TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Multilingvální systémy rozpoznávání řeči a jejich efektivní učení

Autoreferát disertační práce

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

Autor práce: **Ing. Radek Šafařík**
Vedoucí práce: prof. Ing. Jan Nouza, CSc.

Liberec 2020

Obsah

Seznam zkratek

ACC	Vyhodnocovací metrika správnosti (Accuracy)
AM	Akustický model (Acoustic Model)
ASR	Automatické rozpoznávání řeči (Automatic Speech Recognition)
BE	Běloruština
BG	Bulharština
BS	Bosenština
CNR	Černohorština
CZ	Čeština
DNN	Hluboké neuronové sítě (Deep Neural Networks)
DOM	Objektový model dokumentu (Document Object Model)
ERR	Vyhodnocovací metrika chybovosti (Error Rate)
FM	Fakulta Mechatroniky, informatiky a mezioborových studií
G2P	Převod grafémů na fonémy (Grapheme to Phoneme)
GMM	Vícemodální gaussovský model (Gaussian Mixture Model)
HTK	Nástroj pro tvorbu skrytých markovských modelů (Hidden Markov Model Toolkit)
HTML	Značkovací jazyk por webové stránky (Hypertext Markup Language)
HR	Chorvatština
IPA	Mezinárodní fonetická abeceda (International Phonetic Alphabet)
LID	Identifikace jazyka (Language Identification)
LM	Jazykový model (Language Model)
LVCSR	Rozpoznávání spojité řeči s velkou slovní zásobou (Large Vocabulary Continuous Speech Recognition)
MFCC	Kepstrální příznaky řeči (Mel-Frequency Cepstral Coefficients)
MK	Makedonština
ML-ASR	Multilingvální systém rozpoznávání řeči
OOL	Části promluvy v cizím jazyce (Out of Language)
OOV	Slova mimo slovník (Out of Vocabulary)
PL	Polština
PREC	Vyhodnocovací metrika přesnosti (Precision)
REC	Vyhodnocovací metrika senzitivity (Recall)
ReLU	Aktivační funkce neuronových sítí (Rectified Linear Unit)
RU	Ruština
SK	Slovenština
SL	Slovinština
SR	Srbština
TUL	Technická univerzita v Liberci
UK	Ukrajinština
WER	Metrika vyhodnocování přesnosti rozpoznávání (Word Error Rate)
XML	Rozšiřitelný značkovací jazyk (Extensible Markup Language)
YLD	Vyhodnocovací metrika výtěžnosti (Yield)

Úvod

Systémy automatického rozpoznávání řeči (angl. Automatic Speech Recognition systems, ASR) slouží k převodu signálu mluvené řeči do podoby vhodné pro další zpracování počítačovými programy. V případě diktovacích, přepisovacích či překladových systémů má výstup podobu textu, ale například u různých hlasově ovládaných aplikací to mohou být interní symboly či příkazy, které jsou pak převáděny na příslušné akce.

Počátky výzkumu v této oblasti sahají do 60. let 20. století a jsou spojeny s rozvojem prvních výkonnějších počítačů. Během dalších dvou dekád došlo k rychlému rozvoji metod efektivní reprezentace řečového signálu pomocí spektrálních (a později kepstrálních) příznaků. Slova se podařilo dekomponovat do malého počtu stavebních jednotek (odvozených od hlásek) a ty reprezentovat pomocí matematických modelů (nejčastěji to byly skryté Markovovy modely) a celé věty pak poskládat na základě pravděpodobnostních přístupů založených nejčastěji na n-gramových modelech [1][2][3].

Díky tomu bylo možné již na začátku 90. let představit první komerční programy určené zejména pro diktování do počítače. Ty ještě spoléhaly na vstřícný přístup ze strany uživatele. Avšak s rozvojem dalších robustnějších metod se použití rozšířilo i na oblast automatického přepisu televizních a rozhlasových zpráv, přepis jednání (např. v parlamentu) a posléze i na méně kvalitní záznamy např. telefonních hovorů [2][3].

Ve stejné době se objevily i první dialogové systémy, v nichž byla, kromě rozpoznávání, použita i hlasová syntéza. V současné době se se systémy rozpoznávání řeči setkáváme v mnoha mobilních aplikacích, třeba v rámci hlasového vyhledávání (např. VoiceSearch od Googlu), u hlasových asistentek (např. Siri od firmy Apple) nebo konverzačních a chatovacích programech (např. Alexa od Amazonu). Většina těchto aplikací výrazně šetří čas uživatele, neboť hlasová interakce je mnohem rychlejší a přirozenější než práce s klávesnicí, myší a obrazovkou. Pro osoby s některými typy tělesného postižení se navíc jedná o jedinou možnost, jak používat moderní techniku [2][3].

