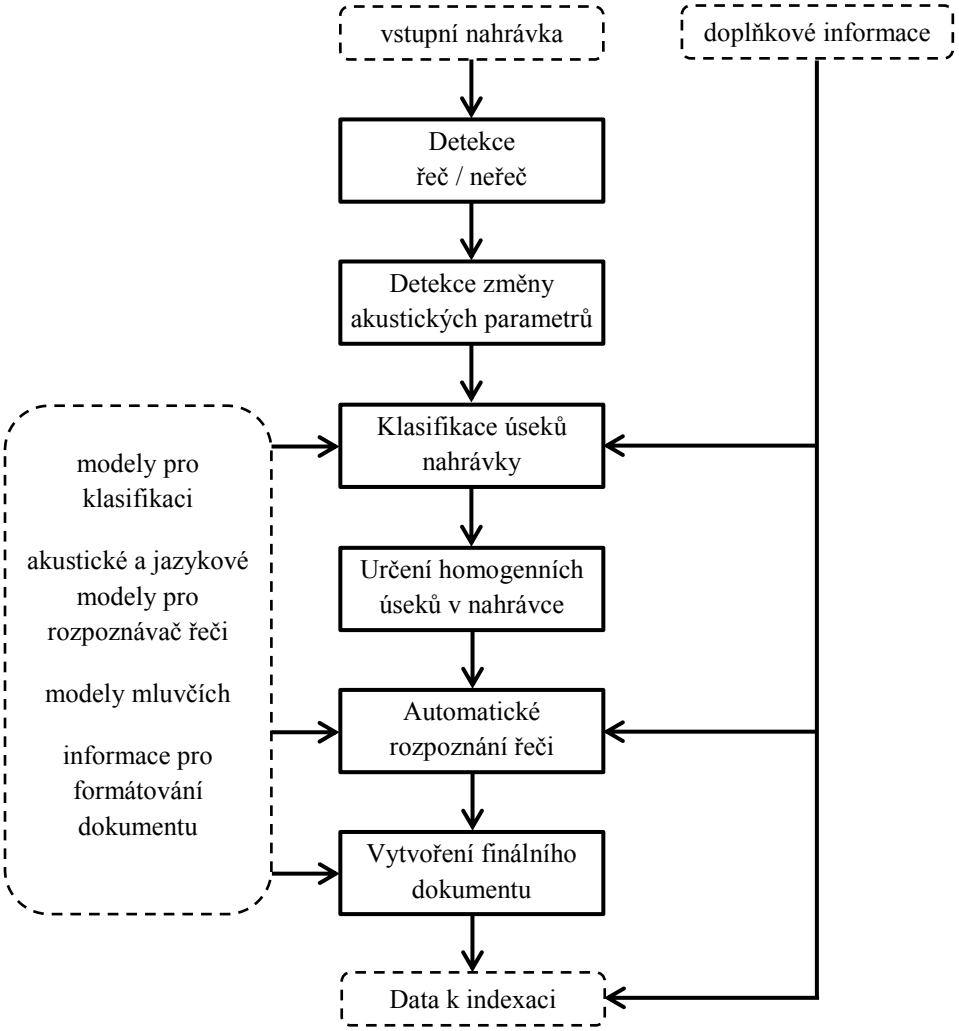
3 Aktuální stav problematiky

První úspěšné snahy o automatické zpracování multimediálních archivů obsahujících mluvené slovo za využití systémů rozpoznávání řeči lze datovat do druhé poloviny 90. let 20. století. Za první prakticky použitelný výsledek lze považovat systém vyvinutý v USA [4, 5, 6]. Systém SpeechFind umožňoval základní segmentaci nahrávky v míře nutné pro její rozpoznání. Následně jednotlivé segmenty rozpoznal a výsledky rozpoznání (text a časové značky), spolu s meta-daty získanými při inventarizaci nahrávky, zaindexoval (a tím zpřístupnil pro vyhledávání a navigaci v nahrávce).

Další významné kroky byly podniknuty od roku 2005 v Holandsku [7, 8] a v rámci mezinárodního projektu MALACH [1, 2]. MALACH je ambiciózní projekt, který má za cíl shromáždit rozhovory s pamětníky holocaustu a takto vytvořený archiv zpřístupnit. Jeho výjimečnost spočívá ve velkém počtu jazyků (ale také dialektů a přízvuků) které se v archivních nahrávkách vyskytují a ve značné míře emocionality promluv. Pamětníci jsou již pokročilého věku, což má vliv na srozumitelnost jejich řeči. Oporou při zpracovávání nahrávek jsou protokoly pořízené spolu s nahrávkou, které obsahují značné množství meta-dat, ale i důležitá vlastní jména apod.

Časový odstup mezi výzkumem probíhajícím v USA a Evropě lze zdůvodnit dvěma významnými faktory. První důvod spočívá v tom, že pro dosažení dostačujícího pokrytí slovní zásoby obsahuje anglický slovník cca 65.000 položek, zatímco tvaroslovně bohatší evropské jazyky jich vyžadují až statisíce. To se odráží v potřebě výrazně většího výpočetního výkonu, respektive ve značném nárůstu doby zpracování obdobně dlouhých dokumentů. Druhý důvod časového odstupu spočívá v pozdějším zahájení digitalizace významných archivů mluveného slova (např. digitalizace archivu ČRo byla zahájena v roce 2003).

Společným atributem všech systémů, které byly pro zpracování nahrávky dosud navrženy, je využití systému rozpoznání řeči ke zjištění obsahu nahrávky a provedení nejrůznějších kroků k zajištění maximální přesnosti tohoto přepisu. Obecný rámec provedení rozpoznání je zobrazen na obr. 3.1. Rozdíly mezi systémy lze nalézt především v posloupnosti kroků, které dělí nahrávku na jednotlivé segmenty a zjišťují optimální nastavení rozpoznávacího systému pro zpracování těchto segmentů. Rozdíly lze nalézt i v množství zjišťovaných doplňkových informacích a přesnosti, s jakou je určený přepis lokalizován a indexován. Většina systémů pak zobrazuje textový obsah přepisu ve formě jakýchsi “titulků“. To znamená, že z úloh strukturalizace nahrávky téměř neprovádí kroky související se zobrazením či prezentací získaného přepisu, ale chápou přepis jenom jako doplňkový materiál původní nahrávky a podklad umožňující vyhledávání v archivu.



Obrázek 3.1: Rámcové schéma rozpoznání nahrávky a strukturalizace dokumentu

Některá schémata se od zobrazeného rámce odlišují tím, že mají k dispozici ručně vytvořený přepis nahrávky. V takovém případě není nutné použití systému rozpoznání spojité řeči, postačuje doplnit do existujícího přepisu časové značky. Úloha se obvykle nazývá nucené zarovnání nahrávky s přepisem (anglicky forced alignment). Nástroje pro její provedení jsou odvozeny z rozpoznávačů spojité řeči.

3.1 Metody počítačového zpracování řeči

V následujících odstavcích budou shrnuty principy fungování klíčových systémů počítačového zpracování řeči. Jako první uvedeme rozpoznávač spojité řeči. Nedodržíme tím sice pořadí, v jakém jsou nástroje použity při zpracování nahrávky (obr. 3.1), ale umožní nám to zavést základní pojmy a terminologii nutnou k dalšímu výkladu.

3.1.1 Systém pro rozpoznání spojité řeči (ASR)

Systémy rozpoznávání (spojité) řeči (ASR - automatic speech recognition) jsou komplexní nástroje, které aplikují znalosti z oblasti počítačového zpracování signálů a také z oblasti zpracování (přirozeného) jazyka. Jejich základní myšlenkou je zpracovat vstupní signál (akustický záznam promluvy) a převést ho do textové podoby. K tomu nejprve převádí digitalizovaný signál do prostoru příznaků, v němž se dekodér snaží přiřadit signálu nejpravděpodobnější obsah – fonémy, slova až celé promluvy.

Statistický přístup k úloze rozpoznání řeči spoléhá na kombinaci akustického procesoru a lingvistického dekodéru [9]. 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 (3.1)

$$\hat{W} = \arg\max_W P(W|\mathbf{O}) = \arg\max_W \frac{P(W)P(\mathbf{O}|W)}{P(\mathbf{O})} \tag{3.1}$$

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 (3.2).

$$\hat{W} = \arg\max_W P(W)P(\mathbf{O}|W) \tag{3.2}$$

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

Převod signálu do prostoru parametrů lze stručně popsat takto. Nejprve je signál rozdělen do kratších úseků – tzv. rámců (v dalším výkladu budu používat hojně využívaný anglický termín frame, neboť překlad rámec nepovažuji za optimální). Tyto framy volíme s ohledem na stacionaritu příznaků uvnitř framu (chceme, aby se příznaky uvnitř rámce příliš neměnily) a také s ohledem na délku fonetických jednotek jazyka, které chceme modelovat (frame musí být kratší než tyto jednotky). Obvykle se proto volí délka framu okolo 20 ms a hranice framu se postupně v signálu posouvají (posuv se běžně volí polovina délky framu). V okamžiku, kdy je signál rozdělen na jednotlivé rámce (jejichž pořadí a index odpovídá časové lokalizaci detekovaných jevů v nahrávce), je možné přistoupit k parametrizaci obsahu jednotlivých framů.

Parametrizace obvykle vychází ze spektrálního, nebo kepstrálního popisu signálu a společného předpokladu, že minimální přenosové pásmo nutné pro zachycení informace v řečovém signálu je 4 kHz (minimální vzorkovací frekvence 8 kHz). Používaných parametrizací je celá řada: Mel-frekvenční Kepstrální Koeficienty (MFCC), Perceptual Linear Prediction (PLP), Linear Prediction Coefficients (LPC), tzv. banky filtrů [10] či bottle-neck příznaky[11]. Z koeficientů získaných parametrizací a případně tzv. delta-příznaků (diferencí mezi příznakem v aktuálním framu a framy předcházejícími) jsou formovány příznakové vektory, které popisují daný frame ve zvoleném příznakovém prostoru.

Podle příznakových vektorů chceme identifikovat akustické jednotky, ze kterých se skládá lidská řeč. Tyto jednotky jsou specifické pro každý jazyk a za nejmenší stavební jednotku řeči považujeme tzv. foném. Fonémy můžeme modelovat jako sadu nezávislých jevů (context independent - CI), pak hovoříme o tzv. monofonech, nebo bereme v úvahu vztah fonému a jeho okolí (context dependent – CD). V případě CD popisu je nejrozšířenější popis pomocí tzv. trifonů. Předpokládá se vliv předcházejících a následujících fonémů na foném modelovaný (modelujeme foném v kontextu jeho okolí, obvykle pomocí tří stavů). Během procesu trénování popisu jednotlivých akustických jednotek může být zjištěna výrazná podobnost některých jednotek a jejich modely jsou pro účely sloučeny. Po tomto sloučení získáváme sadu modelovaných stavů řeči, které se v případě trifonového akustického modelu nazývají senony (svázané stavy akustického modelu), jejichž věrohodnosti jsou vstupem dekodéru rozpoznávače.

Jednotlivé stavy jsou modelovány dvěma základními přístupy. První přístup modeluje stav jako směs gaussovských rozložení (GMM) příznakových vektorů. Druhý přístup využívá hluboké neuronové sítě (DNN), na jejichž vstupu jsou příznakové vektory a její výstupní vrstva vyčísluje věrohodnost jednotlivých fyzických stavů.

Samotná promluva je obvykle modelována jako skrytý markovský proces (HMM), u kterého předpokládáme lineární posloupnost jednotlivých stavů (lze buď setrvat ve stavu, nebo přejít do stavu následujícího). Ve fázi trénování modelů jsou pak pro všechny stavy (fyzické stavy, které tvoří fonetiku jednotlivých slovníkových položek) určeny pravděpodobnosti setrvání ve stavu/přechodu na stav následující. Uvažujeme-li u GMM modelu rozsáhlá trénovací data, je možné popsat j-tý fyzický stav rozpoznávače M gaussovskými rozloženími, kdy každé rozložení (obvykle označované jako mixtura) má vlastní střední hodnotu $\bar{\mathbf{o}}_{jm}$, rozptyl a váhový koeficient c_{jm}. Pravděpodobnostní hustota $b(j, \mathbf{o_i})$, že frame popsaný příznakovým vektorem $\mathbf{o_i}$ (délky R) přísluší j-tému stavu rozpoznávače, je popsána vztahem (3.3)

$$b(j, \mathbf{o_i}) = \sum_{m=1}^{M} c_{jm} \frac{1}{\sqrt{(2\pi)^R det\Sigma_{jm}}} \exp[(\mathbf{o_i} - \bar{\mathbf{o}}_{jm})^T \Sigma_{jm}^{-1} (\mathbf{o_i} - \bar{\mathbf{o}}_{jm})] \tag{3.3}$$

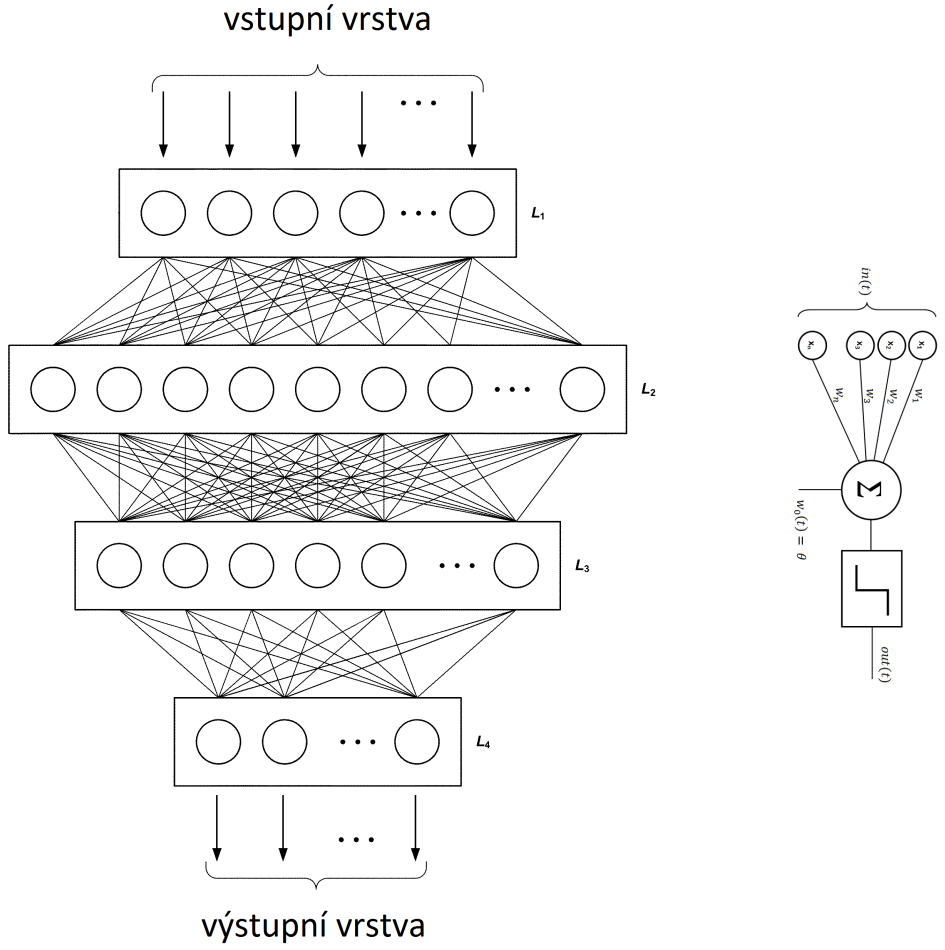
kde m zastupuje jednotlivé mixtury GMM modelu, Σ_{jm} je kovarianční matice příznakových vektorů určená během trénování modelů. Argument exponenciely odpovídá druhé mocnině tzv. Mahalanobisovy vzdálenosti.

GMM modely jsou generativního charakteru (plný pravděpodobnostní model všech proměnných) a za určitých úprav je lze trénovat jako diskriminativní [12], tj. poskytující model cílových proměnných závislý na dostupných pozorováních.

Druhou rozšířenou možností, jak vyčíslit podobnost příznakového vektoru s modelovaným stavem akustického modelu, je použití neuronových sítí. Současný stav problematiky spoléhá na využití hlubokých neuronových sítí (DNN). Nejčastější aplikací DNN je struktura, na jejíž vstupní vrstvu přivádíme příznakový vektor zkoumaného framu (obvykle i více framů okolních) a jednotlivé prvky výstupní vrstvy určují pravděpodobnost, že se jedná o konkrétní fyzický stav rozpoznávače (senon). Chování sítí je dáno jejich topologií a procesem trénování. Pod pojem topologie DNN lze zahrnout počet a šířku vrstev, ze kterých se síť skládá, použitou aktivační

funkci a případně normování hodnot na výstupu sítě. Hodnoty pravděpodobnosti získané na výstupní vrstvě je pak nutné normovat do rozsahu, který je zpracováván Viterbiho dekodérem, tj. získat obdobu logaritmované věrohodnosti (log-likelihood), kterou generuje GMM model. Ilustrace DNN[1] je na ukázána obr. 3.2.

DNN modely jsou nativně diskriminativní, což je dáno (mimo jiné) strukturou trénovacích dat, kdy pro dostupná trénovací data (příznakové vektory) značíme požadovanou hodnotu výstupní vrstvy (který výstupní neuron – senon – se má daným vstupem aktivovat).



Obrázek 3.2: Ilustrace struktury hluboké neuronové sítě (DNN)

Úkolem Viterbiho dekodéru je nalezení nejpravděpodobnější posloupnosti skrytých stavů (stavů HMM – senonů) v nahrávce. Tato posloupnost odpovídá textu rozpoznanému v nahrávce a skládá se z jednotlivých položek slovníku. Detekovaná posloupnost je závislá na věrohodnosti detekce konkrétních senonů a na pravděpodobnostech výskytu detekovaných slovních sekvencí (jež jsou popsány jazykovým modelem). Vzhledem k obrovské výpočetní náročnosti takové úlohy vychází dekodér

[1]http://www.google.com/patents/US8527276; https://sk.wikipedia.org/wiki/Perceptrón

z principů dynamického programování a obsahuje řadu optimalizací (např. prořezávání dekódovaných variant – tzv. pruning).

Slovník musí obsahovat fonetickou anotaci všech svých položek. Právě fonetická podoba slovníkových položek je "překládána" na posloupnost senonů a je vstupem pro dekodér. Fonetická transkripce je vytvářena ručně (což umožňuje pokrýt nespisovný jazyk a nářečí) na základě sady pravidel nebo pomocí nástrojů strojového učení (např. WFST [13]).

Účelem stochastického jazykového modelu je stanovit pro každou posloupnost slov W její apriorní pravděpodobnost $P(W)$. Pravěpodobnost posloupnosti K slov lze popsat vztahem (3.4):

$$\begin{gathered} P(W) = P(w_1^K) = P(w_1\, w_2\, w_3 ... w_K) = \\ P(w_1)P(w_2|w_1)P(w_3|w_1 w_2)...P(w_K|w_1 w_2 ... w_{K-1}) = \prod_{i=1}^{K} P(w_i|w_1^{i-1}) \end{gathered} \tag{3.4}$$

Pro pravděpodobnost libovolné počáteční části $w_1\, w_2 ... w_k (k \leq K)$ obdobně platí:

$$P(w_1^k) = P(w_1^{k-1})P(w_k|w_1^{k-1}) \tag{3.5}$$

Pravděpodobnost slova w_i je podmíněna pouze historií $w_1 ... w_{i-2}\, w_{i-1}$, což je výhodné pro implementaci v dekodéru systému rozpoznání řeči.

Stochastický n-gramový model aproximuje posloupnost $w_1 ... w_{i-2}\, w_{i-1}$ na základě shody posledních $n-1$ slov posloupnosti. Pojmem n-gram pak rozumíme posloupnost n za sebou jdoucích slov v pozorování jejich náhodného výběru. Nejpoužívanější jsou bigramy (n=2) a trigramy (n=3). Apriorní pravděpodobnost sekvence slov je pak aproximována vztahem (3.6).

$$P(w_1^k) \approx \prod_{i=1}^{k} P(w_i|w_{i-n+1}^{i-1}) \tag{3.6}$$

Jazykový model může být kromě tvaru n-gramového modelu tvořen i pomocí WFST [14] nebo neuronových sítí [15]. WFST využívá svůj nativní paralelismus a s ním spojenou možnost získání M-nejlepších přepisů či tzv. lattice [16].

V následujícím textu narazíme na dvě základní metody, které mohou zvýšit přesnost výstupu systému pro rozpoznávání řeči. První metodou je rozpoznání nahrávky více rozpoznávači a následná kombinace dostupných přepisů. Mezi taková řešení patří např. systém ROVER [17]. Druhou metodou pro zvýšení přesnosti přepisu je adaptace příznakových vektorů na konkrétní nahrávku. Ta vychází z předpokladu, že konkrétní segmenty nahrávky mohou být zatíženy odchylkou od dostupných akustických modelů (hluk na pozadí, frekvenční charakteristiky zařízení zapojených do nahrávacího řetězce, řečová specifika konkrétního řečníka). Adaptační metody pak hledají takovou transformaci příznakových vektorů, která přiblíží příznaky v nahrávce příznakům akustického modelu, a tím vykompenzují specifické (nežádoucí) podmínky nahrávky.

3.1.2 Detekce řečové aktivity (VAD)

Úkolem nástrojů pro detekci řečové aktivity (VAD) je zabránit rozpoznávání úseků nahrávky, které řeč neobsahují. Tím šetří výpočetní výkon a brání možnému rozpoznání neexistujícího obsahu v nahrávce.

Základní přístup kombinuje energetický detektor (pro detekci ticha) a modely možného obsahu nahrávky (např. hudba, řeč-muž, řeč-žena, úzké/plné přenosové pásmo, překrývající se řeč, hluk). Modely tohoto obsahu obvykle využívají stejnou parametrizaci jako systém rozpoznání řeči. Nejobvyklejší jsou proto GMM modely s MFCC nebo Waveletovými příznaky [18]. Nejnovější přístupy kombinují několik různých parametrizací, které jsou vstupem DNN. Ta pak rozhoduje, zda frame obsahuje řečový, či neřečový typ obsahu [19]. Společnou vlastností všech výše zmíněných variant je nutnost mít k dispozici dostatek anotovaných nahrávek všech požadovaných kategorií obsahu nahrávky.

Alternativou mohou být metody, které vycházejí z vlastností samotného řečového signálu a nepotřebují proto trénovací data. Příkladem může být metoda Single Frequency Filtering [20]. Ta detekuje řečovou aktivitu podle absence/přítomnosti fundamentální frekvence řeči v signálu.

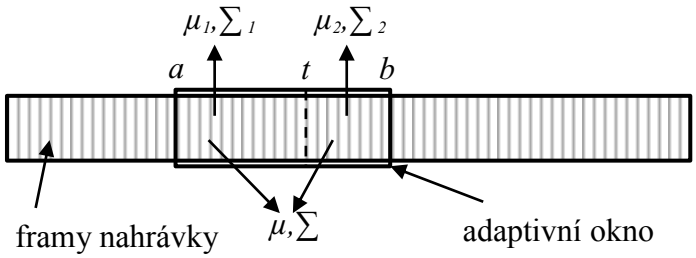
V případě, že detektor nemusí pracovat on-line provádí se vyhlazení výstupu VAD. Vyhlazení může být založeno na redukci dostupných kategorií (mužská i ženská řeč je stále řeč), heuristických pravidlech (např. minimální délka segmentu) či stavových automatech, které optimalizují délku úseků a četnost jejich střídání.

3.1.3 Detekce změny v nahrávce

Chceme-li správně strukturovat nahrávku, klíčovým nástrojem se stává detektor změny akustických parametrů nahrávky. Obvykle se předpokládá změna mluvčího (tzv. speaker change point detection), ale zajímají nás i další změny charakteru signálu (řeč/hudba/píseň/hluk/šum na pozadí). K popisu framů nahrávky se v kontextu hledání bodu změny používají různé mixy příznaků (např. četnost průchodů nulou, energie signálu, MFCC, LPC a další).

Popíšeme-li jednotlivé framy nahrávky pomocí příznakových vektorů, je třeba zodpovědět dvě otázky: 1) jak určit míru podobnosti dvou úseků a 2) jaké úseky nahrávky porovnávat. Porovnání podobnosti dvou sousedících úseků nahrávky obvykle využívá bayesovské informační kritérium (BIC - Bayessian Information Criterion). BIC [21] určuje míru podobnosti dvou sad parametrů – před a za potenciálním bodem změny (a-t;t-b), jak je naznačeno na obr. 3.3. Každá sada parametrů je tvořena příznakovými vektory framů nahrávky v rámci zkoumaného okna.

Pozice bodů a,t a b mohou být v nahrávce umístěny kdekoli a jejich pozice je proto nutné nějak omezit. První možný přístup omezuje vzájemné rozestupy bodů (minimální/maximální délka zkoumaného okna, odstup bodu změny od začátku okna atd.). Druhá možnost omezuje pozice bodů a,t a b na sloty mezi počítačem rozpoznanými slovy. V takovém případě se výpočetní náročnost úlohy výrazně sníží a současně je vyřešena synchronizace výsledků s přepisem.



Obrázek 3.3: Ilustrace hledání změny akustických parametrů v nahrávce v čase t mezi začátkem a a koncem b adaptivního okna

Okno může mít fixní délku (jednu nebo více fixních délek), nebo adaptivní délku. Metody s adaptivní délkou se jeví jako výpočetně méně náročné a přesnější, zatímco analýza pomocí okna s fixní délkou se využívá jako zdroj více "slabších klasifikátorů". K nejnovějším trendům patří rozhodování pomocí více slabších klasifikátorů, ke kterému se užívají neuronové sítě, jak je ukázáno například v [22]. Jako pomocný příznak změny akustických vlastností v nahrávce lze použít delší úseky ticha (cca 1,5 sekundy), které jsou obvykle způsobeny změnou mluvčího, nebo zdroje zvuku [23].

Po nalezení bodů změny v nahrávce je možné (avšak ne nutné) přistoupit k tzv. diarizaci nahrávky [18]. Cílem diarizace je určit, jestli některé úseky nahrávky byly proneseny stejným mluvčím. V prvním kroku je každému úseku nahrávky přidělena unikátní identita. V následujících iteracích jsou vyhledávány nejpodobnější páry úseků nahrávky, které jsou v případě dostatečné shody sloučeny, a podobnost párů je vyhodnocena znovu. Podobnost je typicky určena pomocí BIC, ale může být využito i více různých metrik podobnosti. V případě použití více metrik se obvykle staví hierarchický systém, kdy jedna metrika předklastruje úseky nahrávky a druhá metrika slouží k určení výsledné diarizace dokumentu [24].

3.1.4 Klasifikace charakteru úseků nahrávky

Předpokladem pro klasifikaci obsahu nahrávky je znalost bodů změny (3.1.3). Cíle klasifikace (skrze ni i strukturalizace dokumentu) jsou dva:

- zajistit správné nastavení ASR systémů – AM a LM
- zajistit meta-data pro zobrazení a indexaci dokumentu

Výběr jazykového modelu (language model – LM) závisí na jazyku promluvy (případně jeho tématu), akustický model (acoustic model – AM) je určován podle šířky přenosového pásma, jazyka, pohlaví mluvčího a odstupu signálu od hluku (SNR). Chceme-li odlišit akusticky rozdílné jazyky (jejich fonémové sady jsou výrazně odlišné), je běžné použít GMM modely jednotlivých jazyků [25]. V případě, že jsou si jazyky akusticky podobné (např. čeština a slovenština, španělština a italština), je nutné najít specifické řešení vhodné pro danou kombinaci. Správný odhad tématu promluvy může vést ke zpřesnění výsledného přepisu, jak je ukázáno v [26]. Nutnou podmínkou je dostatek anotovaných dat jak pro trénování systému schopného určit téma, tak pro stavbu konkrétních jazykových modelů.

Určení parametrů přenosového pásma (přítomnost telefonů a podobných zařízení v nahrávacím řetězci) lze provést buď porovnáním energie v různých frekvenčních pásmech, nebo natrénováním klasifikátorů (GMM, neuronové sítě). Podobné přístupy se používají i k odhadu míry zašumění signálu (ASR).

Pohlaví mluvčího lze určit podle výšky jeho/jejího hlasu [27], nebo na základě natrénovaných modelů (typicky opět GMM). Pro určení identity mluvčího existuje nepřeberné množství kombinací modelů mluvčích a metrik jejich podobnosti (nejnovější směr výzkumu opět využívá hluboké neuronové sítě k porovnání podobnosti různých modelů s daty popisujícími nahrávku [28]).

3.2 Existující systémy

Existující systémy vytvořené ke zpřístupnění multimediálních archivů lze rozdělit do dvou skupin. První z nich využívá existující ručně vytvořený přepis a zabývá se propojením obsahu přepisu s audio (video) souborem (RadioOranje [8], InForMedia [5], TaiwanNews [29]). Druhá skupina zpřístupňuje obsah nahrávek na základě počítačového rozpoznání jejich obsahu (MALACH [1, 2], SpeechFind [30], SPRACH [6]). Na první pohled by se sice mohlo zdát, že nucené zarovnání přepisu pořadu se zvukovou stopou nemá mnoho společného s automatickým zpřístupněním archivů. Řada modulů (např. segmentace nahrávky, zpracování doplňkových informací a indexace výsledků) je však v obou případech velmi podobná.

3.2.1 Využití existujícího textového přepisu

Ačkoliv se úloha zarovnání existujícího textu s nahrávkou může jevit jako velmi specifická, řada dnešních zpravodajských webů je vhodná pro její nasazení. Některé pořady (televizní i rozhlasové) obsahují buď úplné přepisy svého obsahu, přepisy zajímavých částí, nebo odkazy na obdobné zprávy. Jelikož se znění některých zpráv přebírá doslovně, mohou i dílčí přepisy vést ke zlepšení výsledného přepisu. Významnou roli hraje možnost rozšíření slovní zásoby o nové položky z takových textů.