Automatické zpracování mluvené (i textové) podoby jazyka má ale jednu specifickou vlastnost - je závislé právě na daném jazyku: na jeho písmu a kódování, hláskovém inventáři, na slovníku a výslovnosti, na syntaxi a gramatice, a v nepo-

slední řadě i na společenském a historickém kontextu. Z těchto důvodů byly první systémy rozpoznávání řeči vyvíjeny vždy pro konkrétní jazyk, čemuž se přizpůsobovaly i použité metody a přístupy. Kromě techniků a programátorů byli součástí výzkumných týmů také experti na lingvistiku a fonetiku. Takový výzkum a vývoj byl finančně náročný a v začátcích si ho mohly dovolit pouze velké firmy (např. IBM či Microsoft), či významné akademické instituce, a většinou se soustředil pouze na velké světové jazyky (zejména angličtinu, francouzštinu, španělštinu, japonštinu, apod.)[3].

Teprve později se podařilo lépe vymezit a oddělit části systému závislé na konkrétním jazyku od těch nezávislých, což následně umožnilo efektivnější přenos poznatků a algoritmů (a později dokonce i natrénovaných modelů) do dalších jazyků. Přesto i dodnes platí, že každý jazyk má své specifické vlastnosti, které ovlivňují např. nezbytnou velikost slovníku, převod mezi psanou a vyslovovanou formou slov, vazby mezi větnými členy, formátování, apod.

Je ovšem také pravda, že současné informační technologie a zejména existence internetu s obrovským množstvím veřejně přístupných (textových a mluvených) dat, umožňují, aby se jazykově závislé moduly učily přímo z těchto dat. Moderní metody strojového učení tak do velké míry umožňují nahradit práci jazykových expertů a významným způsobem zkrátit dobu vývoje systému určeného pro konkrétní jazyk. Což zároveň znamená, že lze těmito technologiemi velmi rychle pokrýt i menší jazyky.

Zaměření práce

Tato práce se zabývá výzkumem a implementací metod, které umožňují rychlý vývoj jazykově závislých modulů pro systémy automatického rozpoznávání řeči. Vznikla na pracovišti (Laboratoř počítačového zpracování řeči na Technické univerzitě v Liberci[1]), kde již od poloviny 90. let pracuje tým, který se touto problematikou zabývá. Během dvou desítek let zde byly vytvořeny všechny základní moduly sloužící pro sestavení a provozování systému ASR, který může být nasazen jak v on-line, tak i off-line režimu a hodí se pro zpracování dat buď ze souboru na disku, nebo přímo z mikrofonu či dokonce z internetového streamu (např. televizního či rozhlasového vysílání).

V době, kdy jsem do týmu přišel, byl již systém schopen pracovat s mluvenou češtinou a slovenštinou, a probíhaly práce na zvládnutí polštiny a chorvatštiny. Zatímco vývoj češtiny (a samozřejmě celého rozpoznávacího řetězce) trval více než 10 let, slovenštinu se podařilo zpracovat cca za 3 roky, a u dalších jazyků se už vývoj základní verze pohyboval kolem jednoho roku. V té době byl vytyčen cíl zvládnout během několika let všechny slovanské jazyky s tím, že vývoj každého z nich by neměl trvat více než několik měsíců.

[1] https://www.ite.tul.cz/speechlab/

Bylo proto nutné seznámit se se všemi těmito jazyky, najít jejich společné a zároveň i odlišné rysy, sestavit pravidla pro vytváření textových korpusů, slovníků a jazykových modelů, vytvořit převodníky mezi ortografickou a fonetickou podobou slov, navrhnout a implementovat metody pro automatické získávání trénovacích dat, a to s různou mírou supervize, a v neposlední míře také vytvořit prostředí pro objektivní testování vyvinutých modulů.

Součástí mé práce byl proto návrh, ověřování a základní implementace všech výše uvedených postupů, jakož i tvorba mnoha pomocných nástrojů, bez nichž by se výzkum a vývoj neobešel, ať už se jednalo například o programy pro hromadné stahování textových a akustických dat, nástroje na jejich analýzu a automatické zpracování, tvorbu a optimalizaci fonetických inventářů a výslovnostních generátorů, moduly pro zpracování čísel a zkratek, až po finální natrénování akustických a jazykových modelů pro každý jazyk.

Zároveň je však třeba říci, že jsem se nemusel zabývat vývojem těch částí rozpoznávacího systému, které jsou jazykově nezávislé. Měl jsem tedy k dispozici již hotové moduly zpracovávající akustický signál a transformující ho na příznakové vektory a dále pak velmi efektivně pracující dekodér převádějící sekvence příznakových vektorů na textový výstup. Tyto klíčové části systému navržené jinými členy týmu jsem tak mohl využívat pro svou práci, což ji výrazně urychlilo, na druhou jsem je musel používat v takové podobě, v jaké byly naimplementovány, bez možnosti do nich zasahovat, což někdy určovalo volbu mých metod a přístupů.

Motivace výzkumu a vazba na praxi

Moje práce byla součástí dvou velkých výzkumných projektů řešených na pracovišti. Oba byly financovány Technologickou agenturou České republiky. Jednalo se o tyto projekty:

- TA04010199 „MULTILINMEDIA - Multilingvální platforma pro monitoring a analýzu multimédií“ (2015-2017),
- TH03010018 „DeepSpot - Multilingvální technologie pro detekci a včasné upozornění“ (2018-2021).

Hlavním cílem obou projektů bylo zvládnout přepis a následnou analýzu televizních, rozhlasových a internetových pořadů ve 13 slovanských jazycích, a to češtiny, slovenštiny, polštiny, ruštiny, ukrajinštiny, běloruštiny, slovinštiny, chorvatštiny, srbštiny, bosenštiny, černohorštiny, makedonštiny a bulharštiny. V současné době je již většina z nich předána partnerovi projektu, firmě Newton technologies, a.s., která využívá rozpoznávací systém a jeho jazykové moduly v rámci on-line monitoringu několika desítek stanic provozovaných v těchto zemích.

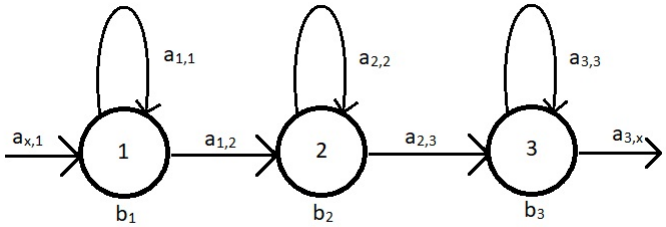
1 Interdisciplinární základy a současný stav poznání

Výzkum a vývoj v oblasti počítačového zpracování řeči má víceoborový charakter. Kromě technických a přírodovědných disciplín, jako jsou akustika, zpracování signálů, matematické modelování, teorie rozhodování, či strojové učení, hrají významnou roli také poznatky ze společensko-vědních oborů, zejména lingvistiky a fonetiky. U systémů, které mají pracovat ve vícejazyčném prostředí, je tato role ještě mnohem důležitější. V této kapitole proto budou krátce představeny základní poznatky a terminologie užité v této práci společně se současným stavem poznání.

1.1 Textová a zvuková podoba jazyka

U každého jazyka rozlišujeme podle [4],[5] a [6] mluvenou a textovou podobu. Základní významovou jednotkou psané formy je *slovo*. Slova sestavená do *vět* pak vytvářejí sdělení. Slova jsou zapisována pomocí *znaků* dané *abecedy* (budeme je označovat též jako *grafémy*), které dávají slovu jeho *ortografickou* podobu. V ohebných jazycích jsou slova modifikována pomocí *morfologických* pravidel, která ze základní formy (nazývané *lemma*) vytvářejí odvozené slovní formy. Sestavování smysluplných a srozumitelných vět ze slov se řídí pravidly *gramatiky* daného jazyka. V současné době velkého rozmachu informačních technologií je důležitým nástrojem pro analýzu a zpracování textů v daném jazyce textový *korpus*, tedy velmi rozsáhlý a dostatečně reprezentativní soubor textů. Jeho statistickým zpracováním lze sestavit seznam nejčastěji používaných slov a vytvořit tak reprezentativní *slovník* (nazývaný též *lexikon*), s kterým může být dosaženo požadované úrovně pokrytí psaných (a do velké míry též mluvených) textů. Statistickými nástroji lze vyjádřit též vztahy mezi slovy ve větách a to na základě četnosti jejich výskytů v rámci za sebou jdoucích slovních sekvencí. Takto popsaný mezislovní kontext se označuje jako *jazykový model*.