RadioOranje [8] je příkladem archivu historicky významných nahrávek. Pro holandskou veřejnost jsou natolik zajímavé, že již dříve byly pořízeny (více či méně) kompletní přepisy těchto rozhlasových projevů. Autoři systému se snaží využít existující přepisy pro získání co nejpřesnějšího textového přepisu a současně využívají maximum dostupných "doplňkových informací", aby umožnili efektivní vyhledávání ve výsledném archivu. Přepisy se nemusí doslovně shodovat s obsahem nahrávky (některé části mohou zcela chybět, anotátor mohl stylisticky reformulovat některé fráze, mluvená řeč obsahuje různé nespojitosti apod.). Autoři proto navrhují robustnější řešení, schopné vyrovnat se s těmito fenomény. Navržené řešení spočívá v rozpoznání nahrávky systémem rozpoznání řeči a následném zarovnání rozpoznaného textu s manuálním přepisem. K tomu využívají algoritmus vycházející z principů dynamického programování (vychází z algoritmu Minimum Edit Distance [31]). CI-GMM-HMM rozpoznávač řeči používá monofonové modely a rozšířenou slov-

ní zásobu získanou z ručního přepisu. V případě, že se ruční přepis a rozpoznaný text výrazně liší, používají autoři pro indexaci rozpoznaný text. Vzhledem k tomu, že zpracované dokumenty mají formát projevu, nemusí se autoři věnovat otázkám strukturalizace nahrávky (změny mluvčího apod.).

O něco komplexnější zadání je řešeno v projektu **InForMedia** [5]. Ten je zaměřen na monitoring médií. Jedná se o jednotný vyhledávací systém, který zpracovává televizní vysílání, rozhlasové zpravodajství a textové zprávy z internetových portálů. Jako vstupní přepis používají autoři skryté titulky a obsah teletextu. Segmentaci textu pak provádějí pomocí ručního formátování obsaženého v textu (převážně podle interpunkce). Data získaná z teletextu jsou zarovnána klasickou implementací nuceného zarovnání (není robustní vůči reformulacím). Jednotlivé načasované fráze jsou pak tématicky klasifikovány metodami "term frequency" a "inverse document frequency", což umožňuje detekci změny tématu (čili hranice jednotlivých zpravodajských příspěvků).

TaiwanNews [29] řeší velmi komplikovanou úlohu, kdy neexistence jednotné psané formy tajvanštiny neumožňuje přímou aplikaci metod rozpoznávání řeči na zpravodajské pořady ani vyhledávání, protože jednotliví uživatelé se neshodnou na správné psané formě slov. Autoři místo toho provádí mezijazykové párování médií, konkrétně čínských textů a tematicky podobných (případně zcela shodných) zpráv v tajvanštině. Standardní čínština je známa všem potenciálním uživatelům. Celý systém pracuje v následujících krocích:

- rozdělení vysílání zpráv na jednotlivé "příběhy",
- rozpoznání obsahu jednotlivých příběhů,
- překlad obsahu čínských zpráv,
- zarovnání s dostupnými texty zpráv => výběr nejbližšího obsahu,
- zaindexování výsledků.

Rozdělení zpravodajské relace na jednotlivé zprávy využívá znalosti obecné struktury tohoto vysílání (znělky, předěly, zprávy, reklamy, typická délka trvání vstupu atd.). Autoři natrénovali GMM model jednotlivých typů obsahu. Současně vytvořili HMM model celkového průběhu zpravodajské relace. Parametrizace použitá pro vytvoření GMM modelů využívá četnost průchodů nulou (ZCR), krátkodobou energii signálu, spektrální tok (spectral flux) a MFCC.

Rozpoznávač tajvanštiny je založen na detekci slabik, ze kterých jsou skládána jednotlivá slova a podle autorů dosahuje přesnosti cca 55 %. Čínské texty jsou přeloženy po jednotlivých slovech a dekomponovány na slabiky odpovídající výstupu rozpoznávače. Následně je provedeno zarovnání textu a výstupu rozpoznávače pomocí algoritmu MED. Dokument s nejvyšší dosaženou shodou je spárován se zprávou. Navzdory poměrně nízké přesnosti rozpoznávače řeči a nutnosti provádět překlad se autorům daří přiřadit správný čínský článek 85 % zpracovaných zpráv.

3.2.2 Využití automatického rozpoznání řeči

V následující pasáži se již budu zabývat systémy, které ke zpřístupnění obsahu nahrávky využívají systém rozpoznání řeči. Popsané systémy zpracovávají dva různé druhy archivů. Prvním jsou archivy historických nahrávek (MALACH, SpeechFind), druhým jsou pak soudobé zpravodajské pořady (InForMedia, SPRACH). U historických archivů je úloha o to komplikovanější, že nelze zaručit kvalitu nahrávacího řetězce, prostředí vzniku nahrávky a v případě MALACHu je situace komplikována i stářím řečníků, které s sebou přináší určité obtíže s výslovností. Zpracování soudobých zpravodajských médií má oproti tomu k dispozici řadu doplňkových informací, obvykle kvalitní nahrávky a dostatek dat pro trénování jazykových modelů.

Nejnáročnějším projektem zaměřeným na zpřístupnění velmi variabilních nahrávek je **MALACH**. Jedná se o rozsáhlou sbírku 52.000 rozhovorů s pamětníky holocaustu pořízené Shoah Visual History Foundation[2]. Nahrávky byly pořízeny ve 32 jazycích a celková délka kolem 116.000 hodin je zřejmě největším archivem svého druhu. Vzhledem k velké variabilitě nahrávacích podmínek, přízvukům mluvčích a hlavně jejich stáří je aplikace systému rozpoznání řeči velmi komplikovaná. Nahrávky obsahují silně atypickou slovní zásobu (často pocházející z více jazyků) a mluvčí, již pokročilého věku, nemají právě precizní výslovnost. Navíc se řeč stává často emocionální a obsahuje nadprůměrné množství různých nespojitostí (váhání, opakování se, pláč). V tomto přehledu zmíním systémy vytvořené pro angličtinu, češtinu a maďarštinu.

Schéma strukturalizace nahrávky vytvořené pro zpracování anglických a českých nahrávek [2] se řídí modelovým schématem zobrazeným na obr 3.1. Prvním krokem je detekce řečové aktivity v nahrávce. Ta je následně rozdělena na úseky pronášené jedním mluvčím, provede se rozpoznání s adaptací na mluvčího a promluvy se dodatečně rozdělí podle detekovaného tématu. Závěrečná segmentace a klasifikace probíhá nad hranicemi větných celků určených v předchozím kroku. Pro nalezení hranice změny tématu a jeho určení jsou použity postupy blíže popsané v [32, 33].

Pro účely trénování měli autoři k dispozici 200 hodin anglických přepisů (800 mluvčích) a 84 hodin českých přepisů (336 mluvčích). Ke každému pořadu je dostupný protokol o nahrávání, ze kterého je možné získat doplňkové informace (např. vlastní jména, která se mohou v nahrávce vyskytovat). Nahrávání bylo prováděno pomocí dvou mikrofonů. To autorům umožnilo oddělit stopu s nahrávkou dotazovaného a stopu s nahrávkou tazatele. Tím jsou v nahrávce detekovány regiony zájmu a potlačen vliv promluv tazatele (např. současný hovor tazatele a tázaného).

Akustické modely pro anglický systém využily MFCC parametrizaci (24 příznaků, okénko 25 ms, posun 10 ms) a MLLR adaptaci na mluvčího. Jazykový model byl trénován přímo z protokolů o nahrávání. Český model využil PLP příznaky s 1. a 2. diferencí. Pro natrénování českého jazykového modelu byla použita i data mimo doménu (zpravodajství). V obou případech zůstala velká část slovní zásoby nepokryta slovníky (přes 8 %). Důvodem je výše zmíněné velké množství specifických vlastních jmen (jmen osob, názvů míst), které pocházely z řady různých jazyků.

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

Ačkoli se chybovost (WER) výsledných přepisů pohybuje okolo 40 % (způsobená vysokým množstvím nepokryté slovní zásoby a výslovností pamětníků), autoři považují tuto úspěšnost za dostatečnou pro indexování obsahu (tématu jednotlivých segmentů, rozpoznaných slov a jejich časových značek).

Autoři [1] předpokládají použití stejného systému zpracování nahrávky, jak již byl představen v předchozích odstavcích. Soustředí se proto na dosažení co nejvyšší přesnosti přepisu spontánní maďarštiny. Porovnávají výsledky použití konvenčního jazykového modelu založeného na jednotlivých slovech s použitím morfémů (a s tím souvisejícího vyššího řádu jazykového modelu). Fonetický přepis je určen pravidly a doplněn o slovník výjimek, zejména slov původem z jiných jazyků. Jedná se opět o GMM–HMM rozpoznávač řeči užívající PLP parametrizaci. Pro účely trénování akustických a jazykových modelů byly pořízeny přepisy 104 rozhovorů (26 hodin akustických nahrávek). Jazykový model užívá data pouze z těchto přepisů. Výsledná WER je opět cca 40 %.

SpeechFind [30, 34] je systém určený ke zpřístupnění National Gallery of the Spoken Word[3], archivu obsahujícího nejrůznější nahrávky pořízené v průběhu 20. století. Obsahuje politické projevy a debaty, záznamy rozhlasového a televizního vysílání a specifické nahrávky jako např. vysílání NASA. Jedná se tudíž o velmi heterogenní směs pořadů a lze u nich předpokládat i postupný vývoj jazyka (potřebných jazykových modelů). Systém lze rozdělit do tří vrstev:

- inventarizace nahrávek, získání pomocných dat a meta-dat,
- segmentace nahrávky a systém rozpoznání řeči,
- databáze dokumentů s uživatelským rozhraním.

Inventarizace nahrávek plní tři úlohy: 1) stáhne nahrávku a standardizuje její formát v souladu s potřebami systému, 2) získává meta-data důležitá pro indexaci (původ nahrávky, datum vzniku atd.), 3) získává pomocná data potřebná pro zvýšení přesnosti rozpoznávání – např. slova mimo slovní zásobu LVCSR (v dalším textu označovaná jako OOV – Out of Vocabulary), která mohou být získána např. z popisků nahrávky.

Segmentační nástroj má za úkol rozlišit tři typy změn v nahrávce: 1) změnu mluvčího, 2) změnu vlastností přenosového pásma a 3) změnu hlukových podmínek na pozadí řeči. Aby bylo možné detekovat všechny požadované změny v nahrávce, využili autoři velmi bohatou směs příznakových vektorů (PMVDR [35], SZCR, logaritomované koeficienty bank filtrů - FBLC). U některých předpokládají schopnost detekovat směny mluvčích, zatímco jiné mají analyzovat spíše pozadí řeči a přenosové pásmo. Jako měřítko podobnosti dvou segmentů používají BIC.

Autoři používají GMM-HMM rozpoznávač řeči Sphinx 3 s akustickým modelem trénovaným na podmnožině 200 h nahrávek. Systém dosahuje WER 25-40 % při méně než 1,5 % OOV. V dostupných popisech systému je důraz kladen na využití

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

meta-dat při vyhledávání (information retrieval) a na otázky související se zpracováním nahrávek z různých zdrojů (např. licenční podmínky přehrávání nalezených dokumentů atd.).

Výše zmíněný projekt **InForMedia** zpracovává i rozhlasové zpravodajství. Pro tyto pořady nejsou k dispozici skryté titulky a autoři proto používají rozpoznávač spojité řeči. Konkrétně nasadili GMM-HMM rozpoznávač Sphinx 2, využívající MFCC parametrizaci s 1. a 2. diferencí příznaků. Detaily o zdroji akustických dat nebo korpusů pro jazykové modely nejsou uvedeny. Segmentace je řešena vyhledáním nízkoenergetických úseků v nahrávce (ticha), které jsou hledány v přibližně 30s odstupech. Délku úseku 30 s považují autoři za optimální pro prezentaci nalezených výsledků v uživatelském rozhraní.

SPRACH [6] je systém určený ke zpracování nahrávek anglicky mluvených zpravodajských pořadů. Systém provádí nejprve segmentaci nahrávky, následně rozpozná jednotlivé úseky pomocí několika různých akustických modelů a nakonec přepisy sloučí do jedné výsledné hypotézy. Pro segmentaci je použit nástroj vyvinutý na Cambridge University [36, 37]. Jedná se o komplexní systém, který funguje v několika vrstvách:

- Segmentace nahrávky
- Vyřazení segmentů obsahujících hudbu
- První průchod rozpoznávačem, po kterém určí pohlaví mluvčích
- Likvidace dlouhých segmentů obsahujících ticho
- Vyhlazení segmentace a klastrování

Po provedení segmentace (a klasifikace některých segmentů jako neřečových) jsou řečové úseky rozpoznány. Autoři použili 200 h audio nahrávek k natrénování tří akustických modelů: CI-RNN-HMM, CD-RNN-HMM a CI-MLP-HMM. Pro všechny rozpoznávače je použit stejný jazykový model (korpus o rozsahu 450.000.000 slov, 65.000 slov ve slovníku). Výslovnosti byly generovány automatickým nástrojem založeným na rozhodovacích stromech [38]. Po té, co všechny rozpoznávače vygenerují možné přepisy segmentu (ve formě lattice), je konečná hypotéza určena systémem ROVER [17].

Autoři se při popisu systému příliš nezabývají otázkami indexace výsledků a zahrnutí doplňkových informací do vyhledávacího indexu. Soustředí se především na vhodnou segmentaci nahrávky a dosažení co nejpřesnějšího přepisu. V tomto ohledu se jedná o nejkomplexnější a technologicky nejpokročilejší ze všech představených řešení (ačkoli ho lze datovat již do roku 1999).

Všechny výše zmíněné systémy byly navrženy okolo roku 2000. V té době již byl k dispozici dostatečný výpočetní výkon pro první snahy o zpracování velkých archivů nahrávek mluveného slova. Současně dosáhly technologie zpracování řeči přesnosti,

která byla pro takové úlohy nezbytná. Až na výjimky byly tyto technologie nasazeny na anglická data (pro dosažení stejné míry pokrytí slovní zásoby stačí v angličtině výrazně menší slovník než např. u jazyků slovanských, což vede k výrazně menším nárokům na výpočetní výkon).

SpeechFind ukazuje, že je důležité využít veškerá dostupná doplňková data a meta-data, aby bylo možné vytvořit co nejkvalitnější vyhledávací databázi. SPRACH je příkladem systému s velmi pokročilou segmentací nahrávky a rozpoznávačem řeči. Většina systémů se zaměřila "pouze" na rozpoznání obsahu nahrávek a jeho indexaci, samotná prezentace výsledků (a k tomu nutné dodatečné úpravy) je prakticky opomíjena. Jediným aplikovaným postupem je dělení nahrávky na krátké úseky (cca 30 s) obvykle nalezené pomocí neřečových událostí v nahrávce (hluky, delší ticho). Veškerý obsah (s výjimkou textového přepisu) je tak obsažen pouze ve vyhledávací databázi a neusnadňuje uživateli orientaci v dokumentu (například zobrazením identity mluvčího, jazyka úseku, označením hudby v nahrávce).

V kontextu strukturalizace dokumentu lze tedy konstatovat, že představené systémy zajišťují podmínky pro správnou funkci systému rozpoznání řeči a zajišťují informace potřebné pro indexaci výsledků a vyhledávání. Současně je jen velmi málo pozornosti věnováno vhodnému zobrazení dokumentu a optimalizaci čitelnosti (orientaci v dokumentu).

4 Cíle práce

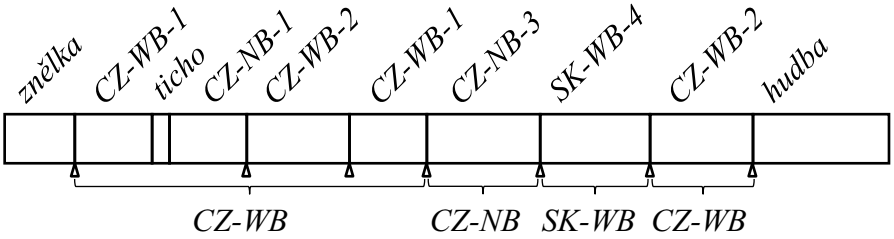
4.1 Úloha strukturalizace přepisu - tvorba informačně bohatého dokumentu

Hlavním cílem této práce je navrhnout, implementovat, experimentálně vyhodnotit a vzájemně porovnat dvě schémata strukturalizace počítačem rozpoznaného dokumentu. U obou navržených schémat požadujeme "plnou" strukturalizaci dokumentu, tj. proces, který zajistí: 1) podmínky pro správnou funkci ASR, 2) informace potřebné pro indexaci, 3) podklady pro správné zobrazení výsledného dokumentu a 4) čitelnost přepisu a orientaci v něm. Výsledný dokument vzniklý strukturalizací přepisu označujeme jako informačně bohatý, neboť zastřešuje řadu dílčích informací: textový přepis nahrávky, identitu mluvčího (jméno nebo pohlaví), jazyk promluvy, charakter přenosového pásma (telefonní vstup, nahrávka ze studia, podíl hluku na pozadí), charakter úseků bez řečového obsahu (znělky, hudba).

Jak naznačuje obr. 4.1, segmentace nahrávky probíhá na různých úrovních, které jsou dány potřebami ASR a zobrazení dokumentu. Jednotlivé úrovně segmentace (a jimi definované regiony nahrávky) se liší, hranice segmentů jsou však podmnožinou jedné společné sady slotů. Právě v segmentaci nahrávky a vzájemném provázání jednotlivých úloh, které finální segmentaci ovlivňují, spočívá první porovnání obou navržených schémat. Druhou porovnanou kategorií jsou přínosy a nevýhody použití různých konfigurací systému rozpoznání řeči (ASR).

První strukturalizační schéma (schéma s izolovaným rozhodováním – sekce 6.1) řeší většinu dílčích úloh bez ohledu na výstupy ostatních nástrojů, ačkoli jejich výsledky se týkají výše zmíněné společné sady slotů. Toto schéma současně vychází z použití CD-GMM-HMM LVCSR (později označovaného jako *LVCSR-GMM*), který pro dosažení požadované přesnosti přepisu používá adaptaci úseků nahrávky na mluvčího, což klade specifické požadavky na řazení kroků strukturalizace nahrávky.

Druhé navržené schéma (schéma s kumulovaným rozhodováním – sekce 6.2) shromažďuje informace z dílčích subsystémů a provádí rozhodování najednou, s využitím maximální dostupné informační základny. Současně je ve schématu použit CD-DNN-HMM LVCSR (později označovaný jako *LVCSR-DNN*), který dosahuje požadované přesnosti přepisu i bez adaptace, což umožňuje výraznou změnu pořadí modulů strukturalizace a úsporu výpočetního času.



Obrázek 4.1: Úrovně segmentace nahrávky využité při strukturalizaci dokumentu

Obě schémata rozdělují obsah nahrávky do následujících základních kategorií, které musí být reflektovány i anotací testovacích dat:

- promluva (s různým přenosovým pásmem a jazykem promluvy, případně hlukem a hudbou na pozadí),
- ticho (nebo jiný delší neřečový segment),
- delší úsek se sporadickým výskytem řečových událostí (může vzniknout v důsledku přítomnosti hudby, nebo silného hluku na pozadí),
- znělka / gong (krátký neřečový předěl podobný hudbě).

Chceme-li porovnat obě navržená schémata a najít tak nejvhodnější topologii strukturalizačního schématu, musíme shromáždit sadu testovacích nahrávek a stanovit vhodné metriky. Jak naznačuje předcházející výčet základních typů obsahu nahrávky, testovací data musí obsahovat komplexní anotaci řady informací navázanou na časové značky řečových i neřečových událostí. Pro tvorbu referenčních dat je cílem vytvořit maximálně automatizovaný mechanismus a zavést takový formát pro jejich uložení, že jednotlivé experimenty bude možné vyhodnotit pomocí stejného referenčního souboru. Automatizace provázání časových značek s obsahem referenčních anotací má za cíl umožnit vznik rozsáhlých testovacích sad. Na základě vyhodnocení silných a slabých stránek otestovaných strukturalizačních schémat můžeme navrhnout jejich úpravy, respektive úplně nové schéma, které by minimalizovalo nedostatky schémat popsaných v této práci (a díky rozsáhlé testovací sadě jeho přesnost ověřit).

Jak již bylo řečeno v sekci 2, práce popsaná v následujících kapitolách je motivována projektem (*NAKI*), jehož cílem je zpřístupnění publicistických a zpravodajských pořadů z archivu ČRo. Splnění cílů definovaných ve výše zmíněné pasáži lze chápat jako aplikační cíl této práce.

4.2 Shrnutí cílů práce

Cíle této práce lze stručně shrnout v následujících bodech:

- provést rešerši existujících metod a řešení v oblasti zpracování a zpřístupnění velkých multimediálních archivů,
- navrhnout schémata strukturalizace nahrávky (s ohledem na přesnost přepisu, informační obsah a vyhledávání),
- 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,
- analyzovat a popsat vazby uvnitř navržených strukturalizačních schémat a stanovit klíčové požadavky na jednotlivé moduly,
- porovnat výsledky dosažené navrženými schématy,
- připravit navržené postupy a nástroje k reálnému nasazení,
- vyhodnotit poznatky z reálného provozu.

5 Moduly a nástroje vyvinuté pro strukturalizaci dokumentu

V následující kapitole (6) představím navržená (a realizovaná) schémata strukturalizace počítačem zpracovaného dokumentu. Obě představená schémata jsou vystavěna ze společné sady nástrojů (modulů), které vykonávají jednotlivé úkony nezbytné pro zpracování nahrávky a strukturalizaci vytvořeného dokumentu. Představení dílčích modulů v samostatné kapitole nám umožní soustředit se při popisu schémat na jejich celkovou koncepci, bez potřeby popisovat chování jejich komponent. Předběžnou představu o provázanosti jednotlivých nástrojů lze získat z obr. 3.1. Přesnější popis strukturalizačních jednotek a vzájemné vazby mezi nahrávkou, výstupem LVCSR systému a výsledným přepisem jsou diskutovány v následující sekci.

Řada nástrojů integrovaných ve strukturalizačních schématech je výsledkem činnosti kolektivu Laboratoře počítačového zpracování řeči (5.3,5.4.2,5.5), která probíhá již asi 15 let a účastnily se jí téměř dvě desítky pracovníků. U těchto nástrojů poskytuji popis základních principů jejich funkce, detailní informace jsou k dispozici v odkazované literatuře. U nástrojů, na jejichž vývoji jsem se přímo podílel nebo byly vytvořeny výlučně pro potřeby strukturalizace dokumentu, je popis obsáhlejší.

Při popisu jednotlivých nástrojů budu věnovat pozornost vstupním datům, která potřebují pro svou činnost (neboť do značné míry determinují řazení nástrojů ve strukturalizačních schématech). U vstupních a výstupních dat se nebudeme zabývat konkrétní datovou strukturou, ale jejím informačním obsahem. Řídicími parametry nástrojů pak rozumíme nastavení prahů, nejrůznější doplňkové informace a hlavně modely, které nástroje používají. Závěrem popisu každého z nástrojů je shrnutí vstupních/výstupních dat a řídicích parametrů.

5.1 Strukturalizační jednotky a jejich vazby

Proces strukturalizace dokumentu musí vyhledat a zohlednit vazby mezi zvukovou nahrávkou dokumentu, přepisem dokumentu pořízeným LVCSR systémem a výsledným strukturalizovaným dokumentem. Ve všech třech úrovních lze dokument hierarchicky rozdělit na nižší celky, které si však nejsou vzájemně ekvivalentní. Nejprve budou popsány vazby mezi nahrávkou dokumentu (vstupním číslicovým signálem) a výstupem LVCSR systému. Následně budou popsány 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 se kterým jsou lokalizovány události v nahrávce a tím určují i přesnost časování výsledného dokumentu. Jeden frame nemá žádný přímý ekvivalent ve výstupu LVCSR systému ani ve výsledném dokumentu (je příliš krátký). Hierarchicky výše je postaven *segment* nahrávky. Segmentem rozumíme homogenní úsek nahrávky (respektive vstupního signálu), přičemž homogenita může být určena na různé úrovni. Základním kritériem homogenity je dělba segmentů na řečové-neřečové, dále lze rozlišovat jazyk segmentu, charakter přenosového pásma a pohlaví, či identitu mluvčího. Segmentů tedy existuje více kategorií, čím je homogenita více konkretizována, tím jemněji jsou segmenty děleny. Nejvyšším celkem na úrovni vstupního signálu je celá *nahrávka* dokumentu.

Na úrovni výstupu LVCSR systému je nejnižším elementem přepisu *událost*. Rozlišujeme události dvou charakterů. Prvním možným charakterem je řečová událost (mezi které počítáme nejen rozpoznaná slova či fráze, ale i součásti foneticko-akustického inventáře spojené s tvorbou řeči – např. nádech, váhavý zvuk). Druhým charakterem je neřečová událost, do které zahrnujeme detekci všech zbylých položek foneticko-akustického inventáře (např. hluky modelující hudbu, kašel a další). Posloupnost událostí detekovaných LVCSR systémem v konkrétním rozsahu časových značek vstupního signálu lze postavit na úroveň segmentu signálu. Výstup LVCSR systému však sám o sobě segmentován není – celý vstupní signál je rozpoznán jako jeden spojitý segment (celá časová osa nahrávky je pokryta (ne)řečovými událostmi).

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*. Za promluvu považujeme homogenní řečový projev jednoho řečníka (homogenní jak ve smyslu jazyka, tak charakteru přenosového pásma). Promluvu lze proto položit na úroveň segmentu vstupní nahrávky (uvažujeme-li stejnou úroveň homogenity). Na promluvu jsou (díky její deklarované homogenitě) vázány všechny klasifikace nahrávky (jazyk promluvy, identita mluvčího, charakter přenosového pásma). Promluvy jsou zastřešeny přímo výsledným dokumentem (kterému odpovídá přepis nahrávky).

Kontejner pro uložení finálního dokumentu (popsaný v sekci 5.9) pak dostupnou informaci částečně redukuje. Jako nejnižší element chápe *frázi* – nejkratší řečovou událost opatřenou časovými značkami. Fráze tedy nejčastěji odpovídá jednomu slovu, může ale odpovídat i cele číselné entitě, nebo naopak jednotlivým částem jmenné entity. Sada frází tvoří *paragraf* přepisu, který odpovídá promluvě. Proto jsou na paragraf vázány všechny informace o promluvě. *Kapitola* odpovídá celému dokumentu (některé volitelné úrovně dělení mezi kapitolou a paragrafem nejsou při automatickém zpracování dokumentu využity).

Redukce informace zmíněná v předešlém odstavci je zcela dostačující pro indexaci přepisu do databáze – odpovídá chápaní dokumentu jako sady "titulků", jak je tomu u systémů představených v sekci 3.2. Současně nám předchozí popis umožňuje tvrdit, že všechny významné události v nahrávce (změny řečníka a dalších atributů stejně jako přítomnost interpunkce) lze lokalizovat do jedné společné sady časových značek. Tato sada časových značek jsou začátky a konce řečových událostí a budeme je v dalším textu označovat jako *sloty*.

Vzájemný vztah elementů definujících výsledný strukturalizovaný dokument je zachycen na následujícím obrázku (obr. 5.1).

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 5.1: Elementy zapojené do tvorby strukturalizovaného dokumentu

5.2 Modul parametrizace akustického signálu

Úkolem modulu parametrizace je přiřadit signálu (akustické stopě dokumentu nebo jejímu úseku) reprezentaci pomocí zvolené příznakové sady. Vstupní audio signál je nejprve konvertován do formátu standardu akustických modelů – PCM Wave (jednokanálová nahrávka, vzorkovací frekvence 16 kHz, 16 bitů na vzorek). Konverze je prováděna knihovnou FFmpeg[1], která umožňuje konverzi většiny rozšířených audio formátů. Po konverzi je signál rozdělen na jednotlivé framy (používáme framy dlouhé 20 ms, s překryvem 10 ms). Každý frame je parametrizován (v našem případě 39 MFCC příznaků – 13 statických a jejich první a druhá diference). Příznakové vektory jsou normalizovány odečtením střední hodnoty – buď globálně (cepstral mean subtraction – CMS), nebo plovoucím oknem (floating CMS). Délku plovoucího okna volíme 2 sekundy.

[1] https://www.ffmpeg.org

Rozhraní nástroje

vstupní data: číslicový signál – zvuková stopa dokumentu
výstupní data: posloupnost příznakových vektorů (39 MFCC)
řídicí parametry: vzorkovací frekvence nahrávky, délka framu a jeho překryvu, požadovaná parametrizace (39 MFCC), normalizace (CMS nebo FCMS s délkou plovoucího okna)

5.3 Systém pro rozpoznání spojité řeči

Systém pro rozpoznání spojité řeči je klíčovou komponentou zpracování zvukových nahrávek. V této práci používáme rozpoznávač spojité řeči vyvinutý na Ústavu informačních technologií a elektroniky FMMIS TUL [3]. Rozpoznávací systém používá vlastní Viterbiho dekodér a byl optimalizován tak, že může pracovat jak v off-line režimu, tak on-line (úlohy diktátu). Pro práci v on-line režimu byl rozpoznávač doplněn o schopnost aplikovat vážené konečné stavové automaty (WFST) na výstupní text, což umožňuje provádět post-processing textu. V obou režimech je systém schopen práce s velkými slovníky (cca 500.000 položek), což ho řadí do kategorie LVCSR (Large Vocabulary Continuous Speech Recognition).

Systém rozpoznání řeči byl původně vytvořen pro rozpoznávání češtiny. V posledních letech jsou postupně vytvářeny modely i pro rozpoznávání dalších jazyků (slovenština [39, 40], polština [41], chorvatština [42]), ruština [43]). V této práci jsou pro nás důležité dva jazyky – čeština a slovenština. U obou jazyků používáme stejnou fonetickou abecedu [44], která obsahuje 42 fonémů a sadu 6 neřečových událostí – hluků (mezi nejdůležitější patří ticho, nádech, váhavý zvuk a různé typy hluků). Ačkoli jsou akustické modely obou jazyků odlišné, umožňují jejich fonetické inventáře vzájemné namapování fonémů v případě nedostatku dat pro trénování akustických modelů. Fonetický přepis (G2P) je založen na vstupních slovnících, systém umí generovat výslovnost i pomocí strojově trénovaných WFST [45]. Akustické modely jsou trifonové.