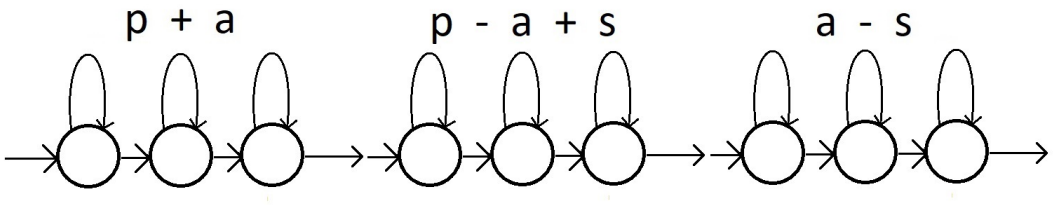
Podobně jako v psané formě, i u mluvené podoby je dle [7] a [8] základní významovou jednotkou slovo. To je sestaveno z *hlásek* daného jazyka (v této práci též nazývané *fonémy*). Hlásky jsou zvuky tvořené činností vokálního traktu a dělí se na *samohlásky* (charakterizované nepřerušovaným proudem zvuku vytvořeného hlasivkami) a *souhlásky*, jejichž charakter je časové proměnný, obsahuje šumovou složku a je ovlivněn překážkami v hlasovém traktu. Charakter jednotlivých hlásek se dále liší v závislosti na okolním kontextu, kdy např. hláska /*a*/ zní jinak (a má



Obrázek 1.1: Třístavový model hlásky

tudíž i jiný spektrální průběh), vyskytuje-li se před ní (či za ní) sykavka, nosovka či explozíva. Fonetici tyto kontextové varianty označují jako *alofony*. Při počítačovém zpracování se tato variabilita řeší tím, že modely hlásek bývají vícestavové (nejčastěji třístavové) a jednotlivé varianty se modelují a trénují zvlášť jako tzv. *trifony*, tj. kontextově závislé modely s různým pravým a levým okolím [1][9].

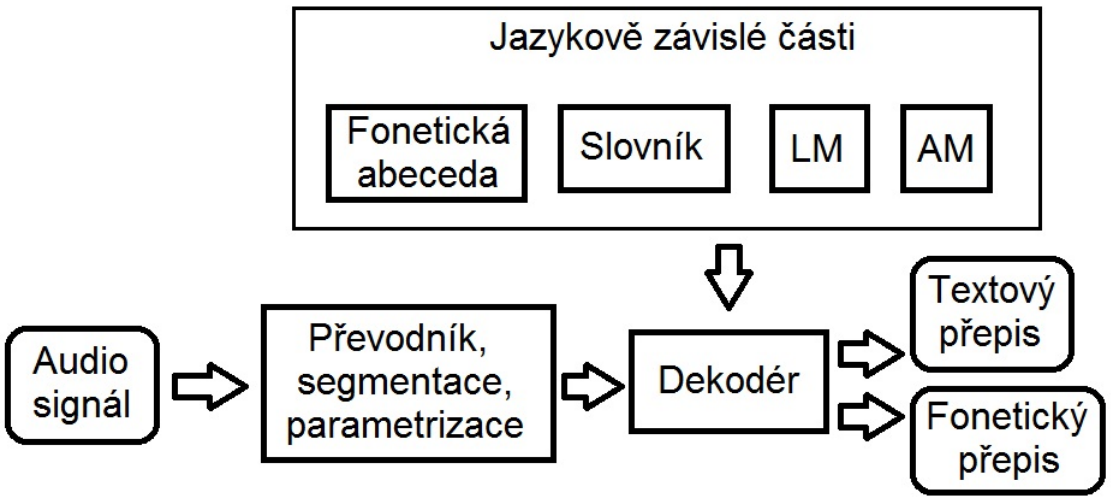
Zvuková podoba řeči má mnohem větší míru variability než textová. Kromě již zmíněné kontextové variability hlásek hraje (dokonce ještě větší) roli osoba řečníka a způsob, jakým mluví a jak vyslovuje. Jediným způsobem, jak alespoň částečně eliminovat vliv těchto faktorů na systém rozpoznávání řeči, je vytvořit dostatečně robustní *akustický model* všech hlásek a jejich variant natrénovaný na velmi rozsáhlém souboru nahrávek pořízených od tisíců různých osob a v různých situacích a akustických podmínkách. V posledních dvou dekádách se nejčastěji používaným typem stal skrytý Markovův model (Hidden Markov Model, HMM). U hlásek má nejčastěji třístavovou levo-pravou strukturu (viz obrázek 1.1) se dvěma typy parametrů: a) přechodovými pravděpodobnostmi mezi stavy a b) stavovými výstupními pravděpodobnostmi. Výstupní pravděpodobnosti bývají reprezentovány buď směsmi gaussovských rozložení (Gaussian Mixture Model, GMM) nebo, v současnosti mnohem častěji hlubokými neuronovými sítěmi (Deep Neural Networks, DNN). Mluvíme pak o akustickém modelu typu GMM-HMM nebo DNN-HMM. HMM libovolného slova se sestaví jednoduchým zřetězením příslušných hláskových (trifonových) modelů (viz obrázek 1.2) [1][9].