Rozpoznávač řeči je v této práci použit 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 ([46]).

5.3.1 LVCSR-GMM

GMM modely používají MFCC parametrizaci (13 příznaků s 1. a 2. diferencí), framy mají délku 20 ms, 10 ms překryv. Na příznakové vektory je aplikována plovoucí normalizace odečtením střední hodnoty (floating cepstral mean subtraction). Modely jsou trénovány nezávisle na pohlaví (smíšený model pro všechny mluvčí).

Volitelným režimem LVCSR-GMM je adaptace na mluvčího (respektive na akustické podmínky nahrávky). 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 [47], odvozenou z Maximum Likelihood Linear Regression [48]. Její podstata spočívá v nalezení transformační matice, která převádí rozšířený příznakový vektor na adaptovaný příznakový vektor (lépe odpovídající akustickým modelům).

5.3.2 LVCSR-DNN

CD-DNN-HMM konfigurace rozpoznávače nahrazuje GMM modely a následné určení věrohodností fyzických stavů (senonů) Viterbiho dekodéru hlubokou neuronovou sítí, která přímo generuje věrohodnosti stavů (po normalizaci a logaritmování). Trénovací data vychází z trénovacích dat GMM modelů – indexy anotovaných senonů jsou použity jako požadované výstupy DNN. Vstupní příznakový vektor sítě je složen z příznakových vektorů klasifikovaných framů (délka 20 ms, překryv 10 ms). Používáme okolí 5 framů před i za zkoumaným framem. Parametrizace je shodná s LVCSR-GMM – 39 MFCC příznaků pro každý frame. V našem případě má DNN 5 skrytých vrstev širokých 1024 neuronů, aktivační funkcí je sigmoida. Jednotlivé neurony výstupní sítě jsou přímo přiřazeny jednotlivým senonům a jejich výstupní hodnota se přepočítává na věrohodnost senonu.

5.3.3 Přehled použitých modelů pro LVCSR systém

V této práci jsou použity různé akustické (AM) a jazykové (LM) modely a jim odpovídající slovníky (VOC) pro systém rozpoznání řeči. 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 model (NB – narrowband). Podle konfigurace akustické části rozpoznávacího systému rozlišujeme GMM a DNN modely. Použité kombinace jsou zobrazeny v následující tabulce 5.1, spolu s množstvím trénovacích dat a velikostí slovníků. V následujících kapitolách je provedeno několik porovnání LVCSR-GMM a LVCSR-DNN. U všech těchto porovnání je AM trénován ze stejné sady trénovacích dat, LM a VOC jsou pro obě konfigurace kompatibilní.

Rozhraní nástroje

vstupní data: audio dokument (nebo jeho segment) reprezentovaný příznakovými vektory (39 MFCC)
výstupní data: posloupnost řečových a neřečových událostí (každá s určenými časovými značkami a ortografickou a fonetickou reprezentací)
řídicí parametry: fonémová sada, slovník, jazykový model, akustický model, případně data pro adaptaci na mluvčího

Tabulka 5.1: Přehled velikosti slovníků, LM, množství trénovacích dat pro AM 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		—

5.4 Segmentace nahrávky: klasifikace řeč–neřeč a diarizace nahrávky

V nahrávkách archivních pořadů se setkáváme s různým charakterem jejich řečových a neřečových úseků. Počínaje plynulou promluvou přes řeč s větším množstvím nespojitostí (nádechů, pauz a váhání) a krátké promluvy oddělené hudebními předěly se dostáváme k úsekům se sporadickým výskytem řečových událostí (obvykle s velkým množstvím hluku na pozadí) až k čistě neřečovým úsekům (hudební produkce, zvuky z průmyslového provozu). Při segmentaci nahrávky je tedy nutné nalézt neřečové segmenty a v řečových úsecích najít body, kde se změnil mluvčí (případně přenosové pásmo či jazyk promluvy).

Cílem segmentace nahrávky je rozdělit nahrávku na homogenní segmenty (rozlišit neřečové segmenty a jednotlivé promluvy). Nalezení promluv (které vyžaduje rozlišení mluvčích) se tak velmi podobá úloze diarizace nahrávky. Diarizací nahrávky rozumíme určení "kdo-kdy-hovořil" čili rozdělení nahrávky na segmenty a označení promluv každého mluvčího jeho identifikátorem. Výsledek diarizace je tedy podobný požadované segmentaci nahrávky (rozdílem může být situace, kdy mluvčí změní jazyk, kterým hovoří). Diarizace nahrávky je poměrně komplikovaný úkol, který popíšeme se značnou mírou zjednodušení (oddíl 5.4.2).

Segmentaci nahrávky lze rozdělit do tří vrstev:

- detekce řečové aktivity,
- nalezení bodů změny v nahrávce,
- určení homogenních segmentů nahrávky.

Pokud segmentaci nahrávky zahájíme rozpoznáním nahrávky LVCSR systémem, můžeme následné hledání bodů změny v nahrávce omezit pouze na začátky a konce rozpoznaných řečových událostí. Takový postup snižuje výpočetní nároky všech potřebných klasifikací. Současně je provedena synchronizace výsledků klasifikátorů s přepisem (obecně nelze zaručit, že časové značky generované nezávislým detektorem změny by se shodovaly s hranicemi následně rozpoznaných řečových událostí).

Úvodním krokem segmentace nahrávek je detekce řečové aktivity v nahrávce a lokalizace segmentů, které lze označit za neřečové. Hranice (časové značky začátků a konců) neřečových segmentů jsou buď hranicí samotné nahrávky, nebo musí tvořit hranici mezi řečovým a neřečovým segmentem. Díky tomuto faktu je možné lokalizovat všechny ostatní změny charakteru nahrávky (body změny) pouze uvnitř úseků označených jako řečové. Do bodů změny počítáme změnu mluvčího (zjištěnou diarizací nahrávky), změnu jazyka a změnu přenosového pásma. Následné shlukování nalezených úseků nahrávky definuje konečné segmenty nahrávky (s ohledem na požadovanou úroveň homogenity segmentů).

5.4.1 Klasifikace řeč–neřeč

Přímo z dostupného výstupu LVCSR systému lze extrahovat regiony s (ne)řečovou aktivitou v nahrávce (časové známky (ne)řečových událostí v nahrávce). Toho využíváme při klasifikaci úseků nahrávky na řečové a neřečové. Samotný modul lze zařadit jak za detekci bodů změny v nahrávce (aplikovat ho na již rozdělenou nahrávku), tak před detekci bodů změny (můžeme segmentačnímu nástroji podvrhnout pozměněný výstup rozpoznávače, ve kterém je v časovém úseku odpovídajícímu neřečovému segmentu rozpoznán hluk).

Klasifikátor řeč–neřeč rozlišuje řečové a neřečové úseky podle podílu řečového a neřečového obsahu mezi začátkem a koncem plovoucího okna (pracujeme nad hranicemi rozpoznaných událostí a a b). Za řečový obsah považujeme řečové události (slova, fráze) a některé neřečové události související s řečí (nádech, váhavý zvuk). Mezi neřečový obsah počítáme všechny zbylé typy hluků a dlouhé úseky ticha.

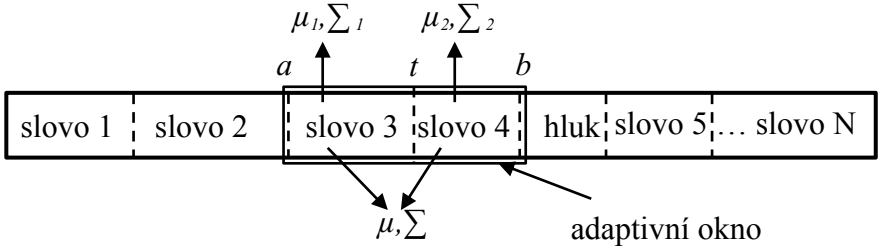
Podíl řečového a neřečového obsahu je počítán v rámci plovoucího okna fixní délky (čas b je první konec rozpoznané události, který má minimální požadovanou vzdálenost od času a). Je-li nalezen úsek nahrávky s dostatečně malým podílem řeči, jsou jeho hranice zpřesněny pomocí druhého (kratšího) plovoucího okna. Délky posuvných oken i hraniční poměry řečového a neřečového obsahu byly určeny na základě experimentů s vývojovými daty [49].

Tento postup samozřejmě 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.

5.4.2 Detekce bodů změny a diarizace nahrávky

Proces diarizace nahrávky se člení do tří vrstev (které jsou obdobné jako pro segmentaci nahrávky):

- detekce řečové aktivity,
- nalezení bodů změny mluvčího,
- shlukování úseků nahrávky.



Obrázek 5.2: Detekce změny mluvčího adaptivním oknem omezeným na hranice událostí detekovaných LVCSR systémem

Detekce řečové aktivity je provedena na dvou úrovních. Nalezení dlouhých neřečových segmentů je provedeno klasifikátorem popsaným v předchozí části. Detekce neřečových událostí uvnitř řečových segmentů využívá výstup LVCSR systému. Nalezení bodů změny v nahrávce využívá body změny mluvčích (určené konečným výstupem diarizačního systému).

Změny mluvčího jsou nalezeny 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 [50, 51]. Podobnost prověřovaných intervalů je vyhodnocena pomocí BIC zavedeného vztahy (5.1) a (5.2). 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 (v našem případě 13 příznaků – nejsou použity 1. a 2. diference) a α představuje penalizační koeficient (v našem případě volíme $\alpha = 1$).

$$BIC = (N_1 + N_2)log(|\Sigma|) - N_1 log(|\Sigma_1|) - N_2 log(|\Sigma_2|) - \alpha P \tag{5.1}$$

$$P = \frac{1}{2}((d + \frac{1}{2}(d(d+1)))log(N_1 + N_2) \tag{5.2}$$

Posledním krokem diarizace je shlukování úseků nahrávky (snaha přiřadit všem promluvám jednoho mluvčího stejné ID). Provádí se následujícím postupem:

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 [24] 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).

Omezíme-li možné body změny mluvčího na hranice řečových událostí v nahrávce (jak naznačuje obr. 5.2), je možné snížit výpočetní nároky a zvýšit přesnost diarizace. V [24] jsou tyto přínosy vyčísleny následovně: snížení výpočetní náročnosti z 0,14 RT na 0,02 RT (RT značí Real-Time faktor – podíl délky provedení výpočtu a délky trvání zpracovaného signálu) a nárůst přesnosti detekce bodů změny (kvantifikovaná pomocí *F-measure*, definované v sekci 7.2), která se změnila z 66,6 % na 77,2 %. Chybovost lokalizace bodů změny (s tolerancí 20 ms) se snížila z 44,9 % na 6,5 %.

Po provedení diarizace nahrávky je možné snadno vyhledat v nahrávce všechny body (sloty), ve kterých dochází ke změně v nahrávce. Jak bylo zmíněno výše, změna zahrnuje charakter segmentu (řeč–neřeč), jazyk promluvy, charakter přenosového pásma a identitu mluvčího (určení jazyka promluvy, identity mluvčího a charakteru přenosového pásma diskutujeme v následující sekci). Požadavky na homogenitu nahrávky se mohou lišit podle toho, který z kroků zpracování nahrávky je strukturalizačním schématem aktuálně prováděn. Například, pokud chceme rozpoznat segmenty za použití správného jazykového a akustického modelu, ID mluvčího je pro segmentaci irelevantní. Proto je pojem segment v této práci použit v kontextu různých úrovní homogenity - různě detailní segmentace nahrávky.

Rozhraní nástroje

vstupní data: přepis nahrávky (časové značky událostí a odlišení řečových od neřečových událostí), audio dokument reprezentovaný příznaky (13 MFCC)
výstupní data: úseky nahrávky s indexy mluvčích nebo označením za neřeč
řídicí parametry: model prostoru pro i-vectorovou reprezentaci mluvčích

5.5 Klasifikace řečových segmentů nahrávky

V následujících odstavcích budou popsány nástroje, které slouží k přesnějšímu určení obsahu jednotlivých řečových segmentů nahrávky. Konkrétně se budeme zabývat určením jazyka promluvy, klasifikací šířky přenosového pásma a určením identity (respektive pohlaví) mluvčího.

5.5.1 Určení jazyka promluvy

Obvyklým postupem pro rozpoznání jazyka promluvy (LID – language identification) je natrénování společného akustického modelu všech jazyků (pokrývá fonémovou sadu všech rozpoznávaných jazyků). Jazykový model je postaven na úrovni n-gramů fonémů a pokrývá tak 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 [52]. Přesnost tohoto přístupu může být ohrožena u krátkých promluv, u kterých se snažíme rozlišit jazyky podobné jak fonémovou sadou, tak slovní zásobou (např. řada českých a slovenských slov jsou homofony).

Proto systém navržený na našem pracovišti [53] výše zmíněný postup rozšiřuje. Pro všechny rozpoznávané jazyky (v našem případě češtinu – CZ a slovenštinu – SK) je připraven společný akustický model (stejně jako u předchozí metody) 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 (nastavení penalizací preferuje setrvání v aktuálním jazyce). Aby nedošlo ke zvýhodnění některého jazyka, jsou slovní zásoby limitovány (všechny na stejné množství nejčastějších slov). Jednotlivé slovníkové položky na sebe vážou informaci o tom, kterému jazyku přísluší, případně že se foneticky stejná položka vyskytuje v obou jazycích. Takto připravené AM a LM mají jednu zásadní výhodu – nehodnotí pouze akustickou informaci v určitém (akustickém) kontextu, ale kontext je delší a zapojuje do modelu vyšší celky (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. 5.3). 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 5.1 (klíč CZ+SK).

V [53] je experimentálně vyčíslen vliv velikosti použitých slovníků na přesnost LID. K vyhodnocení je použito 1.000 českých a 1.000 slovenských úryvků řeči (228 minut / 31.214 slov) s minimální délkou 6 slov. Přesnost diarizace v závislosti na velikosti slovníku ukazuje, že pro slovníky větší než 20.000 položek již systém pracuje s chybou menší než 1,6 %. Slovníky použité v této práci operují s 50.000 slov pro každý jazyk a LID modul pracuje s přesností cca 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 5.3: Ilustrace určení jazyka promluvy–čeština (CZ), slovenština (SK), slovo společné pro slovníky obou jazyků (COM)

Rozhraní nástroje

vstupní data: audio dokument (nebo jeho segment) reprezentovaný příznakovými vektory (39 MFCC); případně informace o rozdělení nahrávky na dílčí promluvy
výstupní data: řečové a neřečové události v nahrávce (s ortografickou i fonetickou podobou, časovými značkami a přiřazeným jazykem ke každé řečové události); segmenty nahrávky s přiřazeným jazykem promluvy
řídicí parametry: dvojjazyčné AM a LM pro systém rozpoznání řeči, slovníky obohacené o přiřazení jazyka (případně statutu sdíleného slova)

5.5.2 Klasifikace šířky přenosového pásma

Jednou ze základních klasifikací úseků nahrávky je identifikace šířky přenosového pásma. Cílem je určit, jestli nahrávací řetězec použitý při vzniku daného úseku nahrávky zajistil dostatečnou šířku přenosového pásma (min. vzorkovací frekvence 16 kHz – *WB*), nebo jestli některá jeho část tento předpoklad nedodržela (např. přenos signálu po telefonní lince, starší nahrávací zařízení, nevhodný mikrofon – *NB*). Vstupní informací pro nástroj jsou hranice segmentů, které mají být klasifikovány a řečová aktivita (určená z přepisu nahrávky).

Postupně byly navrženy dva nástroje pro klasifikaci šířky přenosového pásma. První z nich byl založený na podílu energie v jednotlivých částech spektrogramu nahrávky. Druhý pak odlišuje obě kategorie pomocí GMM modelů.

Nástroj vycházející ze spektrogramu nahrávky využívá předpoklad, že většina nahrávek s úzkým přenosovým pásmem vznikla z telefonního přenosu, takže je omezena vzorkovací frekvencí 8 kHz. Porovnává proto energii signálu v pásmu 0-4 kHz s celkovou energií v pásmu 0-8 kHz, jak naznačuje vztah (5.3). Překročí-li podíl energie (Band Energy Ratio - BER) stanovenou mez, je úsek nahrávky označen za úzkopásmový.

$$BER = \frac{\sum_{f=0Hz}^{4kHz} E(f)}{\sum_{f=0Hz}^{8kHz} E(f)} \tag{5.3}$$

Klasifikátor přenosového pásma nasazený ve strukturalizačních schématech využívá k rozlišení jednotlivých kategorií přenosového pásma GMM modely (MFCC parametrizace shodná se systémem rozpoznání řeči). Při tvorbě testovacích a trénovacích dat byl využit nástroj popsaný v předchozím odstavci. U GMM modelů předpokládáme vyšší robustnost vůči různým variacím nahrávacího řetězce (např. použití přenosných magnetofonů, diktafonů a podobných zařízení). Současně dává prostor k případnému natrénování modelů schopných určit specifická nahrávací zařízení (respektive celé nahrávací a ukládací řetězce).

Rozhraní nástroje

vstupní data: úseky nahrávky reprezentované příznakovými vektory (39 MFCC)
výstupní data: šířka přenosového pásma úseku nahrávky
řídicí parametry: GMM modely plného a úzkého přenosového pásma

5.5.3 Určení pohlaví mluvčího

Pro určení pohlaví mluvčího jsou využity GMM modely podobné těm, které používá VAD modul systému rozpoznání řeči (39 MFCC parametrů). Lze je zjednodušeně nazvat muž (M), žena (F) a obecný hluk (X), ticho není třeba nijak dále klasifikovat. Obecný hluk sice nepatří mezi pohlaví, třetí model zahrnující směs neřečových událostí a písniček však dává klasifikátoru možnost nepřiřadit nesprávné pohlaví neřečovému úseku nahrávky. K detekci řečové aktivity se opět využívá výstup systému rozpoznání řeči.

Rozhraní nástroje

vstupní data: úsek nahrávky popsaný příznakovým vektorem (39 MFCC), přepis nahrávky pro VAD
výstupní data: přiřazené pohlaví mluvčího (M/F/X)
řídicí parametry: GMM modely pro muže, ženu a obecný neřečový úsek nahrávky

5.5.4 Identifikace mluvčího

Úkolem nástroje pro identifikaci mluvčího je určit, který mluvčí (z množiny modelovaných mluvčích) pronesl zkoumanou promluvu. Úloha verifikace mluvčího pak ověřuje, jestli rozpoznaný mluvčí je skutečně hledaným řečníkem, nebo jestli byl pouze nejpodobnějším mluvčím ze sady dostupných modelů.

Modul pro identifikaci mluvčích je založen 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" [54], 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 (5.4), kde x_1 a x_2 značí referenční a zkoumaný i-vector.

$$CDS = \frac{{x_1}' x_2}{\|x_1\| \, \|x_2\|} \tag{5.4}$$

Pro lepší funkci identifikace mluvčích (a pro úsporu výpočetního výkonu) je vhodné limitovat vstupní sadu modelů mluvčích pouze na relevantní podmnožinu všech známých mluvčích (mluvčích, pro které máme k dispozici modely). K tomu lze využít informace zpřístupněné během strukturalizace dokumentu. Prvním kritériem je pohlaví mluvčího. To je v našem případě určeno pomocí GMM modelů (stejně jako šířka přenosového pásma). Druhým kritériem je šířka přenosového pásma zkoumaného úseku nahrávky (stejný mluvčí "zní " jinak v telefonu a jinak přes standardní nahrávací řetězec, což se nutně projeví v příznakovém vektoru popisujícím jeho promluvu a modelu potřebném pro jeho rozpoznání). Třetím kritériem je jazyk promluvy, kdy předpokládáme, že databáze mluvčích obsahuje informaci o všech jazycích, kterými je daný mluvčí schopen hovořit a vyloučit ho z promluv

v jazyce, kterým nehovoří. Čtvrtým kritériem jsou doplňkové informace o pořadu spolu s informacemi obsaženými v databázi mluvčích. Doplňkové informace umožňují například přidat do sady modelů všechny moderátory konkrétního publicistického pořadu, stejně jako omezit sadu mluvčích na základě doby, kdy byla nahrávka pořízena a výskytu mluvčích v čase. Příkladem budiž vyloučení politických disidentů v době vlády komunistického režimu.

Z výše popsané metodiky výběru modelů mluvčích vyplývá, že pro mluvčí modelované v našem systému mohou 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 [55, 56].

Verifikace mluvčích, která je obvyklou součástí systémů identifikace mluvčího, je v našem případě zjednodušena. Skóre získané navrženým mluvčím je porovnáno s bezpečnostním prahem. Pokud je dosažené skóre příliš malé, použije se místo jména mluvčího pohlaví. Výjimkou jsou situace, kdy je zkoumaná promluva krátká a okolní úseky mají bezpečně rozpoznaného stejného mluvčího. V takovém případě je bezpečnostní práh změněn na základě kontextu okolních úseků nahrávky.

Rozhraní nástroje

vstupní data: úsek nahrávky, o níž předpokládáme, že náleží jednomu mluvčímu (promluva ve formě audio stopy), přepis úseku pro určení řečové aktivity
výstupní data: nejpravděpodobnější mluvčí z dostupné sady modelů, skóre odpovídající věrohodnosti určení mluvčího
řídicí parametry: model "total variability space", modely známých mluvčích

5.6 Doplňková parametrizace

V následujících odstavcích budou popsány pomocné nástroje pro určení fundamentální frekvence řeči (melodické linky) a energie signálu. Motivací pro výpočet těchto parametrizací nahrávky je extrakce některých příznaků, které jsou složkami větné prozodie. Tu se snažíme využít při stanovení konečné segmentace nahrávky i při doplnění interpunkce do rozpoznaného textu. I při doplňkové parametrizaci je použit synchronizační mechanismus vycházející z výstupu rozpoznávače, jak je popsán v části 5.4. Příznaky popisující prozodii jsou extrahovány na úrovni framů přepisu, jejich trendy však zkoumáme na úrovni řečových událostí (slov).

5.6.1 Krátkodobá energie signálu

První doplňkovou parametrizaci představuje krátkodobá energie signálu. Výpočet je proveden 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). Každému slovu je přiřazena průměrná hodnota energie $\bar{E}$ a normovaná diference energie E_{nd} (5.5), která vyjadřuje míru "kolísání" krátkodobé energie v průběhu slova.

$$E_{nd} = \frac{max(E) - min(E)}{\bar{E}} \tag{5.5}$$

5.6.2 Fundamentální frekvence řeči

Druhou doplňkovou parametrizací je detekce fundamentální frekvence řeči ($F0$) jakožto klíčový popis melodie řeči. Pro její určení existuje řada metod, které vycházejí z několika společných předpokladů. Prvním je kvazi-stacionarita řečového signálu – parametry řeči lze považovat za stabilní v rámci 10–30 ms dlouhých oken, což se týká i buzení zvukového traktu. 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 existuje několik zavedených metod [57, 58]. První skupina metod využívá autokorelační funkci řečového signálu, druhá analyzuje signál v kepstrální oblasti a třetí pracuje se spektrogramem nahrávky. V následujících odstavcích jsou tyto metody stručně popsány, čímž definujeme východiska pro námi navrženou (a použitou) metodu.

Autokorelační funkce a Metoda centrálního klipování

Určení fundamentální frekvence ($F0$) znělého úseku řeči pomocí autokorelační funkce patří mezi nejznámější a výpočetně nejméně náročné postupy. Vzdálenost dvou lokálních maxim jednostranné autokorelační funkce odpovídá posunu, se kterým se signál "opakuje" a určuje tedy periodu řečového signálu – převrácenou hodnotu fundamentální frekvence. Posun lze přepočíst na fundamentální frekvenci dle vztahu (5.6), kde F_s značí vzorkovací frekvenci a k značí vzdálenost lokálních maxim (posun maxim o k vzorků). Posledním krokem metody je vyhlazení výsledků plovoucím průměrovacím filtrem.

$$F0 = \frac{F_s}{k} \tag{5.6}$$

Metoda centrálního klipování se od předchozí odlišuje pouze předzpracováním vstupního signálu. V rámci každého zpracovaného framu (normalizovaného signálu X_i) určíme práh a převedeme signál podle následujícího vztahu (5.7). Redukční faktor r se obvykle volí $r = 0{,}8$.

Jak lze odvodit ze vztahu (5.6), frekvenční rozlišení obou těchto metod je silně nelineární. Při vzorkovací frekvenci $F_s = 16$ kHz je rozdíl v délce periody o jeden vzorek při spodní mezi rozsahu roven rozdílu frekvence 0,23 Hz, zatímco při horní mezi rozsahu 9,76 Hz. Tato nelinearita je podstatně výraznější než v případě lidského ucha (jehož frekvenční rozlišení se odhaduje na cca 30 frekvencí na oktávu).

$$\begin{aligned} thr &= r \cdot \max(X_i) \\ X_i[j] &= \begin{cases} X_i[j] = 1 \ ... \ X_i[j] > thr \\ X_i[j] = -1 \ ... \ X_i[j] < -thr \\ X_i[j] = 0 \ ... \ jinak \end{cases} \end{aligned} \tag{5.7}$$

Kepstrální metoda

Kepstrální metoda vychází z předpokladu, že v kepstrální oblasti lze odlišit složky reprezentující vlastnosti řečového traktu od části kepstra, které popisuje excitační informaci. Na úseku nahrávky $sig[n]$ se spočítá reálné kepstrum dle vztahu (5.8). V kepstru je nalezeno maximum v rozsahu indexů, který odpovídá frekvencím 60-400 Hz. Přepočet indexu spektra na fundamentální frekvenci se řídí vztahem (5.6).

$$c_R[n] = Re\{IFFT(\ln(|FFT(sig[n])|)\} \tag{5.8}$$

STFT + image processing

Další metoda vychází ze zpracování řečového signálu pomocí krátkodobé Fourierovy transformace (STFT). Nejprve se vypočte spektrogram řečového signálu a zpracovaná oblast se frekvenčně omezí. Spektrogram se prahuje a ze signálu se ponechá pouze první harmonická složka. Na obrázek se následně aplikuje dvourozměrný mediánový filtr a maxima jednotlivých sloupců jsou prohlášena za fundamentální frekvenci.

Navržená metoda: STFT + dynamický dekodér

Porovnáme-li výše zmíněné metody, postupy využívající autokorelaci signálu stejně jako kepstrální metoda jsou výpočetně nejméně náročné, jsou však náchylnější k šumům na pozadí řeči. Metoda vycházející ze STFT je naopak odolnější vůči šumům, její výpočetní nároky jsou však vyšší. Současně se v [57] ukazuje, že spektrální metoda má nižší přesnost. Tento rozdíl lze však částečně odvodit z poměrně velkého kroku frekvenčního spektra (frekvenční osa je rozdělena lineárně). Tím pak, i v případě, že algoritmus vybere nejbližší frekvenci k frekvenci referenční, roste "chybovost" algoritmu. Výpočetní náročnost pak do určité míry souvisí se zpracováním dat jako obrazu – aplikace dvourozměrných filtrů.

Vzhledem k tomu, že se ve své práci zaměřuji na zpracování reálných dat, u kterých je nutné předpokládat přítomnost většího množství šumu na pozadí řeči, rozhodl jsem se modifikovat právě metodu vycházející z STFT (a v ní eliminovat časově náročné operace). Nepřesnost metody lze pak do určité míry kompenzovat zvýšením rozlišení spektrogramu. Toho lze dosáhnout buď doplněním signálu nulami, interpolací ve spektru, nebo metodami jako Zoom FFT. V námi navrženém algoritmu aplikuji právě doplnění signálu nulami.

Námi navržená (a aplikovaná) metoda určení F0 sestává z následujících kroků:

1. segmentace audio-nahrávky
2. výpočet STFT nad segmenty odpovídajícími slovům
3. výběr lokálních maxim magnitudového spektra
4. sestavení struktury pro hledání melodické linky a její nalezení
5. detekce "chyby oktávy" porovnáním výsledků sousedních slov

Výpočet STFT provádíme na signálu vzorkovaném F_s=16 kHz, délku oken volíme 20 ms, překryv 10 ms a signál doplníme nulami na délku N=4096. Ze spektra se využije rozsah indexů odpovídající frekvencím 60–600 Hz, pro něž určíme magnitudu spektra. V magnitudovém spektru vybereme 5 lokálních maxim, z nichž obvykle jedna ze tří nejvýraznějších hodnot představuje fundamentální frekvenci.

Výběrem omezeného počtu lokálních maxim magnitudového spektra provedeme krok velmi podobný prahování spektrogramu, popsaný v metodě *STFT + image processing*. Z každého okna signálu nás tak zajímá pouze několik bodů spektrogramu – konkrétně index jejich frekvence a velikost magnitudy, s jakou byly detekovány. Jednotlivá maxima jsou od největšího ohodnocena 10, 9, 7, 5, 5 body pro následné skórování intonační linky. Samotné nalezení melodické linky slova je provedeno postupem vycházejícím z principů dynamického programování, podobným algoritmu dynamického borcení času (DTW). Liší se v tom, že místo celé matice zarovnávaných sekvencí vstupuje do našeho algoritmu pouze sada "průchozích bodů" pro každé okno signálu.

Algoritmus pro hledání nejlépe vyhovující melodické linky nejprve zprůměruje první tři a poslední tři okna signálu daného slova. Každý bod z tohoto průměru představuje potenciální výchozí/koncový bod melodické linky. V každém kroku mezi dvěma okny vypočítáme, jaká vede nejlepší (nejméně penalizovaná) cesta z předcházejícího okna do okna vyhodnocovaného a pamatujeme si, kudy taková cesta vede. Do každého bodu se můžeme z předchozího dostat 1) přímo z kteréhokoli bodu z minulého okna (s penalizací odpovídající změně frekvence mezi dvěma body); 2) maximálně tři po sobě jdoucí okna může pokračovat frekvence (bod) z předcházejícího okna. Penalizace je navržena tak, aby skoky o více než půl oktávy byly výrazně znevýhodněny. Pokaždé, když cesta projde libovolným bodem, je jeho původní ohodnocení (závislé na pořadí maxim magnitudového spektra) přičteno ke skóre cesty a tak propaguje intonační linky procházející přes výraznější složky spektra framu. Ve chvíli kdy algoritmus dosáhne koncových bodů, vyhodnotí se nejlepší (nejméně penalizovaná) cesta a rekonstruuje se zpětným průchodem ohodnocenou cestou. Tato cesta se uloží jako melodická linka daného slova.

Jelikož slova jsou vyhodnocována nezávisle na svém okolí (dalších slovech přepisu), může se stát, že pro ojedinělá slova je detekována nesprávná intonační linka obsahující dvojnásobné frekvence (o oktávu vyšší). Tuto chybu je možné detekovat pomocí porovnání s okolními slovy a tyto chyby jsou snadno korigovány. Samotná oprava je podmíněna fundamentální frekvencí detekovanou u sousedních slov a současnou existencí o oktávu posunuté alternativní intonační linky.

Rozhraní nástroje

vstupní data: úsek nahrávky (akustická data), časové značky řečových a neřečových událostí
výstupní data: energie signálu a melodická linka navázané na události přepisu
řídicí parametry: nastavení penalizace dekodéru melodické linky

5.7 Dodatečné formátování textu

Primárním účelem dodatečného formátování (post-processingu) textu je zlepšení čitelnosti rozpoznaného textu. Schéma post-processingu, se kterým pracujeme, navrhl doktor Žďánský pro formátování výstupu on-line rozpoznávače spojité řeči. Mnou provedené úpravy se týkaly zejména dvou vrstev schématu – zpracování číslovek a doplnění čárek do přepisu. Sekundárním využitím výstupu post-processingu je zjištění, které rozpoznané řečové události (slova) tvoří homogenní celky čili které sloty v přepisu mohou být vyřazeny z procesu segmentace. Příkladem jsou číslovky, fráze jako "společnost s ručením omezeným" či sekvence titulů a vlastních jmen.

Jelikož lze formátování textu rozdělit do určitých podskupin (zvětšování písmen, tituly, čárky, zkratky atd.), je žádoucí, aby jednotlivé podskupiny mohly být nezávisle přidávány/odebírány ze schématu. Dále je žádoucí, aby formulace každého pravidla zachovávala počet událostí (slov) vstupujících a vystupujících z vrstvy, což umožní udržet synchronizaci mezi vstupním přepisem a jeho formátovanou podobou, tím i vazbu mezi nahrávkou a událostmi formátovaného přepisu. Mohou však být odebírány bílé znaky, takže počet slov se opticky sníží. Pro zpracování je definováno povinné pořadí vrstev (viz tab. 5.2). Například, chceme-li detekovat titul před vlastním jménem, můžeme zaručit, že vlastní jména začnou velkým písmenem. Vzhledem k tomu, že pro aplikaci post-processingu používáme vážené konečné stavové automaty (WFST), má modularita několik dalších výhod. Zaprvé usnadňuje změny jednotlivých vrstev a jejich testování, zadruhé je aplikace "série menších automatů" méně výpočetně náročná než kompozice pravidel do jednoho automatu.

Odstranění hluků je první (povinná) vrstva, která odstraní z textového přepisu ortografickou (psanou) reprezentaci neřečových událostí (nádechy, mlaskání, kliknutí, váhající zvuk, hluky). Pokud bychom tuto vrstvu nepoužili, většinu zbylého řetězce by narušila přítomnost nežádoucích tagů.

Vrstva zpracování **číslovek** detekuje slova, která dohromady tvoří čísla, a nahrazuje je číslovkami. Formátování respektuje kontext čísel (finanční částky, fyzikální jednotky, telefonní čísla). **Řadové číslovky** jsou nasazeny experimentálně – v řadě situací ponechá člověk podobu slovní (např. "První argument obhajoby…"). Číslovka by narušila odlišení začátku věty, vrstvu proto omezujeme na číslovky v datech.

Tabulka 5.2: 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)

Velká písmena jsou v některých případech určena pouze na základě kontextu. Příkladem budiž "ústí řeky Labe" versus "Ústí nad Labem". Některá slova (převážně vlastní jména) mají velké písmeno uloženo přímo ve slovníku.

Zkratky jsou poměrně přímočarou vrstvou. Při jejich definici je potřeba pokrýt všechny tvary slov a dodržet výše zmíněný požadavek, aby se jedno slovo zkratky mapovalo na jeden symbol zkratky (případně na prázdný text). Z výše zmíněného "Ústí nad Labem" pak získáme "Ústí n.L.".

Tituly jsou podobným případem jako zkratky. Kromě titulů formátuje tato vrstva také vojenské a akademické hodnosti. Některá pravidla jsou komplikovanější a navazují na předchozí vrstvy (např. požadavek, aby za titulem následovalo vlastní jméno). Důvodem jsou fráze typu "paní inženýrko".

Speciální symboly zahrnují §, ± a formátování dvojteček, čárek, středníků.

Doplnění čárkové interpunkce je realizováno spolu s ostatními vrstvami post-processingu (v souladu s řazením vrstev). Doplnění zbylé interpunkce je ve dvou navržených schématech realizováno odlišným způsobem (viz sekce 5.8).

Další vrstvou pravidel jsou **oborově definovaná pravidla**. Stejně jako se pro ASR užívá vhodný jazykový model a slovník odpovídající rozpoznávané doméně, můžeme i post-processing obohatit o specifická pravidla vázaná na doménu. Oborově definované úpravy jsou vhodné pro domény jako justice a medicína.

V poslední vrstvě post-processingu se aplikují opravy specifických "chyb", kdy například víme, že příjmení moderátora je shodné s běžným podstatným jménem (např. stodola/Stodola, chudoba/Chudoba), a můžeme formulovat konkrétní opravy.

Rozhraní nástroje

vstupní data: ortografická forma přepisu úseku nahrávky
výstupní data: přepis nahrávky s formátovanou ortografickou formou
řídicí parametry: pravidla pro WFST automaty, jazyk úseku nahrávky

5.8 Doplnění interpunkce

Doplnění interpunkce (a velkých písmen) do přepisu je jeden z nejdůležitějších kroků pro zvýšení čitelnosti přepisu. Již při délce věty 15–20 slov je čitelnost textu výrazně snížena a je třeba ho odpovídajícím způsobem opticky dělit. V této práci jsou navržena dvě interpunkční schémata. Obě schémata (5.8.2,5.8.3) doplňují do přepisu tečky a čárky, k čemuž různým způsobem využívají informaci o segmentaci nahrávky, prozodii promluvy, další informace dostupné v daném stádiu tvorby přepisu a různé zdroje apriorní informace. Prozodií rozumíme celou řadu faktorů, mezi něž patří intonace (technicky reprezentovaná fundamentální frekvencí řeči a energií signálu), větný důraz, tempo řeči, pauzy, dokonce i pořadí slov [59].

Doplnění obou druhů znamének je shrnuto v rámci jedné kapitoly, protože interpunkce může být umístěna pouze do omezeného počtu pozic v nahrávce – slotů. Jak jsem již zmínil dříve, sloty mohou být lokalizovány pouze za řečovými událostmi

v nahrávce. Jedinou výjimku tvoří možnost umístit čárku dovnitř kolokace, tj. řečová událost generována systémem rozpoznání řeči může být složena z více slov. Často se jedná o různé předložkové vazby, které mohou generovat čárku, pravděpodobnost, že dovnitř kolokace je nutné umístit tečku je však téměř nulová.

Z principu jsou možné dva přístupy k doplnění interpunkce. První přístup můžeme nazvat lingvistický. Lingvistické metody vychází ze zpracování textového obsahu a jsou optimální pro zpracování psané (spisovné) řeči. Spisovný jazyk umožňuje zapojení mnoha úrovní lingvistické analýzy (např. Ajka[2] nebo Morče[3]), včetně určení role jednotlivých slov ve větě. Nahrávky mluvené řeči se svým charakterem blíží spíše přirozenému jazyku a výkon řady lingvistických nástrojů může být negativně ovlivněn nepřesnostmi přepisu. V takovém případě může být výhodnější druhý přístup, který budeme nazývat statistický.

Než začneme popisovat konkrétní postupy doplnění interpunkce, je zapotřebí zmínit pojem *pauzové interpunkce* a s ní související teoretické předpoklady. Mluvčí (kteří disponují různým rozsahem řečnických dovedností) vědomě využívají některé z výše zmíněných prvků prozodie, aby jejich promluvy byly srozumitelné a aby zdůraznili klíčové informace. Nejčastější je využití pauz v řeči (nádechů a ticha) k naznačení hranic větných celků. Různé jazyky používají různou míru práce s melodií řeči. K naší škodě se středoevropské slovanské jazyky vyznačují spíše nízkou mírou využití melodie. Mluvčí ale obvykle zdůrazňují konce věty určitou změnou melodie, která odpovídá použitému interpunkčnímu znaménku (./?/!). Co je pro všechny jazyky společné je fyziologicky podmíněný pokles fundamentální frekvence řeči v průběhu promluvy (se zmenšujícím se objemem vzduchu v plicích klesá mírně i tlak, který má přímý vliv na buzení hlasivek). Po nádechu v místě interpunkce pak opět promluva začíná na vyšší fundamentální frekvenci.

Výjimku z výše zmíněných předpokladů tvoří některé specifické skupiny mluvčích. V našich experimentech zaměřených na porovnání automatického doplnění interpunkce s lidmi vnímanou interpunkcí [49] jsme, mimo jiné, identifikovali následující skupiny mluvčích: 1) *hlasatelé* (školení mluvčí s vysokou kvalitou promluvy a velkou shodou anotátorů na pozicích interpunkce), 2) *běžní hosté* (s nižší shodou anotátorů, ale platí u nich předchozí předpoklady) a 3) *politici* (ačkoliv vyškoleni, ve snaze nepustit nikoho ke slovu často přerušují promluvy v naprosto nevhodných slotech a narušují tak předpoklady využití pauzové interpunkce).

5.8.1 Doplnění čárkové interpunkce založené na textu přepisu

Navržený mechanismus doplnění čárkové interpunkce vychází ze statistické analýzy jazykových korpusů. Prvním důvodem jsou obrovské časové nároky na vývoj lingvistických analyzátorů. Druhým důvodem je potřeba jazykové přenositelnosti navrženého řešení. Výraznou komplikací použití lingvistických nástrojů je zpracování historických textů (jejich odlišná slovní zásoba, případně větná stavba).

[2] http://nlp.fi.muni.cz/projekty/ajka
[3] http://ufal.mff.cuni.cz/morce

Ačkoli to není z jazykovědného hlediska zcela korektní, lze čárky rozdělit do dvou kategorií podle důvodu, proč je přepisovatel použije. První kategorii tvoří čárky oddělující dva větné celky (typicky jeden podřazený druhému). Takové spojení se často vyskytuje spolu se spojkou (nebo spojující frází), kterou lze statisticky vysledovat (pro modelování jejich výskytu byl použit n-gramový jazykový model). Druhou kategorii tvoří čárky, které neoddělují větné celky, ale jednotlivé větné členy (např. výčet předmětů nebo podmětů ve větě). Syntaktická analýza v takových případech vede ke zpřesnění doplněné interpunkce [60]. Pro určení rolí jednotlivých slov ve větě je již třeba morfologický analyzátor, který obvykle vyžaduje pro svou funkci určené hranice větných celků (pracuje nad jednotlivými větami).

Pro nalezení statistické závislosti mezi konkrétními frázemi a přítomností interpunkce byly shromážděny trénovací a testovací korpusy. S ohledem na určení systému je 80 % českých korpusů tvořeno přepisy zpravodajských pořadů a texty zpráv. Zbylých 20 % obsahuje knižní data (kvůli výskytu archaičtějších obratů, přímých řečí apod.). Český trénovací korpus obsahuje 15.896.044 slov v 1.372.970 větách, testovací korpus má rozsah 6.645.542 slov v 574.045 větách. Slovenský korpus byl složen výhradně z textů zpráv (stažených z internetu) – 330 MB textů. Je o něco větší než korpus češtiny (180 MB), ale méně rozmanitý.

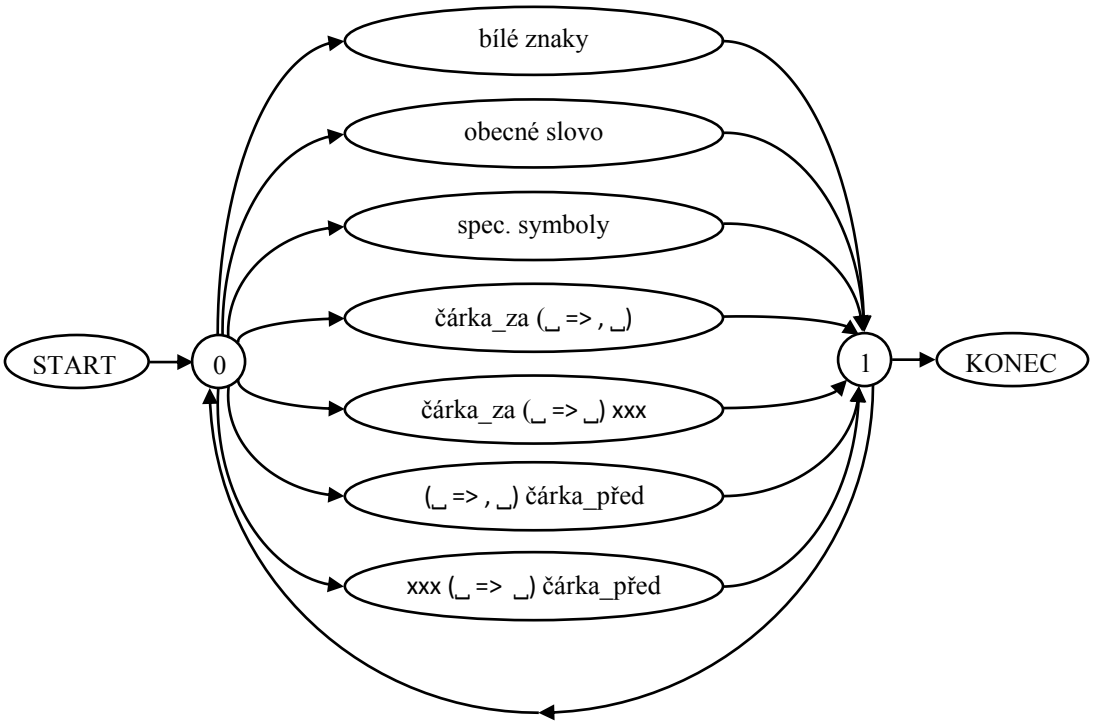
S ohledem na skutečnost, že nás z celého jazyka zajímají pouze sekvence indikující interpunkci, nepřistoupili jsme k modelování n-gramového popisu celého korpusu. Místo toho jsme formulovali selektivní pravidla proměnlivé délky, snadno implementovatelná pomocí WFST. Pravidla byla nalezena následujícím postupem:

V prvním kroku určování čárkovacích pravidel byla z korpusu vybrána slova a dvojice slov, před kterými a za kterými byla pozorována čárka. Ve druhém kroku byly nalezeny všechny výskyty sekvencí, které byly pozorovány spolu s čárkou a bylo zjištěno také jejich okolí (jedno další slovo "směrem k čárce"). To nám umožnilo vyhodnotit pro každou slovní sekvenci (jedno/dvouslovnou), v kolika případech se na zkoumané pozici nachází čárka, nenachází čárka, případně začátek/konec věty. Začátky a konce vět jsou důležité, protože fráze může ukazovat na hranici větného celku, i když daný výskyt fráze není spojen s výskytem čárky.

Pro každou frázi bylo vypočteno, s jakou pravděpodobností generuje čárku. Na základě této pravděpodobnosti se fráze rozdělily do tří skupin:

1. slova/fráze generující čárku pouze ojediněle (ty nejsou centrem našeho zájmu);
2. slova/fráze, u kterých je častější výskyt čárek (cca 30 % případů a více) a
3. slova/fráze s vysokým výskytem čárek (60 % a více).

1. skupina byla rovnou zamítnuta. 2. skupina byla podrobena důkladné analýze, lze-li konkretizovat frázi pomocí jejího okolí a zamezit tak falešným detekcím. Omezení znamená aplikaci čárky, pokud se v okolí nenachází zakázané slovo (nalezené opět analýzou korpusu). Hranice pro přijetí pravidla byla stanovena 65 %. Stejným způsobem byla konkretizována i pravidla z 3. skupiny. Jejich přijetí bylo jisté již od začátku, pouze se ověřila možnost redukce falešných detekcí.



Obrázek 5.4: Struktura WFST automatu pro doplnění čárek do přepisu

Posledním krokem pak byla kontrola, nedochází-li k překryvu pravidel, a ruční doplnění některých variant (např. bylo-li přijato pravidlo ”která”, ”které”, ”kterou”, ”kterých”, bylo přidáno i ”kterého”). Některá příliš konkrétní pravidla byla zobecněna – např. ”ale”, protože anotátoři občas nenapsali čárku na pozici, kde by byla vhodná (nebo alespoň možná) a pravidla by byla zbytečně restriktivní.

Výše popsaným způsobem bylo pro češtinu určeno 1.243 jednoslovných a dvouslovných pravidel, která před sebou generují čárku, a 1.883 prodloužených verzí těchto pravidel (rozšířených o jedno slovo před pravidlem), 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).

Pravidla pro vrstvu generování čárkové interpunkce říkají, že rozpoznaný text může projít automatem buď po jednotlivých slovech a bílých znacích s nějakou váhou (penalizací), nebo může kdykoli projít podsítí definující čárkovací pravidla (s menší penalizací). Nejméně penalizovaná cesta je pak nejlepší – využívá maximum pravidel. Díky tomuto přístupu stačí definovat jak pravidla generující čárku, tak pravidla, která umístění čárky zabraňují (jsou delší) a automat sám zvolí optimální rozmístění čárek. Pravidla jsou graficky znázorněna na obr. 5.4.

V sekci 7.5 jsou výsledky našeho systému pro doplnění čárkové interpunkce [49] porovnány s českým systémem SET [61], který vychází z lingvistické analýzy textů, a se slovenským nástrojem, který používá statistický přístup [62].

Rozhraní nástroje

vstupní data: ortografická forma přepisu úseku nahrávky
výstupní data: ortografická forma přepisu úseku nahrávky s doplněnými čárkami
řídicí parametry: pravidla ve formě WFST automatu, jazyk přepisu

5.8.2 Interpunkční schéma A

Prvním krokem interpunkčního schématu A (které bude v dalším textu značeno *IS_A*) je doplnění čárek do přepisu v rámci provedení post-processingu (viz předcházející oddíl). Ten je aplikován na konečný textový přepis (se znalostí jazyka promluvy).

Nezávisle na čárkovacím schématu jsou provedeny všechny kroky vedoucí ke konečné segmentaci nahrávky (nalezení bodů změny, detekce neřečových úseků, identifikace mluvčích, heuristické opravy segmentace…). Každý konec promluvy mluvčího si pak přirozeně vynucuje ukončení promluvy (odstavce) tečkou. Tím se nahrávka rozdělí na regiony oddělené již doplněnou interpunkcí (čárkami a tečkami na koncích promluv) a naším cílem je optimalizovat čitelnost těchto regionů. K doplnění teček do těchto regionů jsou použity čtyři zdroje informací:

1. trend fundamentální frekvence řeči
2. neřečové události v nahrávce
3. seznam slotů podezřelých, že jsou body změny v nahrávce
4. seznamy slov, v jejichž okolí nemá být tečka umísťována (pro které budeme v dalším textu používat termín blacklist)

Blacklisty vycházejí ze statistického rozboru korpusů a definují 119 slov, za nimiž nemůže být umístěna tečka (např. ”od”, ”pro”, ”jakými”) a 316 slov, před nimiž nemůže být umístěna tečka (např. ”zhruba”, ”byste”). V přepisu jsou nejprve zakázány sloty odpovídající pravidlům v blacklistech a potom jsou postupně aplikovány dostupné zdroje informací.

Využití fundamentální frekvence řeči ($F0$) sleduje trend $F0$ dvě řečové události před a jednu událost za zkoumaným slotem. Reflektujeme tím i předpoklad, že věta musí obsahovat nejméně dvě slova. Fundamentální frekvence je charakterizována průměrnou hodnotou v každém slově $\bar{F}0$ a aktivitou (5.9), která vyjadřuje, jak moc se fundamentální frekvence v daném slově mění. Abychom do slotu umístili tečku, musí fundamentální frekvence prvních dvou řečových událostí klesat a aktivita fundamentální frekvence druhého slova (uprostřed sledované trojice) musí dosáhnout prahu aktivity fundamentální frekvence. Podle charakteru třetí události hledáme dva případy:

- třetí událost v pořadí je hluk (neřečová událost v přepisu – např. nádech)
- třetí událost je řečová a má větší $\bar{F}0$ než obě předchozí

V prvním případě rovnou vkládáme tečku. Ve druhém případě požadujeme pro vložení tečky ještě dodatečné potvrzení. Tím je v našem případě skutečnost, že daný slot byl označen jako potenciální bod změny v nahrávce (a následujícími vrstvami strukturalizace nahrávky byla tato změna zamítnuta).

$$F0_{act} = \frac{max(F0) – min(F0)}{\bar{F}0} \tag{5.9}$$

Informace o slotech podezřelých ze změny v nahrávce je využita ještě jednou – generují tečku, pokud za nimi následuje neřečová událost. Tečky jsou následně vloženy do všech slotů, které jsou následovány neřečovou událostí s délkou minimálně 1 s. Takto definované "bezpečné tečky" a určené čárky rozdělí každou promluvu v nahrávce na sub-regiony s různou délkou.

Je-li délka (počet slov) dostatečně malá (blízká průměrné délce věty), je větný celek chápán jako finální. V opačném případě hledáme takové neřečové události uvnitř dlouhých segmentů, které by ho rozdělily na kratší věty. Na tyto dělicí body je aplikován požadavek, aby mezi nimi byl minimální/maximální počet slov (8/14). Tím dosáhneme větných celků přibližně průměrné délky.

Toto schéma vyžaduje jen velmi malou míru propojení s celkovým schématem zpracování nahrávky. Proto mohla být jeho přesnost vyhodnocena pro obě vypracovaná schémata strukturalizace nahrávky popsaná v sekcích 6.1 a 6.2.

Rozhraní nástroje

vstupní data: přepis nahrávky (řečové i neřečové události), fundamentální frekvence a energie řeči, detekované body změny v nahrávce, určený jazyk promluv
výstupní data: přepis s doplněnou interpunkcí
řídicí parametry: statistický popis jazyků získaný z korpusů (délky vět, blacklisty)

5.8.3 Interpunkční schéma B

Na rozdíl od předchozího interpunkčního schématu, toto schéma (dále označované jako *IS_B*) je úzce propojeno s celkovým schématem strukturalizace nahrávky. Definice slotů je provázána do té míry, že i body změny mluvčího jsou vázány na interpunkční sloty, jejichž počty se snažíme redukovat na základě různých zdrojů informací. Postupy vedoucí k redukci počtu slotů jsou popsány v sekci 6.2. Pro tuto chvíli postačí předpokládat, že doplnění interpunkce je provedeno po jednotlivých promluvách (odstavcích přepisu). Každá promluva má určený jazyk (všechny kroky redukce slotů i IS_B jsou plně implementovány pro oba jazyky) a některé sloty mají díky redukci slotů zakázané vložení interpunkce (tvoří víceslovné nedělitelné sekvence).

Pro každou promluvu je proveden post-processing s doplněním čárek a informace o umístění čárky je přiřazena odpovídajícímu slotu. Podle čárek je promluva rozdělena na úseky, které mohou být čtyř druhů:

- začátek promluvy – konec promluvy
- začátek promluvy – čárka
- čárka – čárka
- čárka – konec promluvy

Každý z těchto úseků představuje jednotku, do níž je zapotřebí doplnit interpunkci (jak tečky, tak čárky). Doplnění interpunkce spočívá v ohodnocení pravděpodobnosti přítomnosti interpunkce v každém slotu, na které navazuje výběr nejvhodnější kombinace znamének. Vyhodnocení vychází z principů dynamického programování a hledá takové umístění interpunkce, které minimalizuje kriteriální funkci (cenu za umístění interpunkčního znaménka).

Ohodnocení pravděpodobnosti přítomnosti interpunkčního znaménka ve slotu

Prvním kritériem, které ovlivňuje pravděpodobnost umístění interpunkce do slotu, jsou trendy krátkodobé energie (E) a fundamentální frekvence ($F0$). Každá položka přepisu je popsána buď jako neřečová událost (které trendy přerušují), nebo sadou čtyř parametrů [$\bar{E}$,E_{nd},$\bar{F0}$,$F0_{act}$] definovaných vztahy (5.5),(5.9). Víceslovné segmenty (vzniklé zablokováním slotu) jsou zpracovány jako sekvence jednotlivých událostí, každá z nich je parametrizována odděleně.

Pro ohodnocení pozorovaných trendů byly natrénovány rozhodovací stromy (pomocí nástroje *scikit-learn*[4]). Stromy byly trénovány pro dvě odlišné situace:

- $slovo_1 \quad slovo_2 \quad [slot] \quad slovo_3$
- $slovo_1 \quad slovo_2 \quad [slot] \quad hluk_3$

Tyto situace nepokrývají všechny analyzované sloty v nahrávce (možné kombinace řečových a neřečových událostí), zahrnují však ty, u kterých lze předpokládat interpunkci. Vstupním vektorem rozhodovacích stromů jsou $\left[\frac{\bar{F0}_2}{\bar{F0}_1} \quad \frac{\bar{F0}_3}{\bar{F0}_2} \quad F0_{act2} \quad \frac{\bar{E}_2}{\bar{E}_1} \quad \frac{\bar{E}_3}{\bar{E}_2} \quad E_{nd2}\right]$, respektive $\left[\frac{\bar{F0}_2}{\bar{F0}_1} \quad F0_{act2} \quad \frac{\bar{E}_2}{\bar{E}_1} \quad E_{nd2}\right]$.

Trénovací data byla získána ze 45 hodin ručně přepsaných nahrávek, kterým byly časové značky doplněny automaticky – metodou nuceného zarovnání (sekce 5.10). Jelikož některé čárky jsou "akusticky neznatelné", jako výskyt interpunkce ve slotu jsou označeny tečky v přepisu. Sloty bez interpunkce pak tvoří druhou kategorii. Při trénování byl použit princip cross-fold validation (80 % dat použito pro trénování, 20 % jako reference). V rámci trénování bylo natrénováno větší množství rozhodovacích stromů, které se lišily jak svou hloubkou, tak vlastnostmi "lístků" (požadavky na cílové skupiny, které se již dále neklasifikují). Vynecháme-li tečky na koncích promluv, tvoří sekvence "slovo slovo. slovo" cca 13 % všech teček, zbytek je tvořen sekvencemi "slovo slovo. hluk". Proto jsou sekvence tří slov trénovány na minimalizaci falešných detekcí, zatímco od slotů následovaných hlukem byla vyžadována

[4] http://scikit-learn.org/

Tabulka 5.3: Shrnutí přesnosti detekce interpunkce pomocí rozhodovacích stromů

	slovo slovo. slovo			slovo slovo. hluk		
	prec[%]	rec[%]	acc[%]	prec[%]	rec[%]	acc[%]
manuální	1,4	45,7	67,6	42,0	40,7	60,9
$strom_1$	1,2	35,4	70,9	39,4	76,7	52,2
$strom_2$	1,3	45,7	66,3	39,0	66,9	53,5
kombinace	1,1	20,3	**80,8**	38,6	**79,5**	50,3

dostatečná míra detekce interpunkce. Protože žádný ze stromů nedosáhl příliš kvalitních výsledků, byly pro každou situaci vybrány dva nejlepší rozhodovací stromy (s odlišnou hloubkou). Dále je pro každou situaci manuálně navržen jeden rozhodovací strom, který se podobá logice popsané v sekci 5.8.2. Každý strom určuje pravděpodobnost (v rozsahu 0,0—1,0), že daný slot obsahuje interpunkci. Aritmetický průměr výstupů těchto tří stromů pak určí celkovou pravděpodobnost přítomnosti interpunkce ve slotu (v tuto chvíli definovanou částečnou informací o prozodii). Jejich charakteristiky jsou shrnuty v tab. 5.3. Metriky použité k vyhodnocení detekce interpunkce pomocí rozhodovacích stromů jsou definovány v sekci 7.2. V tabulce je pro každý rozhodovací strom vyhodnocena *precision* (prec) – daná vztahem (7.5), *recall* (rec), podle vztahu (7.6) a *accuracy* (acc), zavedená vztahem (7.1).

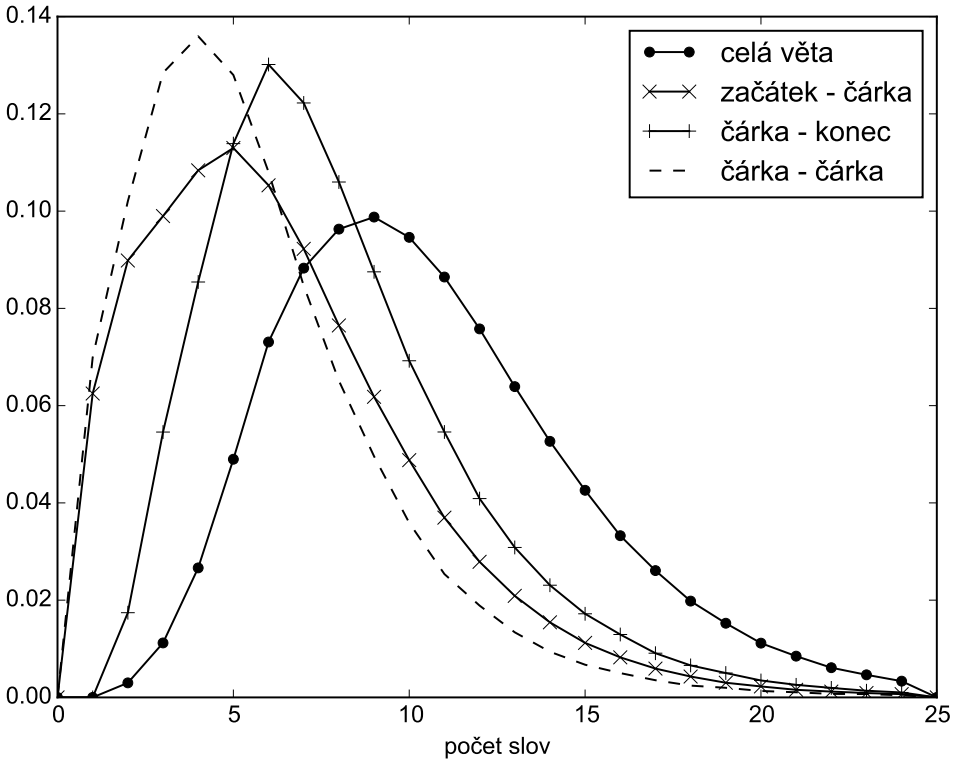
Vyhodnocení klasifikace pro daný slot bylo provedeno tak, že pravděpodobnost větší než 0,5 znamená přítomnost znaménka. Taková informace je zcela nezávislá na kontextu a dalších zdrojích informací. Interpretace této informace závisí na pravděpodobnosti okolních slotů (na kontextu). Pro vyhodnocení byla využita data popsaná v sekci 7.1.