Obrázek 1.2: Trifónový model slova ”pas”

1.2 Modulární systém rozpoznávání řeči

Systém rozpoznávání řeči v klasické podobě (a též v podobě používané v této práci) je vyobrazen na obrázku 1.3. Na jeho vstup přichází zvukový signál a na výstupu se s určitým zpožděním objevuje přepsaný text.



Obrázek 1.3: Schéma modulárního systému rozpoznání řeči

Systém se skládá z několika modulů. První z nich segmentuje signál do tzv. *rámců* (budeme používat i anglický výraz *frame*), dlouhých obvykle 25 ms, a v každém z nich je vypočteno P *spektrálních příznaků*. Ty jsou buď přímo použity při dekódování (v DNN-HMM systému) nebo jsou ještě převedeny na *kepstrální příznaky* (v GMM-HMM systému).

Dekodér postupně zpracovává sekvence těchto příznakových vektorů tak, že počítá pravděpodobnosti, s jakými by byly vygenerovány jednotlivými hláskovými variantami reprezentovanými akustickým modelem. Na základě toho pak průběžně počítá pravděpodobnosti, že se v blízkém rozmezí framů objevuje některé slovo ze slovníku. Zároveň vyhledává nejpravděpodobnější sekvence slov jdoucích za sebou, a to na základě kombinace skóre vzešlého z akustického a jazykového modelu.

Úkolem dekodéru je najít takovou sekvenci slov W, která má největší pravděpodobnost $P(W|X)$ pro pozorovanou sekvenci příznaků X a dané modely. S využitím Bayesovy věty je výpočet pravděpodobnosti definován jako:

$$P(W|X) = \frac{p(X|W) \cdot P(W)}{p(X)} \tag{1.1}$$

kde $p(X)$ je apriorní pravděpodobnost pozorování sekvence příznakových vektorů X a $p(X|W)$ je pravděpodobnost, že pro danou sekvenci slov W bude pozorován příznakový vektor X. Posledně zmíněný člen je nazýván *akustický model*. $P(W)$ je

poté apriorní pravděpodobnost pozorování sekvence slov W nezávisle na příznakovém vektoru X a je nazývána *jazykový model*.

Dekodér se pak snaží nalézt:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} \frac{p(X|W) \cdot P(W)}{p(X)} \tag{1.2}$$

Prostor všech možných slov, ze kterých se může W skládat, je definován *slovníkem*. Ten také obsahuje všechny přípustné výslovností varianty pro každé slovo.

Modul zpracování signálu a dekodér tvoří jazykově nezávislou část rozpoznávacího systému. Slovník (doplněný o výslovnosti všech slov), seznam přípustných hlásek, akustický model a jazykový model jsou naopak ty části, které jsou jazykově závislé a pro každý jazyk musí být předem vytvořeny či natrénovány. Způsob efektivní tvorby těchto jazykově závislých modulů je hlavním tématem této práce.

Tabulka 1.1 zobrazuje základní parametry rozpoznávacího systému vyvinutého na pracovišti a použitého v rámci této práce.