Odhad vlivu kontextu (pravděpodobnosti umístění interpunkce do okolních slotů) byl na testovacích datech ověřen následujícím způsobem. Slot je označen za pozici interpunkce, pokud je jeho pravděpodobnost přítomnosti interpunkce větší než v okolních slotech (dva sloty před a za zkoumaným slotem) a pokud má pravděpodobnost umístění interpunkce hodnotu nejméně 0,5. Jako správná detekce jsou vyhodnoceny situace, kdy slovo před označeným slotem obsahuje interpunkční znaménko (.,?!;:), jinak se jedná o inzerci. Pro přesný popis experimentu je třeba dodat, že na testovacích datech nebylo provedeno úvodní předzpracování přepisu (redukce slotů) a nebere v úvahu ani následující kroky, které ovlivňují určenou pravděpodobnost přítomnosti interpunkce. Míra *recall* takto definované klasifikace je 36,1 %.

Druhým aplikovaným zdrojem informací jsou seznamy zakázaných slov, která zakazují umístění interpunkce. Čeština má definována 114 slov, před kterými se interpunkce penalizuje a 304 slov, za kterými je penalizována interpunkce. Pro slovenštinu bylo nalezeno 69 slov, před kterými se penalizuje interpunkce a 146 slov, za nimiž je penalizujeme. Penalizace je zohledněna snížením pravděpodobnosti přítomnosti interpunkce ve slotu o 0,35. Přítomnost delších hluků za slotem naopak zvyšuje pravděpodobnost interpunkce o faktor 0,20. Váhy jednotlivých zdrojů informací byly nastaveny empiricky na základě experimentů s vývojovými daty.

Vyhodnocení kriteriální funkce

Do algoritmu pro vyhodnocení kriteriální funkce vstupují dvě informace> 1) výše popsaná pravděpodobnost přítomnosti interpunkčního znaménka ve slotu, 2) statistika o typických délkách vět pro zpracovávaný jazyk. Statistiky pro češtinu jsou zobrazeny na obr. 5.5. Statistika slovenštiny je prakticky stejná, což je dáno tím, že jazyky mají téměř totožnou stavbu věty. Analýzou dalších jazyků (nejen slovanských) jsme však ověřili, že jiné jazyky mají charakteristiky velmi odlišné. Z těchto statistik lze určit průměrné/nejčastější délky daných typů větných celků a určit, jaké rozložení (různých délek různých typů úseků) je statisticky pravděpodobnější apod.



Obrázek 5.5: Délky větných celků v češtině

Vyhodnocení kriteriální funkce má za cíl ohodnotit skóre všech možných rozmístění teček, čárek nebo žádné interpunkce do volných slotů zkoumaného segmentu přepisu. Počet všech dostupných kombinací obsazení slotů je již pro 20 slotů neřešitelný "hrubou silou". Komplexita daná samotným počtem možných obsazení slotů je $O = N^3$. Proto je kritérium postupně vyhodnoceno pomocí generátoru, který nemusí udržovat v paměti všechny varianty, ale pouze omezenou sadu nejlepších obsazení slotů. Generátor navíc implementuje vnitřní prořezávání možných obsazení slotů a dostupné sloty jsou omezeny podle pravidel popsaných dále. To umožňuje vyhodnotit nejpravděpodobnější varianty rozmístění interpunkce za přijatelných výpočetních nároků. Omezující pravidla jsou:

- maximální povolený počet slotů obsazených interpunkčním znaménkem (vycházející ze znalosti průměrné délky věty),
- výběr slotů podezřelých z přítomnosti interpunkce (sloty musí přesáhnout minimální hodnotu pravděpodobnosti interpunkčního znaménka a být lokálními maximy) a
- využití širšího kontextu (např. bonusová pravděpodobnost v případech, kdy je mezi sousedními sloty velký počet blokovaných slotů)

Inicializace generátoru probíhá tak, že do prvního podezřelého slotu se vloží tečka/čárka/zůstane prázdný. Generátor pak opakuje následující kroky:

1. Ověření validity variant rozmístění znamének (splnění omezujících pravidel). Je-li varianta validní, ohodnotí se a porovná s nejlepšími výsledky.
2. Pokud již nejsou k dispozici žádné podezřelé sloty, generátor se ukončí.
3. Všechny předchozí varianty rozmístění znamének jsou rozšířeny o tečku/čárku/nic v dalším podezřelém slotu v pořadí.
4. Varianty jsou prořezány a generátor se vrací na krok 1.

Ověření validity rozmístění interpunkce bere v úvahu maximální povolené množství interpunkce a další pravidla, jako např. minimální rozestup teček. Ohodnocení varianty je dáno kriteriální funkcí, která bude popsána dále. Prořezávání variant porovnává skupiny variant, které mají na nejpokročilejším slotu (nejdále od začátku sekvence) stejné interpunkční znaménko. Varianty jsou ohodnoceny, jako kdyby nejpokročilejším znaménkem sekvence končila. Všechny varianty, které mají menší skóre než 0,8 nejlepšího skóre ve skupině, jsou vyloučeny z dalšího zpracování, čímž dochází k prořezání zkoumaných variant.

Jazykové statistiky jsou před aplikací normalizovány (nejčetnější varianta získá pravděpodobnost 1,0) a statistiky jsou rozšířeny o interpolaci ohodnocení delších sekvencí, než jaké byly pozorovány v korpusu (5.10). N značí počet slov ve větném celku, $p(N)$ zastupuje pravděpodobnost výskytu větného celku délky N, $p_{out}(N)$ značí výslednou pravděpodobnost výskytu větného celku délky N. $p_{red}(N)$ označuje pravděpodobnost délky sekvence daného typu po snížení vlivu jazykové statistiky (5.11). Redukovaná pravděpodobnost se používá, protože plný rozsah statistiky se ve výsledné kriteriální funkci chová příliš "agresivně".

$$p_{out}(N) = \begin{cases} \frac{p(N)}{max(p(N))} \; ; & N < 25 \\ 0,60 \; ; 25 \leq N < 30 \\ 0,45 \; ; 30 \leq N < 40 \\ 0,25 \; ; 40 \leq N < 50 \\ 0,15 \; ; N > 50 \end{cases} \tag{5.10}$$

$$p_{red}(N) = 0,8 + 0,2 * p_{out}(N) \tag{5.11}$$

Kriteriální funkce $KF(S)$ (5.12), která přiřazuje zkoumanému rozmístění interpunkce skóre, má dvě složky. První je kumulativní složka $CS(S)$, do níž se započítávají hodnoty pravděpodobnosti umístění interpunkce do slotu $p_{is}(s)$. Pokud má daná varianta rozmístění interpunkce (množina slotů S) ve slotu s umístěno interpunkční znaménko, přičítá se k $CS(S)$ hodnota $p_{is}(s)$, pokud ve slotu interpunkce být nemá, přičítá se doplněk této pravděpodobnosti (5.13). Druhou složkou kriteriální funkce je multiplikativní faktor $MP(S)$, jehož obsahem je součin pravděpodobností jednotlivých větných útvarů $p_{sent}(typ, N)$, které jsou závislé na typu větného útvaru a počtu slotů v něm N (5.14).

$$KF(S) = CS(S) \cdot MP(S) \tag{5.12}$$

$$CS(S) = \sum_{s \in S} \begin{cases} p_{is}(s)\ ;\ s \in \{,.\} \\ 1.0 - p_{is}(s)\ ;\ jinak \end{cases} \tag{5.13}$$

$$MP(S) = \prod_{v \in S} p_{red}(typ_v, N_v) \tag{5.14}$$

Jelikož toto interpunkční schéma využívá redukci počtu slotů, do kterých je možné umístit interpunkční znaménka, provedenou během strukturalizace přepisu, je jeho přesnost vyhodnocena pouze pro schéma s kumulovaným rozhodováním (popsaným v sekci 6.2). Ve schématu s izolovaným rozhodováním (sekce 6.1) totiž k apriorní redukci slotů v textovém přepisu nedochází.

Rozhraní nástroje

vstupní data: strukturalizovaný přepis nahrávky spolu s informacemi o slotech v přepisu a dostupná prozodická informace
výstupní data: přepis s doplněnou interpunkcí
řídicí parametry: statistiky popisující jazyk (délky vět, blacklisty), váhy informačních zdrojů

5.9 Datová struktura pro práci se strukturalizovaným dokumentem

Poslední komponentou nutnou pro popis schémat zpracování nahrávky je datový kontejner pro uložení vytvořeného informačně bohatého dokumentu. Kontejner musí umožnit ukládání všech mezikroků při tvorbě výsledného dokumentu. Je vhodné, aby umožnil manuální prohlížení a případnou editaci mezivýsledků i finálního dokumentu, a je zapotřebí, aby byl snadno převoditelný do formátu dat uložených v databázi archivu. Proto jsme při návrhu datové struktury pro práci s informačně bohatými dokumenty postupovali společně s tvůrci anotačního nástroje *NanoTrans* [63]. Výsledkem je vzájemně kompatibilní datový kontejner (založený na XML tazích), jehož struktura je ilustrována v příloze na obr. A.1.

Datový kontejner je rozdělen do tří sekcí. První z nich obsahuje doplňkové informace a meta-data. Druhá sekce obsahuje strukturovaný přepis nahrávky. Třetí sekce datového kontejneru zastřešuje informace o mluvčích a vazbu na databázi mluvčích.

V první sekci jsou uloženy veškeré dostupné informace o dokumentu (datum vzniku nahrávky, ID dokumentu, stanice a název pořadu a další získaná meta-data) a informace o přepisu (autor přepisu, použité verze akustických a jazykových modelů, verze strukturalizačního schématu). Dále může obsahovat informace o počítači, na kterém bylo zpracování provedeno, přesném čase provedení dílčích úloh apod.

Druhá sekce musí udržet informace o struktuře přepisu, doplňkových informacích a časové značky jednotlivých událostí v přepisu. Přepis je strukturován do čtyř úrovní, které připomínají knihu: kapitola/sekce/paragraf/fráze. První dvě úrovně slouží pouze k rozdělení přepisu. Například kapitoly zpravodajská relace jsou: přehled zpráv, samotné zprávy, sportovní zpravodajství a předpověď počasí. Sekce v kapitole zpráv mohou být kupříkladu zprávy z domova a zprávy ze zahraničí. Kapitoly ani sekce na sebe nemusí vázat žádnou doplňkovou informaci. Paragrafy odpovídají jednotlivým promluvám (respektive celistvým úsekům) v nahrávce. Je na ně navázána většina informací – identita mluvčího, jazyk promluvy, charakter přenosového pásma, kategorie paragrafu (znělka, řečový/neřečový obsah). Hlavní složkou paragrafů jsou pak fráze. Fráze nesou informaci o textu, fonetickém obsahu a časových značkách událostí v nahrávce. S ohledem na kroky post-processingu a strukturalizace nelze tvrdit, že fráze odpovídají událostem v nahrávce (ať už řečovým, či neřečovým). Fráze je spíše nejmenší zobrazenou položkou přepisu dokumentu.

Třetí sekce obsahuje lokální seznam mluvčích, kteří se vyskytují v jednotlivých paragrafech. Každý mluvčí je definován sadou atributů (jméno, příjmení, tituly, pohlaví, jazyk, uživatelské komentáře). Důležité je ID mluvčího, které udržuje propojení konkrétních paragrafů přepisu s globální databází mluvčích. Tato vazba je klíčová pro správnou indexaci dokumentu a vyhledávání v databázi archivu.

Pro ilustraci výše popsané struktury je v příloze (A.2) náhled uživatelského rozhraní nástroje *NanoTrans*, které lze porovnat se strukturou datového kontejneru.

5.10 Automatické zarovnání textu s nahrávkou

Nástroj popsaný v této sekci má dvojí využití. Jednak je součástí strukturalizačního schématu, které vychází z existence ručního přepisu dokumentu (sekce 6.3), za druhé umožňuje automatické doplnění časových značek do referenčních dat. Samotnou úlohu zarovnání textu s nahrávkou lze již dlouho považovat za vyřešenou. Proto se nástroj, který představujeme v následujícím textu, zaměřuje na robustní zarovnání nepřesného či neúplného přepisu s nahrávkou [64, 65]. Tyto odchylky mohou být různě rozsáhlé (od drobných přeřeků a reformulací po přebývající/chybějící úseky délky několika desítek sekund).

Kromě aplikací zmíněných v předešlé pasáži umožňuje automatické doplnění časových značek k existujícímu přepisu řešení úloh jako časování titulků[5] nebo automatická příprava trénovacích dat [66].

[5] https://www.stream.cz

Jednotlivé kroky zarovnání textu s nahrávkou jsou zobrazeny na obr. 5.6. Základem zarovnávacího modulu je rozpoznávač řeči, kterému upravíme jazykový model. Akustický model používáme stejný jako pro rozpoznávání zpracovávaného jazyka.

Prvním krokem je načtení textového přepisu a vytvoření omezeného jazykového modelu pro rozpoznávač řeči. Omezený LM umožňuje pouze přechody mezi po sobě jdoucími slovy přepisu, přeskočení slova nebo přítomnost jedné (nebo více) neřečových událostí mezi dvěma po sobě jdoucími slovy. Určitou volitelnou nadstavbou je svázání krátkých slov (spojky, předložky) s delšími následujícími slovy. Takto svázaná slova se při zarovnání chovají jako kolokace, které jsou standardní součástí slovníků rozpoznávače. Nevýhoda tohoto postupu spočívá ve skutečnosti, že výstupní zarovnaný text obsahuje menší počet "slov" než vstupní přepis (nelze je mapovat 1-na-1).

Druhým krokem, nutným pro rozpoznání nahrávky je opatření slov přepisu fonetickým přepisem (G2P). G2P může vycházet buď ze slovníků, nebo je proveden pomocí vážených stavových automatů (WFST). Speciální moduly umožňují přiřazení výslovnosti číslům, zkratkám (s.r.o., kg, apod.) a speciálním symbolům (např.: @,§). Rozpoznání je pak provedeno s velkým množstvím alternativních fonetických variant. Díky velmi omezenému jazykovému modelu (a slovníku) probíhá rozpoznání rychleji než plnohodnotné zpracování LVCSR systémem (cca 0,1 RT).

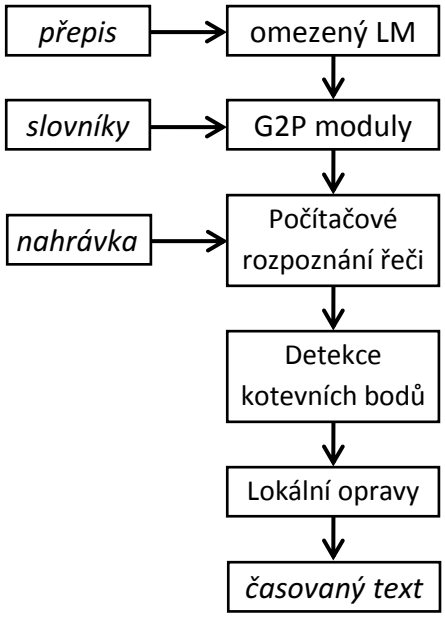
V případě, že textový přepis se neshoduje s nahrávkou, je vysoce pravděpodobné, že některé časové značky přiřazené přepisu rozpoznávačem budou chybné. Náš systém proto detekuje "kotevní body" – úseky přepisu, které podle určitých kritérií považujeme za správně načasované. Kotevním bodem jsou úseky přepisu požadované délky (minimální 15 fonémů), rozpoznané bez přerušení a musí být rozpoznány ve stejném pořadí, v jakém se objevují v přepisu. Druhý požadavek ošetřuje situace, kdy se některé sekvence slov mohou opakovat v přepsaném úseku nahrávky a v úseku, jehož přepis chybí.

Kotevní body rozdělují nahrávku na úseky, jejichž časováním si jsme jistí (ve výstupu značené jako *Hit*) a na obsah mezi kotevními body. Úseky, jejichž časováním si nejsme jistí, jsou podrobeny zarovnání rozpoznané fonetiky s fonetikou zpochybněného přepisu. Zarovnání je provedeno algoritmem navrženým v [31], který ve zbytku práce označujeme jako MED (Minimum Edit Distance).

Edit distance je způsob, jak vyjádřit rozdílnost dvou (textových) řetězců. Rozdílnost řetězců je kvantifikována počtem operací nutných k převedení jednoho řetězce na druhý. Těmito operacemi jsou *substituce* (náhrada znaku), *delece* (odstranění znaku) a *inzerce* (vložení znaku). Shoda obou řetězců na dané pozici se obvykle označuje *hit*. Tuto míru podobnosti lze rozšířit tak, že za znak považujeme celé slovo a řetězce jsou posloupnosti slov. Pokud každá operace dostane svou váhu (součet vah inzerce a delece volíme menší než váhu substituce), můžeme hledat takovou sadu operací, která vede na nejmenší celkovou penalizaci. K nalezení sady operací s minimální penalizací se užívá principů dynamického programování.

Na základě zarovnání fonetiky slov jsou slovům přiřazeny značky (*hit/inzerce/substituce/delece*). Tyto značky nám umožňují interpretovat segment s chybným přepisem (nebo časováním) a jsou proto součástí výstupu. Podle zarovnání fonetiky jsou upraveny časové značky výstupních slov. Větší úseky inzercí

znamenají, že v textu nejspíš chybí přepis části nahrávky, delece naopak značí přebytečný text. Samotná existence značek nám dále umožňuje říci, kterými časovými značkami je si nástroj "jistý" a které časové značky mají sníženou věrohodnost.



Obrázek 5.6: Postup zarovnání textového přepisu s nahrávkou

Rozhraní nástroje

vstupní data: parametrizovaná nahrávka (39 MFCC) a text, pro který hledáme časové značky
výstupní data: přepis obsahující původní text doplněný o časové značky a nejpravděpodobnější fonetickou reprezentaci, neřečové události v přepisu, informace o tom, jestli byla daná slova v nahrávce nalezena
řídicí parametry: AM a LM pro systém rozpoznání řeči, slovník(y), G2P moduly

6 Navržená schémata strukturalizace dokumentu

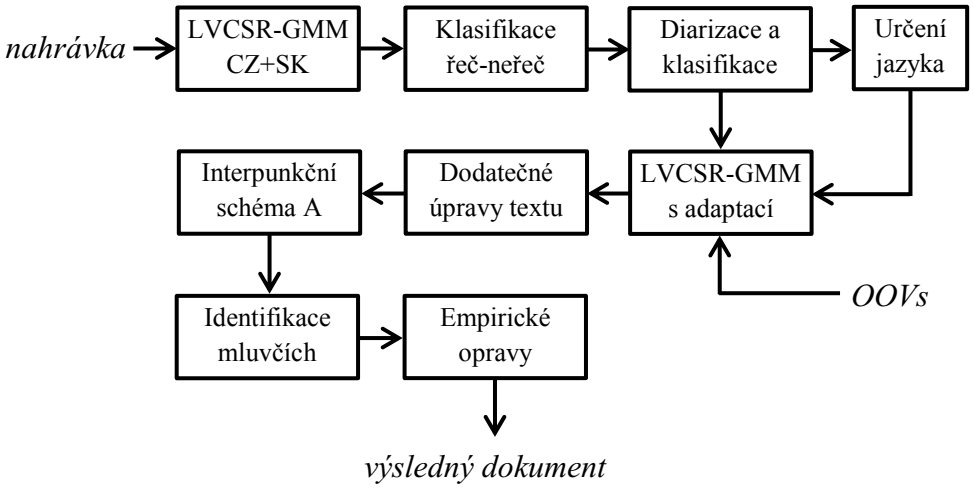
Proces strukturalizace nahrávky je posloupnost řady dílčích kroků, jejichž společným cílem je získat přesný a informačně bohatý přepis inventarizované nahrávky. Strukturalizační schéma lze tedy chápat jako řídicí logiku, která na základě dostupných informací o nahrávce spouští jednotlivé nástroje. Snaží se při tom dosáhnout maximální přesnosti funkce jednotlivých nástrojů a musí při tom respektovat vazby mezi nástroji (zejména jejich pořadí dané požadovanými vstupními daty). Základní úkoly strukturalizačního schématu (na které je kladen důraz v doposud vytvořených systémech – viz kap. 3) jsou: 1) zajištění podmínek pro optimální funkci systému rozpoznání řeči (správný výběr AM a LM, adaptace na mluvčího) a 2) získání informací potřebných pro indexaci přepisu (řečové události a jejich časové značky). Úkoly strukturalizačního schématu v této práci jsou rozšířeny o následující dva úkoly: 3) zajistit informace pro správné zobrazení výsledného dokumentu a 4) optimalizovat čitelnost přepisu a usnadnit orientaci v něm.

V následujících sekcích (6.1 a 6.2) je popsáno nejprve klasické schéma strukturalizace (*Strukturalizace s izolovaným rozhodováním*) a následně navrhujeme nový přístup ke strukturalizaci (*Strukturalizace s kumulovaným rozhodováním*). Poslední sekce (6.3) popisuje strukturalizační schéma nasazené v situaci, kdy je dostupný textový přepis dokumentu.

Rozdíl mezi schématy s izolovaným a kumulovaným rozhodováním spočívá v aplikaci informací produkovaných jednotlivými nástroji. Schéma s izolovaným rozhodováním (stejně jako systémy představené v kapitole 3) aplikuje informaci generovanou konkrétními nástroji (např. detekci šířky přenosového pásma) v okamžiku jejího získání, ale pro další rozhodování se již informace nepoužívá. Oproti tomu, schéma s kumulovaným rozhodováním vychází z teze, že jevy sledované strukturalizací dokumentu jsou vázány totožnou množinou slotů a určité úlohy strukturalizace proto mohou být řešeny společně, na základě informací získaných z více nástrojů. Schéma tedy nejprve shromáždí dostupné informace a teprve potom řídicí logika provede potřebné rozhodnutí (úlohy, které schéma řeší, pak nutně slučují více předchozích dílčích úkolů do komplexnějších otázek).

Obě navržená strukturalizační schémata používají odlišnou konfiguraci systému rozpoznání řeči (*LVCSR-GMM* a *LVCSR-DNN*). Zejména skutečnost, že LVCSR-DNN dosahuje velmi přesných přepisů i bez adaptace na mluvčího, nám umožňuje použít v obou schématech odlišné řazení jednotlivých nástrojů.

6.1 Strukturalizace s izolovaným rozhodováním



Obrázek 6.1: Strukturalizační schéma s izolovaným rozhodováním

Prvním krokem strukturalizačního schématu s izolovaným rozhodováním je rozpoznání nahrávky systémem rozpoznání řeči za využití jazykového modelu, akustického modelu a slovníků, které umožňují pozdější určení jazyka promluvy (sekce 5.3.1 a 5.3.3). Tímto rozpoznáním je provedena i detekce řečové aktivity a detekce hranic slov pro následující kroky. Doba provedení tohoto rozpoznání se pohybuje okolo 0,50 RT (doba provedení všech kroků strukturalizace je vyčíslena pomocí Real-Time faktoru - podílu času potřebného pro provedení výpočtu a doby trvání zpracovaného úseku nahrávky).

Následuje vyhledání neřečových segmentů nahrávky (trvá cca 0,05 RT). To je provedeno na základě výstupu LVCSR systému (sekce 5.4.1). Podle detekovaných neřečových segmemtů je upravena informace o řečové aktivitě pro další nástroje.

Diarizace nahrávky (sekce 5.4.2) a následující klasifikace segmentů (sekce 5.5) začínají detekcí bodů změny v nahrávce a jejich následujícím shlukováním. Každému řečovému segmentu je diarizací přiřazeno ID mluvčího a na základě těchto identit jsou definovány akusticky homogenní segmenty v nahrávce. Segmentům jsou přiřazeny další atributy: pohlaví mluvčího – muž/žena/neznámé, přenosové pásmo – plné(WB)/omezené(NB). Pro segmenty klasifikované jako řeč je určen jazyk promluvy (sekce 5.5.1) - čeština(CZ)/slovenština(SK). Jelikož modely rozpoznávače jsou nezávislé na pohlaví mluvčích (gender-independent), je celá časová osa nahrávky pokryta 5 kategoriemi (CZ-WB/CZ-NB/SK-WB/SK-NB/neřeč). U řečových segmentů umožňuje ID mluvčího provést adaptaci na mluvčího nad celým dokumentem místo nad jednotlivými promluvami. Všechny klasifikace jsou provedeny za cca 0,10 RT.

Klíčovým krokem celého schématu je konečné rozpoznání jednotlivých segmentů nahrávky. Schéma využívá dvouprůchodové rozpoznání s adaptací na mluvčího (sekce 5.3.1). Úseky pro adaptaci jsou získány předcházející diarizací nahrávky. Pro zvýšení přesnosti přepisu je možné přidat do jazykových modelů a slovníků slova mimo slovní zásobu (OOVs), získaná během inventarizace nahrávky. Příkladem může

být název pořadu, stručný popis jeho obsahu, jména hostů a další. Celé rozpoznání s adaptací trvá přibližně 3,10 RT (do času jsou započteny i režie okolo spouštění rozpoznání, přípravy segmentů atd.).

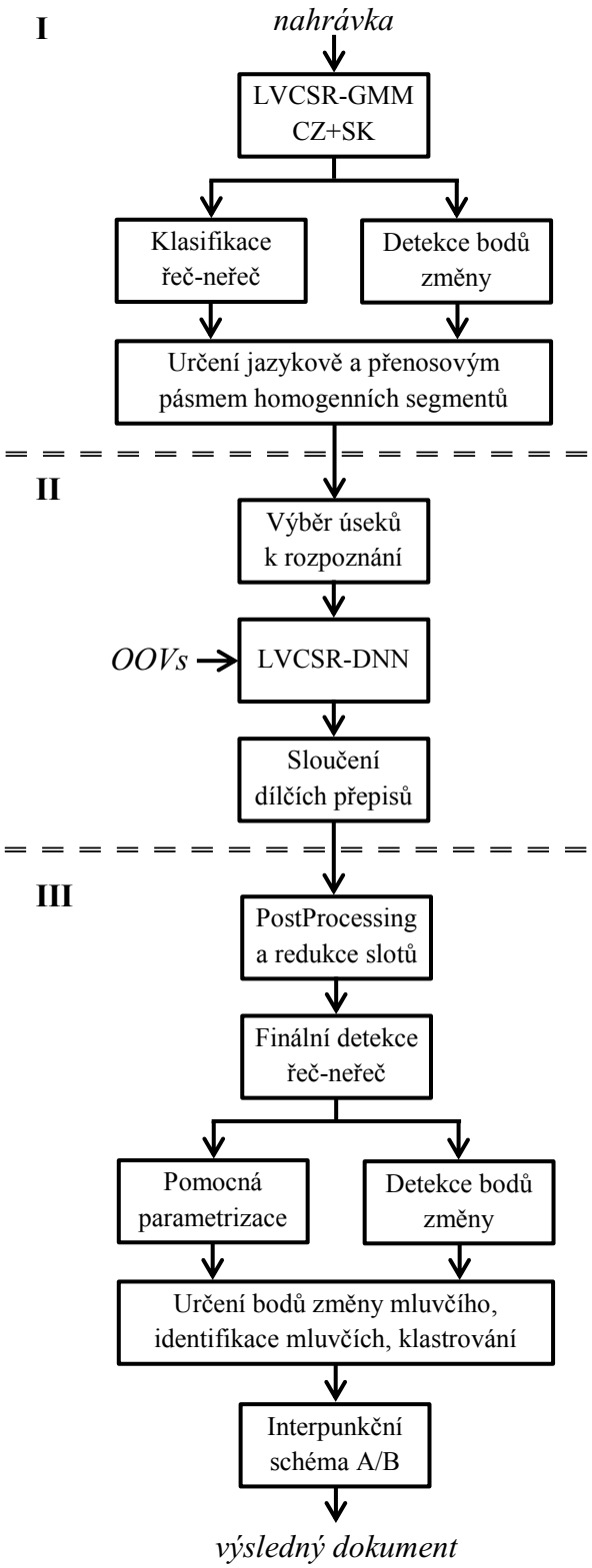
Další kroky zpracování nahrávky již vedou ke zlepšení čitelnosti získaného přepisu. Provedení úprav textu (post-processingu) a doplnění interpunkce (sekce 5.7 a 5.8.2) mění pouze textový obsah přepisu. Ve vyhodnocení experimentů zapojujeme v tomto schématu pouze *Interpunkční schéma A*, protože kroky vedoucí k redukci slotů (prováděné *Interpunkčním schématem B*) ovlivňují i segmentaci nahrávky a takový postup je nekompatibilní s tímto strukturalizačním schématem. Identifikace mluvčích (sekce 5.5.4) přiřadí jednotlivým promluvám mluvčího, pokud je rozpoznán s dostatečnou věrohodností. V opačném případě je k označení mluvčího použito dříve přiřazené pohlaví. Tyto kroky jsou provedeny do 0,10 RT.

Specifický obsah nahrávek může vést k atypickým chybám. Například některé znělky mohou generovat nežádoucí textový přepis, jména některých mluvčích mohou být špatně rozpoznána (převážně u cizinců a OOV), formátování názvů specifických institucí může být odlišné od standardního post-processingu a podobně. Takové chyby mohou být vázány na konkrétní skupiny nahrávek (např. různé díly stejného pořadu). Jsme-li schopni definovat, jak takové chyby korigovat, případně u kterých dokumentů opravu aplikovat, jsou tyto chyby odstraněny v rámci modulu heuristických oprav. Druhou kategorií heuristických oprav jsou opravy segmentace. Příkladem může být situace, kdy mezi dvěma segmenty promluvy téhož mluvčího je detekováno několik slov bez určené identity mluvčího a pohlaví mluvčích všech segmentů se shoduje. Všechny tři segmenty jsou pak sloučeny do jedné promluvy. U některých heuristických oprav nutně hledáme kompromis mezi objektivními vyhodnocovacími metrikami a dopadem na subjektivní čitelnost dokumentu.

6.2 Strukturalizace s kumulovaným rozhodováním

Slabinou schématu s izolovaným rozhodováním je, že jakákoliv chyba vzniklá ve kterémkoliv kroku strukturalizace se propaguje do dalších vrstev (celý proces strukturalizace je sekvenční posloupností jednotlivých nástrojů). Příkladem může být situace, kdy dojde k přílišné segmentaci (falešné body změny mluvčího). V takovém případě se během rozpoznávání zpracovávají segmenty jednotlivě a dochází k přerušení jazykového modelu. Schéma s kumulovaným rozhodováním se snaží této slabině čelit tím, že reformuluje požadované výstupy jednotlivých vrstev schématu a kombinuje v jednotlivých vrstvách informace ze všech relevantních nástrojů.

Strukturalizační schéma s kumulovaným rozhodováním je zachyceno na obr. 6.2. 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. Úloha druhé vrstvy spočívá ve vygenerování finálního textového přepisu nahrávky. Třetí vrstva určuje finální segmentaci dokumentu, doplňuje promluvám požadované informace a provádí formátování textu (PostProcessing). Jako čtvrtou vrstvu by bylo možné chápat export strukturalizovaného přepisu (a všech dostupných informací) do databáze archivu.



Obrázek 6.2: Strukturalizační schéma s kumulovaným rozhodováním

6.2.1 Vrstva I

Úvodním krokem první vrstvy je, stejně jako u předchozího schématu, rozpoznání kombinovaným česko+slovenským systémem rozpoznání řeči (sekce 5.3.1 a 5.3.3). Jeho výstup je podkladem pro provedení dvou vzájemně nezávislých analýz. První z nich je detekce bodů změny v nahrávce (sekce 5.4.2). 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), druhým typem je veškerý hudební obsah (znělky, písničky apod.). Neřečový obsah je detekován stejně jako v předchozím schématu (sekce 5.4.1) – přímo z výstupu systému rozpoznání řeči. Neřečové úseky jsou přeskupeny a definují neřečové regiony v nahrávce. 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 (nepotřebujeme klasifikovat parametry hluků či hudby). Body změny, nacházející se mimo neřečové regiony, ale v jejich těsném okolí (buď pod 0,25 s, nebo ve vzdálenosti 2 slov) jsou zakázány. To vychází z předpokladu, že věta má mít nejméně 2 slova (stejně jako v předchozím schématu). Pokud je bod změny detekován mezi krátkými neřečovými událostmi, je verifikován (detektorem změny aplikovaným nad redukovaným okolím odlišným od plovoucího okna) a na základě verifikace buď zrušen, nebo přesunut za předcházející řečovou událost.

Nad redukovanou sadou bodů změny jsou provedeny klasifikace přenosového pásma a jazyka (sekce 5.5). Každý úsek nahrávky je po provedení klasifikací označen jako neřečový segment, nebo řeč. Řečový segment má označení kombinované z jazyka (CZ/SK/neurčen) a z šířky přenosového pásma (NB/WB). Na základě označení segmentů jsou nalezeny co nejdelší nepřerušené úseky CZ-NB/CZ-WB/SK-NB/SK-WB/neřeč. Provedení vrstvy I trvá cca 0,80 RT.

6.2.2 Vrstva II

První dva kroky této vrstvy jsou poměrně přímočaré. Podle označení úseků získaných předešlou vrstvou jsou vystříhány spojité úseky (které jsou rozšířeny o případné sousední neřečové úseky). To nám umožňuje rozpoznat potenciálně sporné segmenty oběma sadami modelů (podle obou sousedních klasifikovaných úseků). Následuje rozpoznání vystříhaných úseků s pomocí LVCSR-DNN (viz sekce 5.3.2). Ve třetím kroku jsou dílčí přepisy kombinovány. Přepis úseků nahrávky obsahujících řeč je akceptován, segmenty původně označené za neřečový obsah jsou dále analyzovány. Jelikož výchozí rozpoznání LVCSR-GMM nedisponuje modely pro úzkopásmový přenosový kanál (NB), může dojít k záměně řeči za neřečové události. Vstupem pro analýzu jsou přepis navazující na předcházející úsek (”zleva”) a přepis předcházející navazujícímu úseku (”zprava”). Výběr přepisu se řídí následujícími pravidly:

- pokud se řečové události v obou přepisech nepřekrývají, jsou rozpoznané řečové události použity jako přepis
- pokud se detekovaná řeč překrývá a z jedné strany je použit úzkopásmový model, použije se přepis podle úzkopásmového přenosového pásma
- pokud se přepisy překrývají, akceptuje se přepis "zleva" až do přerušení neřečovou událostí a naváže se přepisem získaným z následujícího bloku
- nelze-li nalézt bod přerušení, je použit přepis "zleva", protože navazuje na jazykový model předchozí promluvy

Nakonec je provedeno nové klastrování nahrávky. Je totiž možné, že promluvy se stejnými modely, původně přerušené neřečovou událostí, podle nového přepisu navazují. 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. Vzhledem k překryvům rozpoznávaných segmentů a dalším režiím souvisejícím se spouštěním rozpoznávání trvá provedení vrstvy II cca 1,80 RT.

6.2.3 Vrstva III

Prvním krokem třetí vrstvy je provedení formátování textu (PostProcessingu). Pro následující moduly jsou důležité dva aspekty PostPocessingu. První z nich je označení slov, která patří k sobě tím, že jsou mezi nimi odstraněny mezery (např. "s.r.o.", "n.L."). Druhý aspekt spočívá ve "standardizaci" vstupů pro následující zpracování. PostProcessing zaručí správné použití velkých/malých písmen (což je důležité při zpracování jmenných entit) a zjednodušuje detekci sekvencí jako zkratky nebo číslovky (všechny pády jsou zahrnuty do jedné zkratky, převod "paragraf" na § apod.). Počínaje tímto krokem jsou všechny změny v přepisu (formátování i slučování slov v rámci zkratek či jmenných entit) zaznamenány. To později umožní zlepšit vyhledávání v indexech archivu (bude možné hledat "tři sta" i "300").

Pro **redukci slotů** je kromě PostProcesingu využito několik dalších zdrojů informací. Prvním z nich je vyhledání jmenných entit. Mezi ty patří jména osob (včetně případných hodností, titulů, oslovení a funkcí), jména obcí (pokud obsahují sekvence jako "pod Sněžkou", "nad Labem") a názvy soukromých společností (detekované podle sekvencí typu "a.s.", "s.r.o." atd.). Vyhledávání jmen osob je spuštěno výskytem vlastních jmen (viz tab. 6.1), v jejichž okolí se vyhledají příjmení a další složky jmenné entity. Zbylé jmenné entity jsou hledány v případě výskytu zkratkových částí názvu ("s.r.o.", "n.L."). Jména byla získána jak z naší databáze mluvčích, tak ze seznamů českých[1] i slovenských[2] státních institucí.

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

Tabulka 6.1: 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

Druhým zdrojem informací pro redukci slotů jsou nejrůznější oslovení a fráze ("dobrý den", "paní ředitelko", "předseda vlády" apod.). Fráze jsou definovány výčtem nejčastějších frází, oslovení jsou různými kombinacemi oslovení a funkcí, případně funkce a instituce.

Poslední zdroj informací zpracovává jednotky a speciální symboly navázané na čísla. Konkrétně se jedná o fyzikální jednotky, měny, speciální symboly (procento, paragraf) a o nalezení sekvencí, které mohou být interpretovány jako datumy. Shrneme-li výše zmíněné jmenné entity, získáme podobný seznam jako v [67] s tím rozdílem, že naše implementace se omezuje na použití slovníků a pravidel.

Dalším krokem zpracování nahrávky je konečná **klasifikace řeč-neřeč** (sekce 5.4.1). 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.

Cílem několika následujících modulů je nalezení správných dělicích bodů v nahrávce (ve smyslu oddělení promluv jednotlivých mluvčích). Z předchozích kroků již máme k dispozici informaci o jazyce nahrávky a parametrech přenosového pásma. Chceme tedy rozšířit přepis o identitu mluvčích, nebo alespoň o jejich pohlaví. Veškeré dále popsané kroky jsou již prováděny nad jednotlivými úseky nahrávky, které jsou odděleny buď změnou akustických modelů (jazyka, přenosového pásma), nebo přítomností dlouhého neřečového úseku.

Jednou větví analýzy úseků je **diarizace** (sekce 5.4.2). Tentokrát však nemá k dispozici "předběžný přepis" celé nahrávky, ale finální přepis úseku nahrávky s redukovanou sadou slotů. Druhou větví analýzy je **pomocná parametrizace**, jejímž cílem je zapojit do rozhodování veškeré dostupné informační zdroje:

- trend fundamentální frekvence a krátkodobé energie promluvy (sekce 5.6)
- přítomnost neřečových událostí v okolí
- seznamy zakázaných slov (sekce 5.8.3)
- přítomnost čárky ve slotu (určená PostProcessingem)

Jelikož jednotlivé informační zdroje představují slabé klasifikátory, jsou jejich rozhodnutí shrnuty do jednoho kumulativního skóre. Váhy jednotlivých informačních zdrojů (a jejich "závěrů") jsou shrnuty v tabulce 6.2.

Porovnáním pomocného skóre a výsledků diarizace je možné vytipovat body změny mluvčího, které mohou být falešnými detekcemi (jejich pomocné skóre je

Tabulka 6.2: 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

menší než 1,5 anebo je vzniklý segment kratší než 2 s). Jejich ověření je provedeno dalším spuštěním diarizačního nástroje (fixní okno dlouhé 3 s na obě strany od ověřovaného bodu).

Po ověření bodů změny mluvčího je nahrávka rozdělena na jednotlivé potenciální promluvy a můžeme identifikovat mluvčí. Před samotnou identifikací je určeno pohlaví mluvčích v jednotlivých promluvách a na základě pohlaví mluvčího, jazyka a přenosového pásma jsou vybrány modely mluvčích pro identifikaci (sekce 5.5.4).

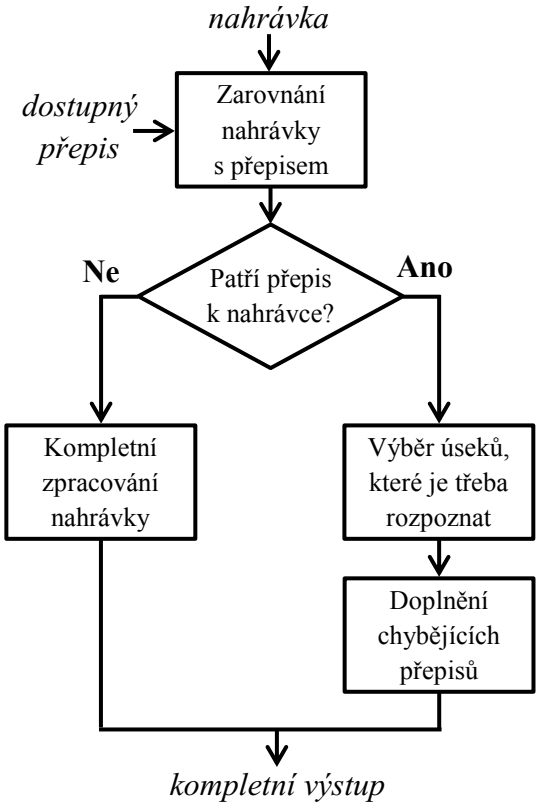
Posledním krokem zpracování nahrávky je doplnění interpunkce do přepisu. Původně bylo pro vícezdrojové schéma zpracování nahrávky navrženo interpunkční schéma B (sekce 5.8.3) vázané na provedení redukce slotů. S ohledem na porovnání schémat byla vytvořena i varianta, která aplikuje interpunkční schéma A (sekce 5.8.2), opět s výhodou redukované sady slotů. Provedení vrstvy III trvá 0,60 RT (z toho cca 70 % času zabírá detekce bodů změny a rozpoznání mluvčích).

6.3 Strukturalizace dokumentu s dostupným textovým přepisem

V přehledu stávajícího stavu problematiky (sekce 3.2.1) jsme zmínili, že je možné použít existující přepis nahrávky a využít ho místo rozpoznání nahrávky LVCSR systémem. Námi vyvinutý nástroj pro zarovnání textu s nahrávkou (sekce 5.10) je možné využít ke stejnému účelu. Navíc disponuje schopností ohodnotit důvěryhodnost vygenerovaných časových značek. Tuto schopnost jsme se rozhodli využít ve schématu využívajícím existující (i neúplný) přepis zpracovávané nahrávky.

Prvním krokem je zarovnání dostupného přepisu s nahrávkou. Na základě výstupu je zjištěno, nakolik se přepis shoduje s nahrávkou. K tomu slouží ohodnocení každého zarovnaného slova (Hit/Substituce/Delece/Inzerce), jak je popsáno v sekci 5.10. Z ohodnocení lze snadno určit 4 kategorie zarovnaných úseků:

- úseky, kde se přepis shoduje s nahrávkou (dostatečná převaha Hitů)
- vložené bloky neřečových událostí (Inzerce hluků)
- úseky v nahrávce, kterým chybí přepis (Inzerce řečových událostí)
- úseky přepisu, které se v nahrávce nenachází (Delece slov ze vstupního přepisu)



Obrázek 6.3: Hybridní strukturalizační schéma disponující přepisem nahrávky

Na základě tohoto dělení obsahu nahrávky lze rozhodnout, je-li shoda přepisu a nahrávky dostatečně vysoká, abychom přepis použili, nebo jestli je lepší použít systém rozpoznání řeči. V případě, že je shoda přepisu s nahrávkou dostatečná, jsou vybrány úseky, které je třeba dorozpoznat. Buď jim přepis chybí úplně, nebo vykazuje příliš malou shodu s nahrávkou. Před samotným vystřižením segmentů je v případě nutnosti (na hranici nejsou detekovány Hity) provedeno lokální zarovnání. Vyříznuté segmenty jsou zpracovány jedním ze schémat strukturalizace dokumentu (sekce 6.1 a 6.2). V závislosti na množství informací obsažených ve vstupním přepisu jsou doplněny požadované klasifikace a může být vytvořen finální výstup.

První výhodou tohoto schématu je možnost vypořádat se s přepisem v případě, kdy vstupní přepis neodpovídá nahrávce. Druhou výhodou je, že dokáže využít i částečně přesné (neúplné) přepisy a současně šetří výpočetní výkon oproti plnému zpracování nahrávky.

Celková doba zpracování nahrávky strukturalizačním schématem s izolovaným rozhodováním (sekce 6.1) je cca 3,85 RT. U schématu s kumulovaným rozhodováním (sekce 6.2) jsme dosáhli určité úspory času a trvá cca 3,20 RT. Oproti tomu, úvodní krok hybridního schématu zabere cca 0,35 RT a v optimálním případě tím zpracování nahrávky končí. V případě obsahu chybových segmentů pak trvá zpracování úměrně délce chybového úseku a použitému strukturalizačnímu schématu. Čili uspoříme min. 2,50 RT správně přepsaných úseků.

7 Experimentální vyhodnocení

7.1 Testovací data

Pro vyhodnocení výsledků dosažených navrženými strukturalizačními schématy bylo sestaveno šest skupin testovacích dat, jejichž základní popis je shrnut v tabulce 7.1. Všechny referenční přepisy jsou strukturované (rozdělené na promluvy s označenými mluvčími, jazykem promluvy a šířkou přenosového pásma). Časové značky byly do referenčních přepisů doplněny automaticky (sekce 5.10). Přepisy pochází ze dvou zdrojů. Prvním je již dříve zmíněný projekt NAKI, druhým jsou přepisy vytvořené společností Newton[1]. Jednotlivé testovací sady jsou vybrány tak, aby pokryly různé zastoupení jazyků v nahrávce, míru připravenosti/spontánnosti promluv, "akustické kvality" nahrávek i různou míru řečnických dispozic mluvčích. Současně se snažíme pokrýt vývoj jazyka výběrem nahrávek z různých časových období.

První množina testovacích nahrávek (*G1_LQ*) je sadou nejstarších nahrávek z archivu ČRo. Jejich akustická kvalita odpovídá technologiím z doby, kdy byly pořízeny. Obsahují proto řadu artefaktů způsobených nahrávacím zařízením a následnými přenosy mezi médii, na kterých byly nahrávky uloženy. Dokumenty mají charakter veřejných projevů, rozhlasových přednášek a rozhovorů s hosty.

Testovací sady *G2_CzSk* a *G3_Cz* zahrnují výhradně zpravodajské relace z období existence Československé a následně České republiky. Nahrávky jsou pořízeny studiově, s určitým podílem telefonních vstupů a projevů přehrávaných ze záznamu. Sada *G2_CzSk* byla vybrána tak, aby byl v každém dokumentu zaručen výskyt slovenštiny, což nám umožňuje vyhodnotit přesnost detekce jazyka.

Sady *G4_modern* a *G5_diskuze* představují moderní nahrávky s minimálním výskytem slovenštiny. Skupina *G4_modern* jsou novodobé zpravodajské relace. Skupina *G5_diskuze* je složena výhradně z diskuzních pořadů, u nichž je vyšší četnost změny mluvčích a promluvy jsou spontánnější (méně připravené). Současně lze předpokládat výskyt úseků nahrávky, kde hovoří více mluvčích současně.

Sada *G6_stream* představuje sadu pořadů z internetového portálu Stream.cz[2]. Tyto pořady jsou charakteristické přítomností nejrůznějších hluků, typicky jediným mluvčím a velkým množstvím znělek a hudby na pozadí. Proto ji spolu se sadou *G1_LQ* považujeme za ukázku akusticky náročných dat, která nám umožní porovnat výkon systémů rozpoznání řeči v příznivých/náročných akustických podmínkách.

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

Tabulka 7.1: Základní charakteristiky připravených sad testovacích dat

	období [od – do]	slov [počet]	promluv [počet]	interp. [počet]	délka [h:m]	pořadů [počet]	neřeč [h:m]	CZ [h:m]	SK [h:m]	WB [h:m]	NB [h:m]	zdroj přepisu
G1_LQ	1926–1953	35.654	1.356	6.588	73:27	50	4:38	68:49	0:02	64:51	3:58	NAKI
G2_CzSk	1971–1998	148.295	3.680	21.702	250:58	50	6:30	217:02	27:26	220:54	23:34	NAKI
G3_Cz	1971–1999	134.200	3.668	19.354	228:53	50	8:13	220:40	0:00	215:24	5:16	NAKI
G4_modern	2001–2010	36.319	1.228	6.243	52:25	37	1:02	50:47	0:36	46:33	4:50	Newton
G5_diskuze	2009–2014	70.676	2.684	16.202	101:13	23	2:18	98:55	0:00	98:55	0:00	Newton
G6_stream	2012–2014	46.555	3.326	9.797	82:54	100	12:49	70:05	0:00	70:05	0:00	Newton

7.2 Vyhodnocovací metriky

Pro vyhodnocení experimentů, prezentovaných v následujících pasážích, bylo zapotřebí zvolit vhodnou sadu vyhodnocovacích metrik. Mezi základní (široce užívané) metriky patří *správnost*, *přesnost* (angl. precision), *úplnost* (angl. recall) a harmonický průměr přesnosti a úplnosti (obvykle značený *F-measure*). Autoři [68] tuto sadu metrik rozšiřují o *Slot Error Rate* navrženou pro vyhodnocení přesnosti doplnění interpunkce. 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* (7.1), druhou je *correctness* (7.2). 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 3 využívá pro vyčíslení přesnosti systémů rozpoznání řeči metriku *word error rate* – *WER*, která je komplementární k *WAcc*. *WAcc* a *WER* jsou zavedeny vztahy (7.3) a (7.4).

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\,[\%] \tag{7.1}$$

$$correctness = \frac{TP}{N} = \frac{H}{N} \times 100\,[\%] \tag{7.2}$$

$$WER = \frac{S + D + I}{S + D + H} = \frac{S + D + I}{N} \times 100\,[\%] \tag{7.3}$$

$$WAcc = 1 - WER = \frac{N - S - D - I}{N} = \frac{H - I}{N} \times 100\,[\%] \tag{7.4}$$

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.

Přesnost (angl. precision), zavedená vztahem (7.5), 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 (7.6), 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 (7.7), 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.

$$precision = \frac{TP}{TP + FP} \times 100\ [\%] \tag{7.5}$$

$$recall = \frac{TP}{TP + FN} \times 100\ [\%] \tag{7.6}$$

$$F\text{–}measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \times 100\ [\%] \tag{7.7}$$

Slot Error Rate – SER lze chápat jako komplement correctness pro interpunkci doplněnou do textového přepisu (7.8). Chybějící i přebytečná interpunkční znaménka mají v této metrice stejnou váhu jako záměny jednotlivých typů znamének. Metrika je koncipována stejně jako *WER*.

$$SER = \frac{S + D + I}{S + D + H} = \frac{S + D + I}{N} \times 100\ [\%] \tag{7.8}$$

Kromě hojně používaných metrik představených výše se ukázalo nutné zavést několik dalších metrik pro vyhodnocení provedených experimentů. První z přidaných metrik adaptuje výpočet *WAcc* pro situaci, kdy jsme schopni přesněji namapovat výsledné a referenční události detekované LVCSR systémem. M událostí výsledku může odpovídat N událostem reference (proto značíme metriku $accuracy_{M2N}$). V rozšíření této metriky (7.9) 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 $ends$).

$$accuracy_{M2N} = \frac{H - I + ends + ws}{N} \times 100\ [\%] \tag{7.9}$$

Posledními upravenými metrikami jsou vážené verze *precision* (7.10), *recall* (7.11) a *F-measure* (7.12), určené pro vyhodnocení detekce bodů změny v nahrávce. Jejich definice předpokládá, že některé typy chyb jsou méně závažné než jiné a mohou se proto započítat s redukovanou vahou w_i. Tyto specifické chyby jsou tři. První z nich je posunutí bodu změny (místo interpretace jako pár FN-FP), počet posunutí značíme *shift*. Druhá je vložení nadbytečného bodu změny do oblasti dlouhé neřečové události (I_{NS}). Třetí specifickou chybou je vložení bodu změny mluvčího do slotu, který obsahuje interpunkci (na konec větného celku), její počet značíme I_{punct}.

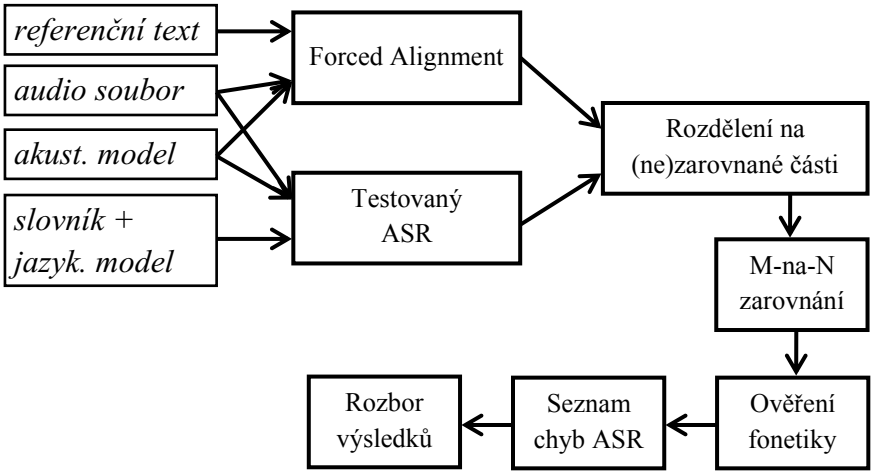
$$precision_w = \frac{TP + w_i \cdot shift}{TP + FP + w_i \cdot shift + w_i \cdot I_{NS} + w_i \cdot I_{punct}} \times 100\ [\%] \tag{7.10}$$

$$recall_w = \frac{TP + w_i \cdot shift}{TP + (FN - shift) + w_i \cdot shift} \times 100\ [\%] \tag{7.11}$$

$$F\text{–}measure_w = \frac{2 \cdot precision_w \cdot recall_w}{precision_w + recall_w} \times 100\ [\%] \tag{7.12}$$

7.3 Vyhodnocení přesnosti rozpoznání řeči s využitím časované reference

Metoda zarovnání výstupu systému rozpoznání řeči s referenčním textem užívající automatické časování reference je zachycena na obr. 7.1. Metoda spočívá v sekvenčním procházení časové osy referenčního přepisu a vyhledání shodných událostí ve zkoumaném přepisu. Tím se celá časová osa nahrávky rozdělí na segmenty (viz obr. 7.3), kde se výsledek shoduje s referencí a na "zbytek" ležící mezi regiony shody. Úseky s rozdílným obsahem jsou zarovnány pomocí M-na-N zarovnání, které je popsáno v následující sekci. Zarovnání se provádí podle textového (nebo fonetického) obsahu regionů (automatické časování doplňuje nejen časové značky, ale i nejpravděpodobnější fonetickou podobu zarovnávaných událostí).



Obrázek 7.1: Postup zarovnání referenčního a rozpoznaného přepisu

M-na-N zarovnání textu

Řada událostí, které mohou být detekovány během rozpoznávání, se nedá namapovat na referenční události 1-na-1 (např.: "protože" a "proto že"). Proto jsme navrhli algoritmus, který umožňuje sofistikovanější zarovnání dvou (textových) řetězců. Zarovnání textů předpokládá, že M-slov reference může odpovídat N-slovům výsledku. Rozšířený obal nástroje pak podporuje uživatelské nastavení citlivosti na velká/malá písmena, ignorované znaky (interpunkce apod.) a další. Samotné jádro zarovnání vychází ze zarovnání fonetických podob řečových událostí v regionu zájmu pomocí MED [31]. MED vygeneruje posloupnost událostí (shod, substitucí, delecí a inzercí), která umožní stanovit míru (ne)shody mezi libovolným párem slov z reference a výsledku. Libovolnému páru slov z reference a výsledku (referenční slovo značíme rf, slovo výsledku rs) odpovídá určitá subsekvence výstupu MED. Shody (hity) mezi nimi ohodnotíme váhou $w_H = 1{,}0$ a substituce $w_S = 0{,}7$. Jelikož chceme určit míru shody, váhu delecí a inzercí uvažujeme nulovou. V subsekvenci výstupu MED

se nachází N_H hitů a N_S substitucí, přičemž celá subsekvence má délku N_{tot}. Výpočet míry shody $S_p(rf, rs)$ je zachycen vztahem (7.13). Výsledek zarovnání a jeho interpretace jsou ilustrovány na obr. 7.2.

$$S_p(rf, rs) = \frac{w_H \cdot N_H + w_S \cdot N_S}{N_{tot}} \tag{7.13}$$

	sedákem	balili	můžeš	na	místě
se	**2.0**	0.0	0.0	0.0	0.0
také	**2.4**	0.7	0.0	0.0	0.0
bavili	0.0	**5.7**	0.0	0.0	0.0
muži	0.0	0.0	**3.4**	0.0	0.0
z	0.0	0.0	**0.7**	0.0	0.0
náměstí	0.0	0.0	0.0	**1.7**	**4.4**

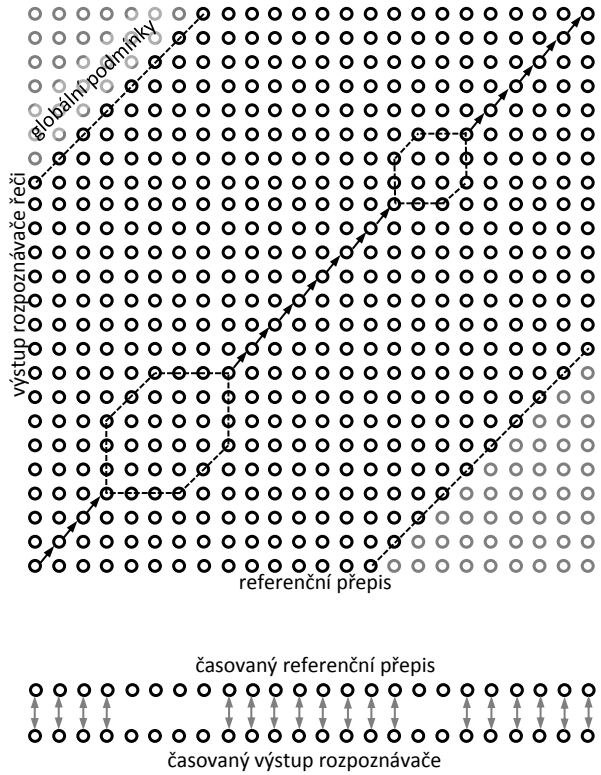
REF: se také bavili muži z náměstí
ASR: sedákem balily můžeš na místě

Obrázek 7.2: Ukázka zarovnání chybového úseku a hodnot skóre S_p

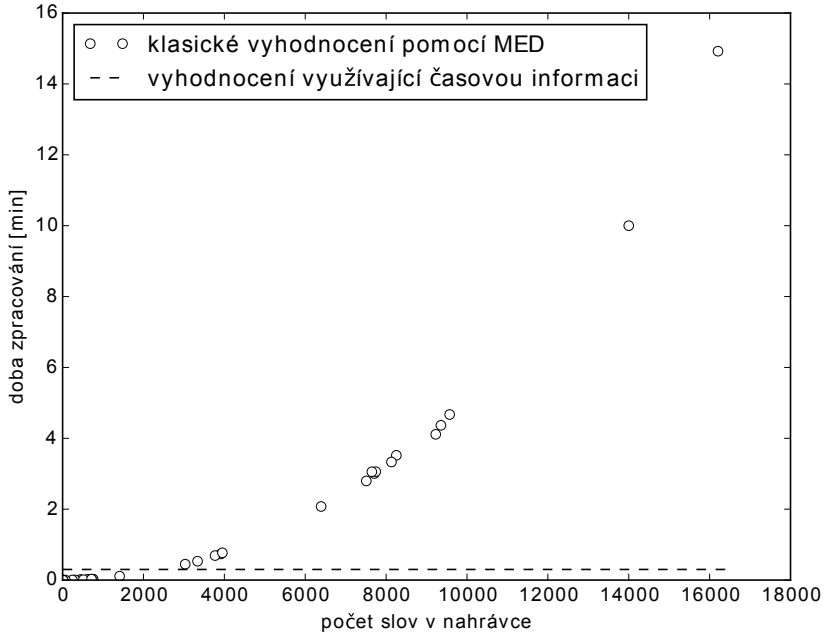
Zarovnání chybových segmentů uvnitř regionů bez shody umožňuje přesnější klasifikaci chyb. Ta je možná na úrovni fonetické (”měly” a ”měli”), nebo na úrovni textové (”takto” a ”tak to”). Analýzou zarovnaných úseků jsme kvantifikovali nejčastější typy chyb. Například chyby v koncovce slov (převážně sloves) tvořily 24,5% všech nalezených chyb LVCSR systému. Také jsme určili některé velmi časté záměny slov, což umožnilo najít chyby ve slovníku, rozšířit slovník o potřebná slova a definovat některé kolokace, které pokryly část záměn krátkých předložek. Konkrétní výčet nalezených chyb a přístup zvolený pro jejich řešení najdete v [69].