Tabulka 1.1: Parametry použitého systému rozpoznávání řeči

Vzorkovací frekvence	16 kHz
Délka rámce	25 ms
Rámcová frekvence	100 hz
GMM příznaky	39 dim. MFCC
GMM počet komponent	32
DNN příznaky	39 dim. Log filter banks
DNN architektura	dopředná pětivrstvá síť
DNN počty neuronů	1024-768-768-512-512
DNN aktivační funkce	ReLU

1.3 Metriky používané pro vyhodnocování přesnosti rozpoznávání

K vyhodnocení přesnosti rozpoznávání je v této práci používána míra *WER* (Word Error Rate), která využívá metody hledání minimální vzdálenosti na úrovni slov pro zjištění počtu operací, potřebných ke změně jednoho řetězce na druhý. Následně je hodnota WER vypočtena pomocí vzorce:

$$WER = \frac{S + D + I}{N} \tag{1.3}$$

Hodnoty S, I a D označují počet záměn (substitucí), vložených slov (inzercí) a odstraněných/nerozpoznaných slov (delecí). N označuje celkový počet slov v refe-

renčním textu. Výsledná hodnota udává slovní rozdíl obou textů v procentech. Ze vztahu 1.3 také vyplývá, že hodnota WER může nabývat hodnot větších než 100 % z důvodu relativně neomezeného množství možných inzercí.

Další důležitou metrikou je míra slov mimo slovník *OOV* (Out of vocabulary). Ta udává (opět v procentech) počet slov v referenčního textu, která nejsou obsažena ve slovníku rozpoznávacího systému, a tudíž nemohou být správně rozpoznána.

V této práci zavádíme ještě parametr označovaný jako *OOL* (Out of language), který udává míru zastoupení řeči v jiném než cílovém jazyce. Tato metrika je vhodným doplňkem k OOV při analýze úspěšnosti reálných nahrávek, např. u zpravodajských pořadů různých stanic.

1.4 Současný stav v dané oblasti a existující řešení

Pro efektivní vývoj ASR systémů pracujících v multilingválním prostředí je třeba vyřešit řadu dílčích úloh, od získání dostatečného množství textových a akustických dat, přes tvorbu slovníků, definování fonémového inventáře a vygenerování výslovnosti pro každé slovo, vytvoření fonetického přepisu ke každé nahrávce, natrénování akustického a jazykového modelu, až po vytvoření prostředí pro objektivní testování hotového systému.

Výzkumné (akademické i firemní) týmy se zpočátku věnovaly hlavně vývoji systémů určených pro svůj vlastní jazyk. Po jeho zvládnutí se pak některé z nich pustily i do dalších jazyků, buď příbuzných, nebo takových, které nabízely významný komerční potenciál. V současné době hraje v této oblasti největší roli společnost Google, která má ve svém portfoliu ASR systémů většinu světových jazyků.

Mezi akademickými institucemi patří k významným týmům tohoto směru např. francouzský tým LIMSI (soustředěný kolem Jean-Luca Gauvaina a Lori Lamel), který již na konci 20. století začal vytvářet jazykové mutace svých systémů pro různé jazyky, a to jak světové [10], tak i tzv. *jazyky s minimálními zdroji* (under-resourced či low-resourced languages). Dalším významným pracovištěm je KIT na německé Karlsruhe Universität spojený zejména se jménem Tanji Schultz. Nejznámějším příspěvkem jejího týmu je vytvoření standardizované databáze nahrávek v mnoha jazycích nazvané Globalphone [11].

Ve výzkumu rozpoznávání řeči sehrály významnou roli veřejně přístupné platformy navržené pro experimentální vývoj v této oblasti. Od 90. let 20. století byl v této oblasti nejznámější produkt Cambridgské univerzity známý pod zkratkou HTK (Hidden Markov Toolkit) [12]. V poslední dekádě převzal vedoucí roli v této oblasti systém Kaldi [13], který jako první umožňoval použití neuronových sítí. Obě programové platformy byly a jsou často využívány pro experimentování s multilin-

gválními systémy.

1.4.1 Textové korpusy a jazykové modely

Textový korpus je základem pro tvorbu slovníku a jazykového modelu. Je potřeba, aby byl tvořen rozmanitými texty v dostatečném množství (řádově stovky MB až jednotky GB), které dostatečně zastupují rozpoznávaný jazyk, případně konkrétní oblast určenou k rozpoznávání. Existuje mnoho textových korpusů volně nebo komerčně dostupných. Jedním z nejznámějších distributorů je asociace ELRA[1] (Europen Language Resources Association). Další možností jsou národní korpusy jednotlivých zemí, jako je např. Český národní korpus[2] [14].

Alternativní možností je tvorba vlastních korpusů přizpůsobených konkrétním potřebám vyvíjeného systému. Přístupy k vytěžování volně dostupných online zdrojů jsou popsány například v [15] nebo [16], kde byl vytvořen univerzální přístup a nástroje pro automatické stahování textů z webových stránek a jejich zpracování. Naproti tomu postup popsaný v této práci není až tak obecný, zaměřuje se na přesnost a využívá multingválních závislostí mezi jazyky pro zefektivnění celého procesu.