Navíc se ukazuje, že některé chyby nemusí mít plnou váhu, případně že jejichž subjektivní dopad na kvalitu přepisu je menší (např. fráze ”protože” a ”proto že”, ”toho je mu” a ”to ho jemu”). Chyby tohoto typu způsobují velký nárůst WER (současně se substitucí generují i inzerce/delece).

V předchozím textu jsme ukázali, že časovaná reference může být s výhodou použita k podrobnějšímu vyhodnocení přesnosti přepisu. Současně nám umožňuje zjistit, s jakou přesností jsou jednotlivým událostem přepisu přiřazeny časové značky. Navíc díky algoritmu popsanému v sekci 5.10 obsahuje reference informaci o přítomnosti neřečových událostí. Umožní nám tím vyhodnotit přesnost rozpoznání (a časové lokalizace) neřečových událostí, což čistě textová reference neumožňuje. Poslední výhodou přístupu je úspora výpočetního času. Jak naznačuje obr. 7.3, využitím časové informace je možné kvadratickou úlohu výpočtu MED redukovat na lineární úlohu doplněnou o lokální přepočty M-na-N zarovnání. Získanou úsporu času zobrazuje graf na obr. 7.4. Úvodní časování reference nepočítáme do času vyhodnocení (provede se jednorázově při anotaci referenčních dat). Již při délce nahrávky 3.000 slov je použití časované reference časově efektivnější.



Obrázek 7.3: Ilustrace úlohy zarovnání referenčních dat s přepisem za využití časované reference



Obrázek 7.4: Porovnání výpočetních nároků výpočtu WER metodou MED a při využití časované reference

7.4 Porovnání použitých konfigurací LVCSR

Abychom byli v dalších pasážích schopni odhad*precision* (Prec) - vztah (7.5), *recall* (Rec) – vztah (7.6) a *F-measure* – vztah (7.7).out vliv strukturalizačního schématu na přesnost rozpoznání dokumentu, provedli jsme porovnání základní výkonnosti obou použitých konfigurací rozpoznávače – LVCSR-GMM a LVCSR-DNN. Vhodnou množinou testovacích dat je G3_Cz, která obsahuje pouze češtinu a reprezentuje nejobsáhlejší složku zpracovávaných archivů – zpravodajské relace. Obě nastavení rozpoznávače jsou porovnány jak v přesnosti rozpoznání řečových událostí (tab. 7.2), tak ve schopnosti detekovat neřečové události v nahrávce (tab. 7.3). Schopnost detekovat neřečové události (hluky) je důležitá pro řadu použitých modulů, které vyžadují informaci o řečové aktivitě v nahrávce. Metrikami použitými pro vyhodnocení rozpoznání řečových událostí jsou *word accuracy* – *WAcc* daná vztahem (7.4), její upravená verze ACC_{M2N} daná vztahem (7.9) a *correctness* – Corr dle vztahu (7.2). Rozpoznání neřečového obsahu je vyhodnoceno pomocí *precision* (Prec) - vztah (7.5), *recall* (Rec) – vztah (7.6) a *F-measure* – vztah (7.7). Doplňkovou informaci pak představuje podíl neřečových úseků, které jsou časovány s odchylkou větší než 50 ms a 100 ms (označeny MT-50 a MT-100).

Tabulka 7.2: 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

Tabulka 7.3: Porovnání přesnosti použitých konfigurací LVCSR – neřečové události

	Prec [%]	Rec [%]	F-measure [%]	MT-50 [%]	MT-100 [%]
LVCSR-GMM	92,6	94,6	93,6	6,4	3,4
LVCSR-DNN	89,0	83,5	86,1	22,8	13,9

7.5 Porovnání nástrojů pro doplnění čárkové interpunkce

V této sekci porovnáme náš nástroj pro dolplnění čárkové interpunkce (sekce 5.8.1) se dvěma současnými systémy. Náš systém bude značen jako *FST*, porovnávaný český systém [61] bude značen *SET* a slovenský nástroj [62] bude značen *SVK*.

7.5.1 Systémy pro doplnění čárkové interpunkce pro češtinu

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 1 mluvčího 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 7.4:

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*)

Tabulka 7.4: Porovnání nástrojů pro doplnění čárkové interpunkce pro češtinu

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

Porovnáním dosažených výsledků lez snadno dojít k závěru, že SET produkuje menší množství nadbytečných čárek než FST. Na druhou stranu FST dokáže detekovat o něco více čarek než SET. Výsledná *F-measure* se příklání mírně k použití FST. Pokusy s kombinováním obou nástrojů nepřinesly výrazně lepší výsledky – mezi oběma nástroji existuje cca 85% shoda na pozicích označených čárkou, což nedává velký prostor pro další zlepšení.

Druhý závěr zohledňuje kvalitu dat vstupujících do nástrojů. Pro oba nástroje platí, že čím větší je množství vstupních informací (ruční značení konců vět) a přesnost přepisu, tím lepších výsledků nástroje dosahují. Důležitý je v tomto ohledu rozdíl mezi čárkováním bezchybného (ručního) přepisu nahrávky a čárkováním výstupu rozpoznávače (v našem případě s cca 10% WER).

7.5.2 Systémy pro doplnění čárkové interpunkce pro slovenštinu

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 [62], 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

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

Tabulka 7.5: Porovnání nástrojů pro doplnění čárkové interpunkce pro slovenštinu

	SVK	FST
precision [%]	95,3	96,3
recall [%]	49,6	49,0
F-measure [%]	65,3	65,0

přepisech v režimu, kdy jsou známy konce vět (zpracování je provedeno po větách). To odpovídá schématu *sent_manual* v předchozí sekci. Testovací data našeho systému představoval korpus 150 MB textů zpráv. Výsledky porovnání jsou shrnuty v tabulce 7.5 a dá se říct, že výsledky obou systémů jsou téměř totožné.

Porovnáme-li výsledky systémů pro zpracování češtiny a slovenštiny (pro odpovídající scénář) lze naléze určité paralely. Oba systémy pro doplnění čárek do slovenštiny mají precision cca 95 %, zatímco jejich české protějšky se pohybují okolo hodnoty 85 %. *Recall* obou porovnávaných systémů pro zpracování slovenštiny je přibližně 50 %, což je srovnatelné s průměrnou hodnotou obou českých systémů (46 % a 54 %).

7.6 Porovnání schémat pro strukturalizaci dokumentu

7.6.1 Značení experimentů

V následujícím textu budeme porovnávat dvě schémata zpracování nahrávky, dva moduly pro doplnění interpunkce a experimenty budou prováděny nad různými množinami testovacích dat. Značení používá následující zkratky: SCh_{IR} označuje strukturalizační schéma s izolovaným rozhodováním popsané v sekci 6.1, zatímco SCh_{KR} značí schéma s kumulovaným rozhodováním popsané v sekci 6.2. V experimentech závislých na použitém interpunkčním modulu značí *_PA* interpunkční schéma A (sekce 5.8.2), analogicky *_PB* označuje interpunkční schéma B (sekce 5.8.3). U většiny experimentů pak rozlišujeme dvě skupiny testovacích nahrávek složené ze základních balíčků (viz sekce 7.1). První z nich (značená *_LQ*) zahrnuje akusticky náročná data G1_LQ a G6_stream. Druhá skupina zahrnuje 4 zbylé testovací skupiny složené z běžného rozhlasového vysílání ("broadcast" *_BC*).

7.6.2 Vyhodnocení detekce bodů změny v nahrávce

Pro vyhodnocení schopnosti schémat detekovat body změny v nahrávce a dále je zpracovat jsou použity metriky *precision* (Prec), *recall* (Rec), *F-measure* a *accuracy* (Acc). Hodnoty uváděné v tabulce 7.6 v závorkách jsou hodnoty vážených metrik ($precision_w, recall_w, F\text{–}measure_w$) definovaných v sekci 7.2, váha $w_i = 0,5$.

U akusticky náročných nahrávek dosahuje přesnější detekce bodů změny SCh_{KR}. Jedním z hlavních důvodů je, že provádí segmentaci nahrávky až při znalosti konečného přepisu nahrávky. Tento přepis je pořízen LVCSR-DNN, který rozpoznává

akusticky náročné nahrávky lépe než první průchod LVCSR-GMM, který používá pro segmentaci SCh_{IR}.

Nahrávky vytvořené ve "standardních" podmínkách jsou o něco lépe zpracovány SCh_{IR}. Důvodem jsou občasné chybějící detekce neřečových událostí LVCSR-DNN (viz tab. 7.2), které způsobují, že neřečové události se zahrnou do popisu mluvčích před a za potenciálním bodem změny a vedou tak k inzerci bodů změny. Jedna nedetekovaná neřečová událost může takto způsobit až dvě inzerce bodů změny (vložení nového segmentu a návrat k původnímu mluvčímu).

Tabulka 7.6: Detekce bodů změny v nahrávce

	Prec [%]	Rec [%]	F-measure [%]	Acc [%]
SCh_{IR}_LQ	28,74 (37,97)	19,96 (24,10)	23,56 (29,48)	97,67
SCh_{KR}_LQ	30,56 (37,26)	56,32 (64,18)	39,62 (47,15)	97,07
SCh_{IR}_BC	73,08 (80,70)	72,13 (81,83)	72,60 (81,26)	99,20
SCh_{KR}_BC	62,74 (69,98)	74,09 (82,64)	67,95 (75,79)	98,97

7.6.3 Vyhodnocení segmentace nahrávky

Kvalita segmentace nahrávky spočívá v několika faktorech shrnutých v tabulce 7.7. Prvním faktorem je správnost (nebo nesprávnost) volby modelů, s jejichž pomocí jsou jednotlivé úseky nahrávky rozpoznány. Úseky lze v našem případě rozdělit do 5 kategorií:

- úseky bez řečové aktivity (*neřeč*) – včetně hudby a znělek
- úseky v češtině nahrané ve studiu (*CZ_WB*)
- telefonní vstupy v češtině (*CZ_NB*)
- úseky ve slovenštině nahrané ve studiu (*SK_WB*)
- telefonní vstupy ve slovenštině (*SK_NB*)

Správnost klasifikace ve smyslu "neřeč – ostatní kategorie" je zachycena ve sloupci *VAD*. Podíl řečových úseků, kterým byly přiřazeny správné modely (jazyk i šířka přenosového pásma), je pak obsahem sloupce *modely*. Sloupce *neřeč⇒řeč* a *řeč⇒neřeč* zachycují množství vzájemných záměn řečových a neřečových úseků. Poslední sloupec (*mluvčí*) vyčísluje, jakému množství řečového obsahu bylo přisouzeno správné pohlaví mluvčího.

Chceme-li správnost klastrování zhodnotit podrobněji, umožní nám to zobrazit referenční a výsledné klastrování ve formě matice záměn. Matice záměn – tab. 7.8 a tab. 7.9 jsou zobrazeny pro testovací sadu G3_Cz. Jednotlivá čísla v matici záměn značí, jaké procento celého přepisu kategorie zahrnuje.

Tabulka 7.7: 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

Tabulka 7.8: Matice záměn klasifikace úseků nahrávky: $SCh_{IR}\_G3\_CZ$ (hodnoty jsou vyjádřeny v % celkového trvání nahrávky)

SCh_{IR} \ ref.	neřeč	CZ_WB	CZ_NB	SK_WB	SK_NB
neřeč	**1,62**	0,05	0,00	0,01	0,00
CZ_WB	1,21	**87,39**	1,37	0,05	0,00
CZ_NB	0,23	7,27	**0,05**	0,00	0,00
SK_WB	0,00	0,07	0,00	**0,00**	0,00
SK_NB	0,00	0,00	0,02	0,00	**0,05**

Tabulka 7.9: Matice záměn klasifikace úseků nahrávky: $SCh_{KR}\_G3\_CZ$ (hodnoty jsou vyjádřeny v % celkového trvání nahrávky)

SCh_{KR} \ ref.	neřeč	CZ_WB	CZ_NB	SK_WB	SK_NB
neřeč	**2,29**	0,21	0,01	0,01	0,00
CZ_WB	0,76	**93,98**	2,02	0,00	0,00
CZ_NB	0,00	0,00	**0,00**	0,00	0,00
SK_WB	0,01	0,40	0,04	**0,00**	0,00
SK_NB	0,00	0,00	0,02	0,00	**0,00**

7.6.4 Vyhodnocení modulů pro doplnění interpunkce

Jak jsme ukázali v [49], volba umístění a typu interpunkčních znamének je nejednoznačná úloha i v případě, že ji provádějí vysoce kvalifikovaní anotátoři (v tomto případě studenti Katedry českého jazyka, FP, TUL). Podstatně vyšší je shoda na samotné volbě pozic pro umístění interpunkce (slotů). Jelikož všechny referenční přepisy, které máme k dispozici, jsou přepsány právě jedním anotátorem, provádíme vyhodnocení teček a čárek zvlášť a poté hromadné vyhodnocení pozic všech slotů s interpunkcí. Pozice slotů nám ukazují, nakolik se systému daří oddělovat jednotlivé větné (významové) celky v promluvě. Sada interpunčkních znamének, které mají anotátoři k dispozici, je bohatší než sada použitá našimi systémy. Za shodu mezi referencí a výsledkem považujeme, pokud počítač umístí čárku do slotu, kde se v referenci nachází čárka, středník, dvojtečka nebo pomlčka. Analogicky je shoda umístění tečky do slotu, kde je v referenci tečka, vykřičník nebo otazník.

Pro odhad výchozího stavu (bez použití interpunkčních schémat) předpokládáme přepis s bezchybnou detekcí bodů změny v nahrávce (korektně určenými promluvami) a bezchybným přepisem. V takovém scénáři je každá promluva ukončena tečkou a žádná jiná interpunkce se v dokumentu nenachází. Protože jednotlivé sady testovacích dat vykazují velkou variabilitu množství interpunkce, rozhodli jsme se je vyhodnotit každou zvlášť (viz tab. 7.10). Metriky použité k popisu jsou *precission* (Prec), *recall* (Rec), *F-measure* a *accuracy* (Acc) popsané v sekci 7.2. Jako doplňkový parametr udáváme množství interpunkčních znamének na 1.000 slov referenčního přepisu (P_1k). K interpretaci metrik je zapotřebí dodat, že samotné metriky nedokážou přesně popsat výslednou (individuálně vnímanou) čitelnost dokumentu. Naše experimenty [49] ukázaly, že uživatelé preferují raději nadbytek interpunkce než chybějící interpunkční znaménka (věta delší než cca 15 slov se již stává téměř nečitelnou, zatímco přílišnou segmentaci uživatel snadno odhalí). Základní přesnost interpunkce je shrnuta v tabulce 7.10.

Tabulka 7.10: Přesnost doplnění interpunkce bez aplikace interpunkčních schémat

	Prec [%]	Rec [%]	F-measure [%]	Acc [%]	P_1k
G1_LQ	100,0	12,24	21,84	86,06	**185**
G2_CzSk	100,0	12,93	22,89	89,63	**146**
G3_Cz	100,0	13,66	24,03	89,88	**144**
G4_modern	100,0	14,54	25,39	88,11	**171**
G5_diskuze	100,0	11,98	21,40	83,82	**229**
G6_stream	100,0	20,18	33,58	86,36	**210**
skupina LQ	100,0	17,12	29,23	86,23	**199**
skupina BC	100,0	13,07	23,13	88,46	**163**

Z tabulky 7.10 lze vypozorovat několik informací o jednotlivých podmnožinách testovacích dat. Za prvé lze říci, že novější nahrávky obsahují více interpunkce (obsahují kratší větné celky). Totéž lze tvrdit i o diskuzních pořadech, kdy se mluvčí rychleji střídají a jsou užívány kratší věty. Také lze říci, že starší rozhlasové relace (G2_CzSk a G3_Cz) obsahují podstatně delší větné celky než ostatní podmnožiny. V tabulce 7.11 jsou pak porovnány různé kombinace schémat zpracování nahrávky a interpunkčních modulů. Jejich "základní výkon" bez doplnění interpunkce je odhadnut v řádcích *skupina LQ* a *skupina BC* tabulky 7.10.

Tabulka 7.11: Porovnání použitých interpunkčních modulů

	tečky			čárky			sloty				
	Prec [%]	Rec [%]	F-measure [%]	Prec [%]	Rec [%]	F-measure [%]	Prec [%]	Rec [%]	F-measure [%]	SER [%]	Acc [%]
$SCh_{IR}\_PA\_LQ$	43,51	59,69	50,33	64,92	38,88	48,64	59,32	56,83	58,04	16,36	85,15
$SCh_{KR}\_PA\_LQ$	40,55	65,41	50,06	68,11	42,85	52,60	58,31	63,09	60,60	16,46	85,35
$SCh_{KR}\_PB\_LQ$	41,45	49,39	45,07	50,13	47,49	48,77	55,42	57,42	56,40	17,46	84,13
$SCh_{IR}\_PA\_BC$	56,47	77,06	65,18	83,09	54,26	65,65	71,20	68,66	69,91	8,95	91,92
$SCh_{KR}\_PA\_BC$	48,15	72,05	57,72	82,41	53,35	64,77	65,88	66,91	66,39	10,25	90,71
$SCh_{KR}\_PB\_BC$	49,04	66,50	56,45	59,61	57,09	58,32	61,20	66,85	63,90	11,43	89,64

7.6.5 Vyhodnocení souvislosti strukuralizace dokumentu a přesnosti automatického přepisu

V této sekci nejprve vyčíslíme přesnost, s jakou jsou ve finálních dokumentech rozpoznány řečové i neřečové události (tabulky 7.12 a 7.13). K tomu použijeme stejné metriky, pomocí kterých jsme porovnávali LVCSR-GMM a LVCSR-DNN v sekci 7.4. Kromě výsledků pro testovací sady "LQ" a "BC" jsou napočítány i metriky pro testovací sadu G3_Cz. To nám spolu se znalostí kvality segmentace přepisu (sekce 7.6.3), která má přímý vliv na volbu akustických a jazykových modelů, umožní popsat klíčové vazby uvnitř schémat srukturalizace nahrávky.

Tabulka 7.12: Porovnání přesnosti rozpoznání řeči v rámci navržených schémat

	WAcc [%]	Corr [%]	Acc_{M2N} [%]
SCh_{IR}_LQ	48,44	61,19	65,37
SCh_{KR}_LQ	49,17	64,42	66,86
SCh_{IR}_BC	83,29	92,11	94,89
SCh_{KR}_BC	81,49	90,18	92,77
SCh_{IR}_G3	84,86	93,49	96,27
SCh_{KR}_G3	82,20	90,40	93,21

Tabulka 7.13: Porovnání přesnosti detekce neřečových událostí

	Prec [%]	Rec [%]	F-measure [%]	MT-50 [%]	MT-100 [%]
SCh_{IR}_LQ	61,70	79,47	69,46	30,63	22,51
SCh_{KR}_LQ	69,04	76,25	72,47	30,53	21,06
SCh_{IR}_BC	86,19	84,64	85,41	20,81	13,02
SCh_{KR}_BC	81,49	81,96	84,53	21,80	13,12
SCh_{IR}_G3	87,55	85,85	86,69	22,69	15,10
SCh_{KR}_G3	89,06	82,07	85,76	23,94	14,85

Analýzou matic záměn mezi jednotlivými typy segmentů (tabulky 7.8 a 7.9) lze určit, že schéma SCh_{IR} správně klasifikuje 89,11 % celé délky nahrávky, zatímco schéma SCh_{KR} správně klasifikuje 96,27 % (součet položek na diagonále matice záměn). Pokud ovšem nasčítáme úseky klasifikované jako řeč se správně určeným jazykem (bez ohledu na šířku přenosového pásma), zjistíme, že obě schémata mají prakticky stejnou šanci rozpoznat správný text (99,03 % u schématu SCh_{IR} oproti 99,06 % u schématu SCh_{KR}). LVCSR-GMM ve SCh_{IR} využívá adaptaci na mluvčího, takže je zde předpoklad dosažení lepších výsledků v důsledku adaptace. LVCSR-DNN ve SCh_{KR} již nemůže dosáhnout žádného zlepšení, naopak chybná volba modelů může přesnost výstupu snížit. Konkrétní změny vyčíslené ve formě hitů (H), substitucí (S), inzercí (I) a delecí (D) jsou zachyceny v tabulce 7.14.

Tabulka 7.14: Změny přesnosti rozpoznání řeči způsobené zapojením LVCSR do strukturalizačního schématu

	H	S	I	D	*WAcc* [%]	Corr [%]
LVCSR-GMM	146.894	17.422	4.506	5.736	82,80	91,30
⇓	⇓	⇓	⇓	⇓	⇓	⇓
SCh_{IR}	150.416	15.003	4.311	5.047	84,86	93,49
LVCSR-DNN	150.993	13.921	3.565	5.852	85,50	96,70
⇓	⇓	⇓	⇓	⇓	⇓	⇓
SCh_{KR}	145.465	16.537	4.219	7.748	82,20	90,40

Porovnáme-li LVCSR-GMM a SCh_{IR}, je zde vidět zlepšení, kterého je dosaženo díky adaptaci rozpoznávače na mluvčího. Tyto výsledky jsou jen mírně horší než základní přesnost LVCSR-DNN. U LVCSR-DNN a SCh_{KR} dochází na první pohled ke zhoršení dosažených výsledků. Jak ukázala podrobná analýza vyhodnocených dokumentů, přibližně 90 % nových "chyb" je způsobeno aplikací post-processingu (formátování textu). Formátování textu je integrální součástí SCh_{KR}, kde napomáhá redukci počtu slotů v přepisu. Typickým příkladem takto vzniklých odchylek výsledku od referenčního textu je zpracování číslovek. Například úprava *tři sta šedesát šest : 366* generuje jednu substituci a 3 delece. Po odečtení tohoto vlivu by SCh_{KR} dosáhlo přesnosti přepisu minimálně srovnatelné se SCh_{IR}. Zbylá množina chyb souvisí především s drobnými chybami v segmentaci, které mohou způsobit, že krátká slova (spojky, předložky) na začátku promluv mohou být rozpoznávačem "přeslechnuty". Těchto chyb je však naprosté minimum.

7.6.6 Shrnutí dílčích experimentů

V této sekci shrneme dílčí porovnání obou navržených schémat zpracování dokumentu (sekce 6.1 a 6.2) a popíšeme nejdůležitější interakce mezi jednotlivými moduly, ze kterých jsou schémata sestavena.

Prvním úkolem obou schémat je detekce bodů změny v nahrávce. Z výsledků, shrnutých v tabulce 7.6 je zřejmé, že při zpracování akusticky náročných dat (LQ) dosahuje lepších výsledků schéma s kumulovaným rozhodováním (SCh_{KR}), zatímco u standardního rozhlasového vysílání (BC) dosahuje lepších výsledků schéma s izolovaným rozhodováním (SCh_{IR}). Důvodem lepšího výkonu SCh_{KR} nad testovací množinou LQ je fakt, že SCh_{KR} provádí konečnou detekci bodů změny až po konečném rozpoznání obsahu nahrávky. Má tedy k dispozici nejlepší dostupnou detekci řečové aktivity. Oproti tomu SCh_{IR} provádí detekci bodů změny s využitím základního přepisu, jehož akustické modely nejsou tak robustní. Současně však SCh_{KR} dosahuje horších výsledků nad standardní testovací sadou (BC). Důvod spočívá v nižší přesnosti detekce neřečových událostí při použití LVCSR-DNN modelů, které je zachyceno v tabulce 7.3. Důsledkem nepřesností detekce neřečových událostí v nahrávce je zahrnutí hluků do popisu porovnávaných řečových úseků (sekce 5.4.2). Tyto hluky započtené mezi popis řeči pak způsobují falešné detekce bodů změny.

V sekci 7.6.3 je vyhodnocena kvalita segmentace zpracovávané nahrávky. Tabulka 7.7 ukazuje, že SCh_{KR} mírně překonává SCh_{IR} jak v celkovém rozlišení řeč/neřeč, tak ve volbě odpovídajících modelů rozpoznávače (akustických a jazykových). Mírně horších výsledků dosahuje SCh_{KR} při označení řečových úseků nahrávky za neřečové. Rozdíl je však zanedbatelný.

Nejdůležitějším kritériem pro porovnání obou schémat je dosažená přesnost přepisu nahrávky, protože od ní se odvíjí pravděpodobnost vyhledání dokumentu žádaného uživatelem. Jak jsme již popsali v sekci 7.6.5, LVCSR-DNN aplikovaný ve schématu SCh_{KR} dosahuje vyšší přesnosti rozpoznání dokumentu než LVCSR-GMM aplikovaný ve schématu SCh_{IR}. U převážně studiových nahrávek (BC) není rozdíl výrazný, ale u dat s nižší kvalitou nahrávek (LQ) se výrazně projevuje vyšší robustnost akustických modelů založených na hlubokých neuronových sítích.

Dalším faktorem, který stojí za zhodnocení, jsou celkové časové nároky na zpracování nahrávky jednotlivými schématy. SCh_{IR} potřebuje přibližně 3,85 násobek délky nahrávky (RT), zatímco SCh_{KR} 3,20 RT. To představuje úsporu 17 %, která je při zpracování velkých archivů znatelná.

Pro doplnění interpunkce jsme vyhodnotili tři kombinace schématu zpracování nahrávky a modulu doplňujícího interpunkci. Jsou to SCh_{IR}_PA, SCh_{KR}_PA a SCh_{KR}_PB. Tabulka 7.11 shrnuje výsledky dosažené jednotlivými kombinacemi nástrojů. Poněkud paradoxně dosahuje nejlepších výsledků kombinace SCh_{IR}_PA, kde je vložení interpunkce silně závislé na neřečových událostech v nahrávce (patrně proto používají lingvisté pojem "pauzová interpunkce" při popisu interpunkce doplněné do mluveného jazyka). Stejný interpunkční modul aplikovaný ve druhém schématu (SCh_{KR}_PA) dosahuje o něco horších výsledků. Na první pohled by se mohlo zdát, že toto snížení přesnosti doplnění interpunkce je způsobeno metodikou "redukce slotů". Proto jsme vyčíslili, kolik slotů je sloučeno a v kolika případech je v zaniklých slotech přítomna interpunkce. Průměrně je zrušeno 1,75 % všech slotů, přičemž 1,55 % ze zrušených slotů obsahuje v referenci interpunkční znaménko. Všechny tyto chyby však nelze připsat samotnému postupu redukce slotů, neboť ne všechnu interpunkci bychom byli schopni doplnit a navíc ne všechny sloty jsou opatřeny správným textovým přepisem. Lze tedy konstatovat, že redukce slotů není hlavní příčinou poklesu přesnosti doplnění interpunkce. Hlavní příčinu spatřujeme, stejně jako u detekce bodů změny, v nepřesnostech detekce řečové aktivity, které zkreslují doplňkovou parametrizaci jednotlivých řečových událostí a tím i pravděpodobnost vložení interpunkce do daného slotu (viz sekce 5.6). Za hlavní příčinu nižší výkonnosti interpunkčního schématu "B" – viz SCh_{KR}_PB – lze označit aplikaci statistického popisu předpokládaných délek větných celků (sekce 5.8.3). Jak nám totiž ukazuje tabulka 7.10, četnost interpunkce v přepisu je závislá nejen na historickém období pořízení nahrávky, ale hlavně na typu pořadu a míře spontánnosti promluv. Snaha namapovat obecný typ promluvy na výše zmíněnou statistiku délky větných celků zpravodajských pořadů je proto nevhodná.

8 Zkušenosti z praktického nasazení

Strukturalizační schémata popsaná v této práci byla postupně nasazena ke zpracování části archivu Českého rozhlasu[1]. Pro provoz celého systému jsou (kromě systému provádějícího inventarizaci nahrávek a strukturalizaci z nich vytvořených dokumentů) potřeba uživatelské rozhraní (pro vyhledávání a zobrazení nalezených dokumentů), databáze již vytvořených přepisů a stream-server, který umožňuje přehrávání a navigaci ve zvukových stopách dokumentů. Uživatelské rozhraní umožňuje formulovat dotazy do databáze archivu a zvolit vhodná omezení vyhledávacích kritérií. Na základě výsledků obdržených z databáze dokumentů (jejíž rozsah je zachycen v tab. 8.1) jsou uživateli zobrazeny pořady, které odpovídají zadanému dotazu, a je možné přistoupit k prohlížení nalezených dokumentů. Přehrávací část uživatelského rozhraní pak zobrazí nalezený přepis a díky propojení se stream-serverem přehrává odpovídající zvukovou nahrávku.

Vyhledávací rozhraní (obr. 8.1) umožňuje nalézt hledaný text v rámci omezujících podmínek. Je možné vybrat mluvčího (jméno, pohlaví, moderátor), jazyk promluvy (češtinu, slovenštinu), časové rozmezí vzniku nahrávky (datum, čas, den v týdnu), rozhlasovou stanici, případně její pořad. Citlivost vyhledávání nastavuje přístup k velkým/malým písmenům a lze zvolit vyhledávání v "low-quality" přepisech (např. ve znělkách, hudebním obsahu či hlučných nahrávkách).