Jazykový model poskytuje odhad pravděpodobnosti sekvence rozpoznaných slov. Nejčastější přístup je pomocí statistického modelování za využití slovních n-gramů (bigramy, trigramy a další). Pravděpodobnosti v n-gramových jazykových modelech jsou běžně určovány pomocí maximálního odhadu věrohodnosti. To činí rozložení pravděpodobnosti závislé na trénovacích datech, a proto je vyžadováno co největší množství těchto dat. Jazykový model je pak tedy vypočítán z textového korpusu pro všechna slova ve slovníku.

V rámci této práce jsem byl odkázán na použitý ASR systém, který pracuje pouze se statistickými bigramovými modely. Ty sice mají horší výsledky oproti trigramovým modelům, nicméně při práci se slovanskými či jinými flektivními jazyky, které mají relativně volný slovosled a velký slovník, není rozdíl tak markantní, jak ukázala interní studie. Naopak využití bigramů snižuje výpočetní náročnost a umožňuje tak systému pracovat v reálném čase. Pro modelování delších mezislovních kontextů mohou být do slovníku přidávány kolokace (častá slovní spojení), čímž se do jisté míry supluje vliv vyššího n-gramového modelu.

1.4.2 Fonetická stránka jazyka a výslovnostní slovníky

Pro každý jazyk existují fonologické studie rozlišující jednotlivé fonémy daného jazyka a případně i jeho různých dialektů. V počátcích vývoje ASR systémů, které měly k dispozici malé množství akustických dat pro trénování, se vyplatilo pracovat s většími počty fonetických jednotek definovanými i na základě okolního kontextu.

[1] http://www.elra.info/
[2] https://www.korpus.cz/

To samozřejmě vedlo k využívání velkých fonetických sad a komplikacím na různých úrovních vývoje, jako například velká nevyváženost trénovacích dat pro jednotlivé modely fonémů či problémy při fonetické anotaci.

S příchodem nových technologií, metod a množství trénovacích dat bylo možno přejít od kontextově nezávislých jednotek (tzv. monofónů) na trifónové modely (případně vyšší), které už v sobě zahrnují i vliv levého a pravého okolí jednotlivých hlásek. To vedlo ke zmenšení fonetické sady a zefektivnění celého systému [17].

Výslovnostní slovník je nezbytná součást rozpoznávacího systému. Ten zaprvé obsahuje množinu všech slov, které mohou být systémem rozpoznány a za druhé tvoří spojení mezi lexikální a akustickou částí systému. Slova do slovníku jsou vybírána z textového korpusu, na kterém je následně trénován jazykový model. Nejčastěji jsou vybírána podle četnosti v korpusu, případně mohou být další důležitá slova dodána ručně. Velikost slovníku závisí na cílovém jazyku. U slovanských jazyků, které jsou vysoce flektivní, se ukazuje jako nezbytné množství slov v řádu stovek tisíc slov.

Pro tvorbu výslovností pro jednotlivá slova ve slovníku jsou využity různé metody tzv. G2P (Grapheme-to-Phoneme) konverze, které mohou využívat přesně daná produkční pravidla, mohou být reprezentovány stavovými automaty, nebo využívat metod strojového učení, například na základě neuronových sítí. Způsob využívající předem daná produkční pravidla vyžaduje určitou znalost fonetiky daného jazyka a ruční přípravu pravidel. Nicméně u ortograficky mělkých jazyků, kam patří i slovanské jazyky, může být tvorba pravidel poměrně snadná. Navíc zde lze využít určitých obecných principů fonetiky, které jsou často sdíleny napříč jazyky [9].

Další přístupy využívají pro fonetickou transkripci metod strojového učení, kdy se systém na určité množině slov s jejich fonetickými přepisy sám naučí vztahy mezi ortografickou a fonetickou podobou. Dnes jsou nejčastěji využívány neuronové sítě různých typů, např. LSTM rekurentní neuronoé sítě [18], či LSTM sítě v kombinaci s konvolučními vrstvami [19]. Tento přístup nicméně vyžaduje již nějaké množství výchozích dat pro trénování a nemůže tak být využit v počáteční fázi, kdy žádná data k dispozici nejsou.