Přehrávání nalezených dokumentů (obr. 8.2) zobrazí strukturalizovaný přepis nalezeného dokumentu a současně s přehráváním příslušné zvukové stopy zvýrazňuje přehrávané řečové události. Navigace v přehrávaném dokumentu funguje obousměrně – jak posunem v časové ose zvukové stopy, tak kliknutím na text řečových událostí. Systém lze nalézt na adrese https://naki.ite.tul.cz.

Tabulka 8.1: 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

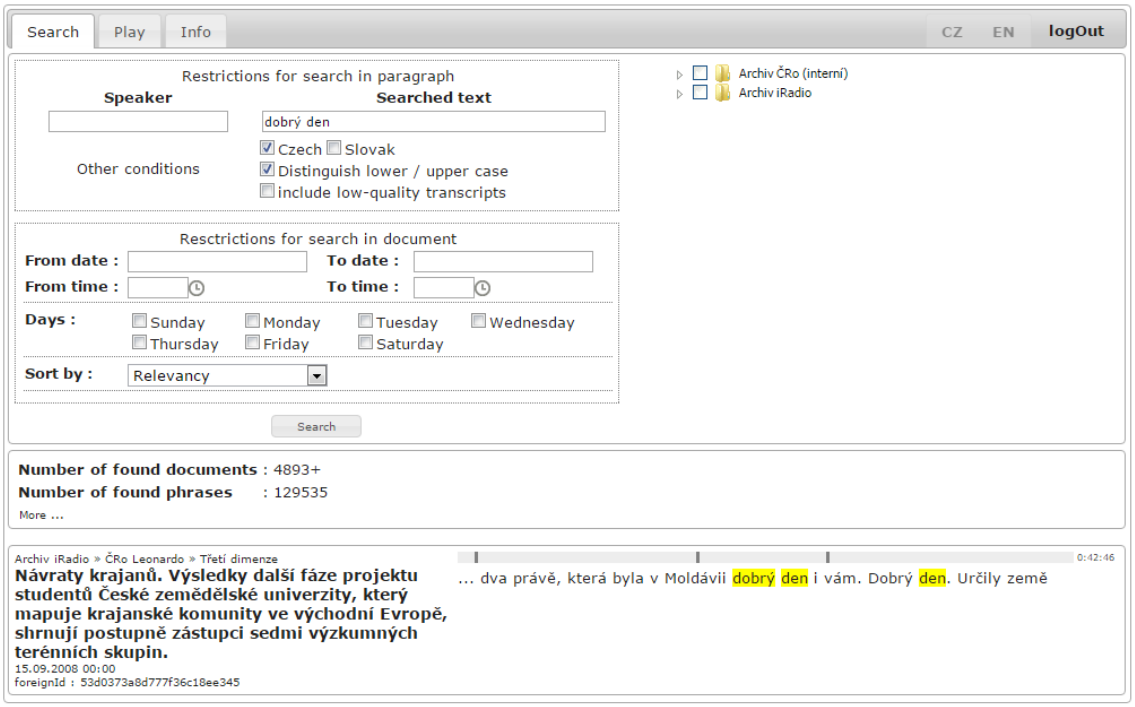
[1]projekt Ministerstva kultury ČR: DF11P01OVV013; Zpřístupnění archivu Českého rozhlasu pro sofistikované vyhledávání

Filtry vyhledávacího rozhraní umožňují uživatelům analyzovat dokumenty z různých úhlů pohledu. Velmi cenná je v tomto ohledu možnost pozorovat vývoj jazyka, slovní zásoby [70] i výslovnosti v čase [71]. Propojení obsahu dokumentu se zvukovou stopou a časovými značkami umožňuje zkoumat vztahy mezi psaným a mluveným jazykem [72]. Výše zmíněné výzkumy využívající archiv ČRo byly provedeny v rámci projektu NAKI, v průběhu tvorby a zpřístupňování archivu. Ačkoli je archiv veřejně dostupný necelý rok, začala ho využívat i další výzkumná pracoviště. Příkladem může být oddělení současné lexikologie a lexikografie ÚJČ AV ČR, jehož pracovníci užívají archiv při ověřování výslovností konkrétních slov při přípravě nového Akademického slovníku současné češtiny. Další probíhající výzkum se zabývá dynamikou slovní zásoby (zejména vlastními jmény a jejich užíváním).

Vyhledávání v archivu samozřejmě využívají i pracovníci ČRo. Velmi cenným nástrojem je archiv pro rešeršisty, kteří vytváří programovou nabídku pro redaktory. S archivem pracují i zvukoví technici a zaměstnanci, kteří provádí hledání v archivech na základě dotazů výzkumníků.

Dosavadní zpětné vazby od uživatelů archivu ukazují, že přebytečnou interpunkci dokážnou uživatelé snadno ignorovat. Během poslechu vyhledané nahrávky lze snadno přehlédnout i chyby v přepisu, pokud se ovšem nejedná o vlastní jména. U segmentace nahrávky na promluvy jednotlivých mluvčích požadují uživatelé hlavně detekci bodů změny mluvčího. Situace, kdy je promluva rozdělena na více částí, uživatelům nevadí, pokud je přerušení promluvy umístěno v bodě, ve kterém končí logický (větný) celek. Tohoto zjištění využívá i zobrazovací rozhraní, které dlouhé promluvy rozděluje (na základě znalosti interpunkce a výskytu dlouhých neřečových událostí), aby bylo dosaženo lepší orientace v přepisu.

Jelikož se řada analýz archivu zaměřuje na vývoj mluveného jazyka a jeho atributy, je potřeba zmínit, jak byly vytvořeny jazykové modely a slovníky pro jednotlivá historická období. Tvorba jazykových modelů vycházela z modelů dostupných pro moderní publicistické pořady. Na základě historických písemných pramenů byl jazykový model adaptován, včetně zařazení nových slovníkových položek (jména význačných osob a institucí) a úpravy výslovností slov, která se u některých slov vyvíjela (zejména u slov cizojazyčného původu). Zdroji pro rozšíření jazykových modelů byly naskenované výtisky Rudého práva (1945-1989), zápisy ze zasedání Národního a Federálního shromáždění, ČNR, SNR a texty veřejných projevů. Krokem navazujícím na získání textových korpusů byla adaptace jazykového modelu, při níž byly využity jak výše uvedené zdroje, tak i metoda nesupervizované adaptace učící se na automaticky získaných přepisech.



Obrázek 8.1: Vyhledávací rozhraní systému NAKI

Obrázek 8.2: Přehrávací rozhraní systému NAKI

9 Závěr

9.1 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 představené v přehledu současného stavu problematiky (kapitola 3) implementují první dva úkony, které zahrnujeme do strukturalizace přepisu. Jedná se o segmentaci nahrávky na homogenní úseky (ne nutně odpovídající promluvám mluvčích) a klasifikaci vlastností těchto segmentů. Cílem je zjistit, které AM, LM a slovníky jsou optimální pro rozpoznání konkrétních úseků nahrávky. Systémy tak dosahují maximální možné přesnosti přepisu, který svou strukturou i zobrazením připomíná spíše titulky než přehledný dokument. Přepis dokumentu (řečové události a jejich časové značky) spolu s klasifikacemi úseků přepisu a meta-daty nahrávky představují veškerou informaci nutnou k indexaci inventarizovaného dokumentu.

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 systému rozpoznání řeči navazují další klasifikace (zejména určení identity mluvčího). Podle atributů přiřazených jednotlivým úsekům nahrávky je přepis strukturalizován, čímž jsou definovány konečné hranice promluv. Čitelnost jednotlivých promluv je pak optimalizována po jednotlivých promluvách (podle nástrojů dostupných pro jazyk promluvy).

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 zdrojů ke zlepšení funkce jednotlivých subsystémů zapojených do procesu strukturalizace nahrávky. Ilustrací může být využití textového obsahu přepisu k redukci možných bodů v nahrávce, ve kterých se

může změnit mluvčí. 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 [73], míra redukce slotů by mohla dosáhnout až 40 %, kdy autoři detekují ustálená slovní přísloví, aforizmy apod. Výsledky SCh_{KR} překonávají SCh_{IR} ve většině parametrů. Výjimkou jsou chyby způsobené chybějící detekcí některých neřečových událostí.

Chceme-li porovnat přesnost rozpoznání nahrávek mezi systémy MALACH, SpeechFind a systémy navrženými v této práci, musíme nejprve zmínit několik zásadních faktů. První skutečností je časový odstup (cca 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 (viz tab. 7.12). Zajímavé pak je, že ignorujeme-li bílé znaky a chyby v koncovce slov (obvykle sloves), dostává se WER pod 10 % (sloupec Acc_{M2N}).

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). LVCSR-DNN také prokázal vyšší robustnost a přesnost rozpoznání řečových událostí v nahrávce než LVCSR-GMM. Porovnání přesnosti obou konfigurací rozpoznávání je shrnuto v tab. 7.2 a 7.12. Současně se ukázalo, že LVCSR-DNN má nižší přesnost detekce neřečových událostí než LVCSR-GMM – viz tab. 7.3.

Nasazená interpunkční schémata ukázala, že pro doplnění čárkové interpunkce lze použít statistické modely vycházející z jazykových korpusů (sekce 7.5). V otázce detekce hranic větných celků se statistický přístup neosvědčil. Oproti tomu, přístup vycházející z prozodických příznaků (zejména neřečových událostí) dosáhl lepších výsledků (viz tab. 7.11). Současně se ukázaly velké rozdíly v množství interpunkce v různých typech pořadů (zpravodajská relace / diskuzní pořad).

Vytvořený vyhodnocovací nástroj (sekce 7.3), který využívá (automaticky) časovanou referenci, nám umožňuje časově efektivnější vyhodnocení řady aspektů strukturalizovaného přepisu. Současně jsme schopni vyhodnotit nejen WER systému rozpoznání řeči, ale i přesnost detekce neřečových událostí. Správně určené neřečové události mají zásadní vliv na strukturalizaci přepisu. Možnost vybrat akustické modely, které dosahují nejen vysoké přesnosti přepisu, ale i spolehlivé detekce neřečových událostí, je proto předpokladem pro dosažení lépe strukturovaných přepisů.

Výsledky práce lze shrnout v následujících bodech:

- V kapitole 3 je shrnut aktuální stav problematiky.
- Byla navržena dvě schémata strukturalizace nahrávky a jedno schéma využívající existující textové přepisy (kapitola 6).
- Byla navržena hierarchická struktura elementů přepisu a jejich vazba na zpracovávanou nahrávku.
- Byly navrženy moduly pro členění textového přepisu nahrávky, včetně doplnění interpunkčních znamének (sekce 6.1,6.2 a 5.8).
- Připravená sada testovacích dat (sekce 7.1) umožňuje vyhodnotit přesnost testovacích přepisů, segmentaci nahrávky a různé úrovně strukturalizace nahrávky.
- Zarovnávací nástroj využívající časovanou referenci (sekce 7.3) umožňuje vyhodnocení všech potřebných metrik, včetně časové přesnosti událostí přepisu.
- V sekcích 7.4 a 7.6.5 jsou porovnány výhody a nevýhody dvou konfigurací akustického dekodéru systému rozpoznání řeči.
- Klíčové vazby mezi moduly použitými pro strukturalizaci dokumentu jsou shrnuty v sekci 7.6.6.
- Výsledky dosažené navrženými schématy jsou shrnuty a porovnány v sekci 7.6.

Nejvýznamnější přínos oproti současnému stavu problematiky (shrnutým v sekci 3) spatřuji v následujících čtyřech bodech. Za prvé, zatímco systémy popsané v sekci 3 produkují pro zpracovaný dokument přepis, jehož strukturu lze nejlépe přirovnat k titulkům (zajišťují zpracování nahrávky ASR systémem a získávají informace potřebné pro indexaci rozpoznaného obsahu nahrávky), strukturalizační schémata navržená v této práci vytvářejí přepis, který umožňuje zobrazení jako celistvý dokument s označením řady doplňkových informací (zajišťují přehledné zobrazení výsledného dokumentu a zlepšují čitelnost a orientaci v dokumentu). Takové zobrazení dokumentu klade mnohem větší nároky nejen na přesnost vytvořeného přepisu, ale hlavně na provázání jednotlivých informací získaných o obsahu dokumentu - na proces strukturalizace dokumentu.

Za druhé, strukturalizační schéma s izolovaným rozhodováním (navržené v sekci 6.1) vychází (stejně jako dříve vytvořená schémata) z postupného izolovaného rozhodování - jednotlivé klasifikační nástroje jsou použity jeden po druhém a jejich závěry jsou okamžitě aplikovány na vytvářený dokument. Oproti tomu, schéma s kumulovaným rozhodováním (sekce 6.2) umožňuje shromáždit dílčí informace a teprve následně provést rozhodnutí (za využití vzájemné verifikace/kombinace jednotlivých informačních zdrojů). Možnost používat k jednotlivým rozhodnutím více (vzájemně souvisejících) informačních zdrojů umožňuje jak snadné zapojení nových informačních zdrojů do procesu strukturalizace dokumentu, tak eliminovat některé chyby generované dílčími nástroji.

Za třetí, využití automaticky časovaných referenčních přepisů (navržené v sekci 7.3) umožňuje časově efektivnější a podrobnější vyhodnocení výsledků systému rozpoznání řeči i dílčích úloh strukturalizace dokumentu. Navržený mechanismus navíc umožňuje vyhodnotit libovolné aspekty strukturalizace dokumentu za využití jedné struktury referenčního dokumentu (není třeba pro každý experiment vytvářet reference se specifickým informačním obsahem). Automatické časování referenčních přepisů je jednou z okolností, které umožnily vytvoření rozsáhlé a různorodé sady testovacích nahrávek.

Za čtvrté, systémy popsané v této práci byly použity ke zpracování vybrané části archivu ČRo, jejíž rozsah převyšuje 100.000 hodin. Zpracované dokumenty pokrývají rozsáhlou časovou periodu (od roku 1926 do současnosti), vyskytují se v nich promluvy ve dvou hlavních jazycích (češtině a slovenštině), nahrávací řetězce jsou velmi různorodé (nejstarší nahrávací technologie, studiové nahrávky, telefonní vstupy) a zpracované pořady jsou nejrůznějších typů (zpravodajství, projevy, diskuze, přímé vstupy).

9.2 Praktické přínosy práce

Strukturalizační schémata popsaná v této práci byla postupně nasazena ke zpracování části archivu Českého rozhlasu (více než 100.000 hodin nahrávek převážně zpravodajského charakteru). 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 (obr. 8.1) a přehrávací rozhraní (obr. 8.2).

Filtry vyhledávacího rozhraní umožňují provádět řadu analýz obsahu archivu (nalezené výsledky lze třídit např. podle období vzniku, jazyka promluvy nebo časové proměnných atributů řeči – např. tempo řeči). Již během vytváření archivu byly vypracovány první studie pozorující vývoj jazyka, slovní zásoby a výslovnosti v čase ([70, 71, 72]). Po zveřejnění archivu ho začali hojně využívat pracovníci Českého rozhlasu i různá výzkumná pracoviště (např. oddělení současné lexikologie a lexikografie ÚJČ AV ČR).

9.3 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 (a výběru) DNN akustického dekodéru (s jehož topologií, aktivační funkcí a složením trénovacích dat probíhá řada experimentů) navržené vyhodnocovací schéma, které využívá časovanou referenci. Tím bychom zajistili maximální přesnost funkce nástrojů, které vyžadují vysokou přesnost detekce řečové aktivity.

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. Toho lze dosáhnout kombinací časové informace (získané fonémovým zarovnávačem) a určením slabik slov (pro které jsme již navrhli prototypový nástroj vycházející z WFST). Takový postup je popsán v [59] a umožnil by například detekovat, na které slabiky (a slova) byl mluvčím kladen důraz.

Třetím krokem, který by mohl snížit počet chyb v přepisech, je větší míra redukce slotů v nahrávce. Lingvistický výzkum ukazuje, že na základě analýzy textového obsahu přepisu lze "svázat" až 40 % slov přítomných v přepisu [73].

Pro optimální zpracování jednotlivých nahrávek by bylo vhodné navrhnout strategii pro rozlišení jednotlivých 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 [29]). 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 *Text, Speech and Dialogue*, 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. 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.

[4] M. Siegler, M. J. Witbrock, S. Slattery, K. Seymore, R. Jones, and A. G. Hauptmann, "Experiments in spoken document retrieval at cmu," 1997.

[5] 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.

[6] 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.

[7] W. Heeren, d. F. Jong, v. d. L. Werff, M. Huijbregts, and R. Ordelman, "Evaluation of spoken document retrieval for historic speech collections," tech. rep., European Language Resources Association (ELRA), 2008.

[8] W. Heeren, R. Ordelman, and F. De Jong, "Affordable access to multimedia by exploiting collateral data," in *Content-Based Multimedia Indexing, 2008. CBMI 2008. International Workshop on*, pp. 542–550, IEEE, 2008.

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

[10] 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.

[11] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks.," in *INTERSPEECH*, vol. 237, p. 240, 2011.

[12] J. Vaněk, *Diskriminativní trénování akustických modelů*. PhD thesis, Západočeská univerzita v Plzni, 2009.

[13] J. R. Novak, "Phonetisaurus: A wfst-driven phoneticizer," *The University of Tokyo, Tokyo Institute of Technology*, pp. 221–222, 2011.

[14] M. A. B. Shaik, D. Rybach, S. Hahn, R. Schlüter, and H. Ney, "Hierarchical hybrid language models for open vocabulary continuous speech recognition using wfst," *Proc. of SAPA*, 2012.

[15] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 5528–5531, May 2011.

[16] M. Fapšo, P. Smrž, P. Schwarz, I. Szöke, M. Schwarz, J. Černockỳ, M. Karafiát, and L. Burget, "Information retrieval from spoken documents," in *Computational Linguistics and Intelligent Text Processing*, pp. 410–416, Springer, 2006.

[17] 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.

[18] S. E. Tranter, D. Reynolds, *et al.*, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[19] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multiresolution cochleagram features for voice activity detection," in *Proc. of Interspeech 2014*, pp. 1534–1538, 2014.

[20] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 4, pp. 705–717, 2015.

[21] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, vol. 8, Virginia, USA, 1998.

[22] L. Valeriano Neri, T. I. Ren, G. D. Cavalcanti, T. I. Jyh, and J. Sijbers, "A combined features approach for speaker segmentation using bic and artificial neural networks," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pp. 4332–4335, IEEE, 2013.

[23] D. Küçük and A. Yazıcı, "A semi-automatic text-based semantic video annotation system for turkish facilitating multilingual retrieval," *Expert Systems with Applications*, vol. 40, no. 9, pp. 3398–3411, 2013.

[24] 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.

[25] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features.," in *INTERSPEECH*, 2002.

[26] J. Echeverry-Correa, J. Ferreiros-López, A. Coucheiro-Limeres, R. Córdoba, and J. Montero, "Topic identification techniques applied to dynamic language model adaptation for automatic speech recognition," *Expert Systems with Applications*, vol. 42, no. 1, pp. 101–112, 2015.

[27] Y. Hu, D. Wu, and A. Nucci, "Pitch-based gender identification with two-stage classification," *Security and Communication Networks*, vol. 5, no. 2, pp. 211–225, 2012.

[28] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] E. Fosler-Lussier and G. Williams, "Not just what, but also when: Guided automatic pronunciation modeling for broadcast news," in *DARPA Broadcast News Workshop*, pp. 171–174, 1999.

[39] J. Nouza, J. Zdansky, P. Cerva, and J. Silovsky, "Challenges in speech processing of slavic languages (case studies in speech recognition of czech and slovak)," in *Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 225–241, Springer, 2010.

[40] P. Cerva, J. Nouza, and J. Silovsky, "Study on cross-lingual adaptation of a czech lvcsr system towards slovak," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pp. 81–87, Springer, 2011.

[41] J. Nouza and M. Boháč, "Using tts for fast prototyping of cross-lingual asr applications," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, pp. 154–162, Springer, 2011.

[42] J. Nouza, P. Cerva, J. Zdansky, and M. Kucharova, "A study on adapting czech automatic speech recognition system to croatian language," in *ELMAR, 2012 Proceedings*, pp. 227–230, IEEE, 2012.

[43] R. Šafařík and J. Nouza, "Methods for rapid development of automatic speech recognition system for russian," in *Electronics, Control, Measurement, Signals and their Application to Mechatronics (ECMSM), 2015 IEEE International Workshop of*, pp. 1–6, June 2015.

[44] J. Nouza, J. Psutka, and J. Uhlír, "Phonetic alphabet for speech recognition of czech," *Radioengineering*, vol. 6, no. 4, pp. 16–20, 1997.

[45] J. R. Novak, N. Minematsu, and K. Hirose, "Wfst-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding," in

10th International Workshop on Finite State Methods and Natural Language Processing, p. 45, 2012.

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

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

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

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

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

[58] 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.

[59] J. Kolář, J. Švec, and J. Psutka, "Automatic punctuation annotation in czech broadcast news speech," in *9th Conference Speech and Computer*, 2004.

[60] C. Cerisara, P. Král, and C. Gardent, "Commas recovery with syntactic features in french and in czech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1413–1416, 2011.

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

[62] 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 International Publishing, 2014.

[63] L. Šeps, "Nanotrans; editor for orthographic and phonetic transcriptions," in *Telecommunications and Signal Processing (TSP), 2013 36th International Conference on*, pp. 479–483, July 2013.

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

[65] 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 Berlin Heidelberg, 2013.

[66] M. Bohac, M. Kucharova, Z. Callejas, J. Nouza, and 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. 2014, no. 1, 2014.

[67] 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.

[68] J. Kolář and L. Lamel, "Development and evaluation of automatic punctuation for french and english speech-to-text," in *INTERSPEECH*, pp. 1376–1379, ISCA, 2012.

[69] 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.

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

[71] 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 *Text, Speech, and Dialogue*, pp. 360–368, Springer, 2013.

[72] 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.

[73] 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.

A Přílohy

A.1 Obsah přiloženého CD

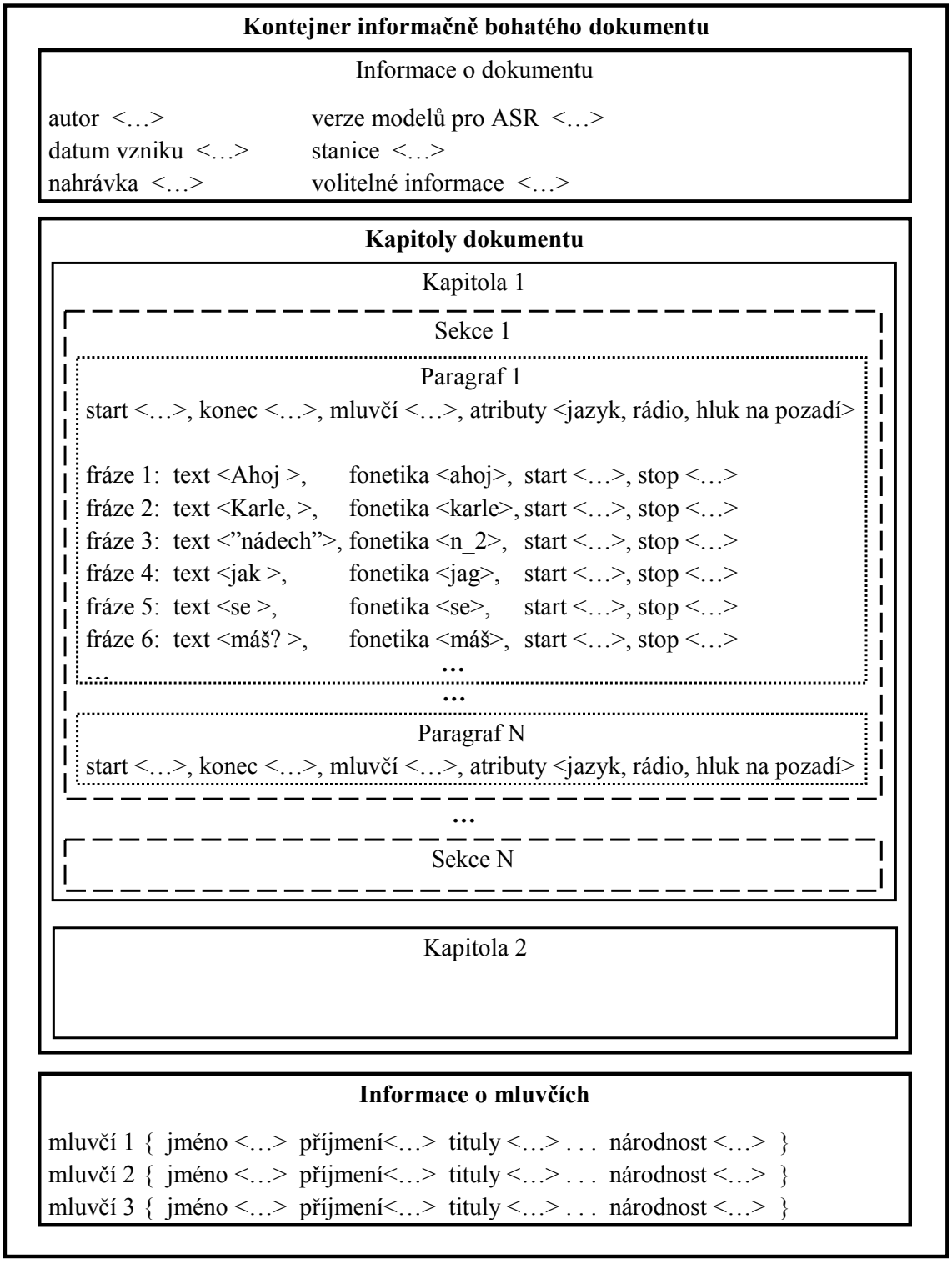
- Text disertační práce

  disertacni_prace_Bohac_2016.pdf

  zdrojové soubory pro LaTeX
- Ukázka referenčních a testovacích dat a výsledků[1]

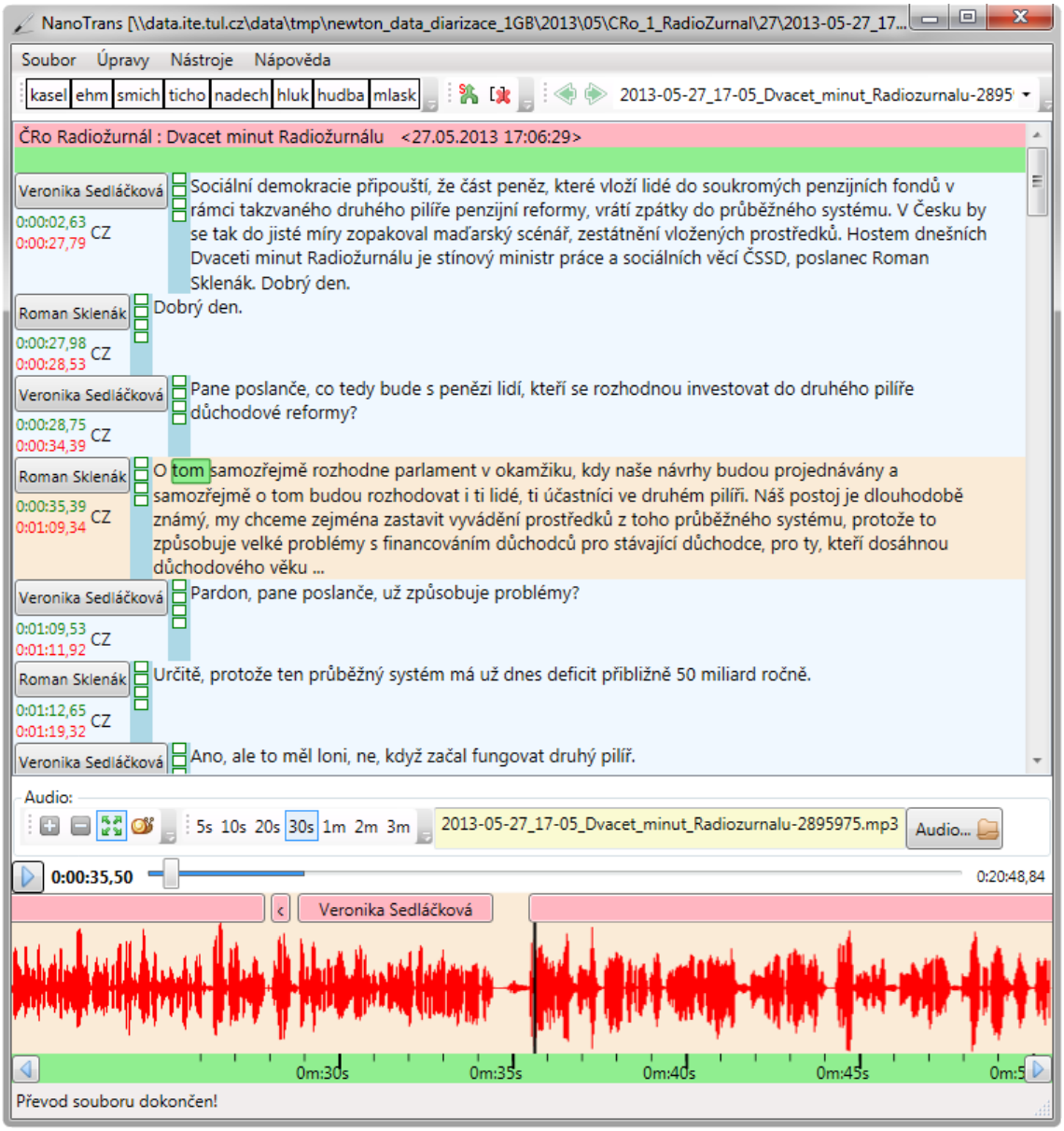
[1] celá testovací sada nemůže být součástí CD, protože její velikost přesahuje 16GB

A.2 Datový kontejner pro strukturalizaci dokumentu



Obrázek A.1: Datový kontejner pro práci s informačně bohatým dokumentem

A.3 Uživatelské rozhraní nástroje NanoTrans



Obrázek A.2: Uživatelské rozhraní anotačního programu NanoTrans

A.4 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.