1.4.3 Akustický model a trénovací databáze

Zatímco obstarat textová data je v dnešní době internetu jednoduché, s vhodnými akustickými daty je situace složitější. Pro vytvoření akustického modelu je potřeba alespoň několika hodin nahrávek řeči společně s jejich co nejpřesnějšími fonetickými přepisy. Řada subjektů dnes nabízí hotové řečové korpusy, ale za nemalou cenu. Příkladem je již zmíněný tým na Karlsruhe Universität a jejich databáze *GlobalPhone* či asociace ELRA. Existují i crowd-sourcingové projekty jako *Amazon Mechanical Turk* [20], kde jsou řečové nahrávky připraveny na serveru a lidé mohou vytvářet přepisy těchto nahrávek za určitý finanční obnos. Dále existují projekty jako na-

příklad VoxForge[3], kde dobrovolníci nahrávají krátké věty v různých jazycích, či projekt Librivox[4], kde lidé předčítají knihy v mnoha různých jazycích.

Další možností je vytvoření vlastního řečového korpusu a to buď nahráváním rodilých mluvčích, což vyžaduje hodně úsilí a zdrojů, nebo automatickým zpracováním volně dostupných dat na internetu. Jako taková data mohou být využity např. audioknihy, pořady s titulky, případně přepisy z jednání parlamentu a podobně. V tomto případě je využito lehce supervizované učení spočívající v aplikaci existujícího ASR systému k zarovnání textu k nahrávce, včetně detekce případných neshod, které jsou následně odstraněny. Takto získaná data jsou využita k trénování nového systému. Tyto postupy jsou iterativní, kdy v každé iteraci jsou vytvářena nová trénovací data, z nich je natrénován nový, vylepšený model, který je následně použit v dalším kroku. Takový způsob byl uplatněn například v [21] pro získání trénovacích dat z anglických audioknih, v [22] pro čínský ASR systém těžící z televizního vysílání s titulky nebo pro jihoafrickou angličtinu za využití pořadů rádiového vysílání s přesnými přepisy [23].

K tvorbě trénovacích dat může být využito i nesupervizovaného přístupu, kdy je pro audio nahrávky vytvořen fonetický přepis za pomoci již existujícího systému či více systémů. Takový způsob byl například aplikován při vývoji polského systému pro zpracování záznamů Evropského parlamentu [24], kde je využita důvěryhodnost (confidence measure) vypočtená dekodérem a stanovený práh, který tato hodnota musí překročit, aby byl přepis považován za přesný.

Tato práce se zaměřuje na získávání trénovacích dat ze zdrojů, které obsahují audio nahrávky a nějaký doprovodný text. Co je v nahrávce obsaženo, není úplně předem známo a o textu se také neví, jestli obsahuje úplný nebo částečný přepis promluvy v nahrávce, či nějakým způsobem promluvu v nahrávce parafrázuje, nebo obsahuje jen popis toho, co se v nahrávce děje. Příkladem jsou zpravodajské weby, kde jednotlivé zprávy či články obsahují audio či video a k němu přidružený text, který popisuje situaci a případně cituje části promluvy. Naším cílem je získat co největší množství takovýchto volně dostupných dat, pokusit se najít alespoň nějaké shodující se části a vytěžit z nich trénovací data.

Ve chvíli, kdy jsou k dispozici data pro více jazyků, je možno je využít k tvorbě multilinguálních systémů. Ty pak mohou být využity k různým účelům a to například k identifikaci jazyka, či k získávání trénovacích dat pro nový jazyk a následnému urychlení vývoje.

[3] http://voxforge.org
[4] htttps://librivox.org

2 Cíle práce

Hlavním tématem předkládané práce je výzkum a vývoj zaměřený na multilingvální systémy automatického rozpoznávání řeči (ML-ASR), a to zejména na jejich jazykově závislou část zahrnující lingvisticko-lexikální moduly (slovníky a jazykové modely) a akusticko-fonetické moduly (fonémové inventáře, výslovnosti a akustické modely). Práce úzce souvisí s projekty řešenými na školícím pracovišti, což ovlivnilo i stanovení cílů a priorit. Ty lze definovat takto:

- Navrhnout co nejefektivnější přístup k vývoji výše uvedených jazykově závislých modulů, který bude po svém nasazení vyžadovat minimum lidské práce, a to jak expertní (zejména v oblasti lingvistické), tak i manuální (například ve formě přepisů, sluchových kontrol či anotací), a který si vystačí s daty veřejně přístupnými prostřednictvím internetu.

- Navrhnout a implementovat kompletní sadu nástrojů, které pomohou zautomatizovat většinu nezbytných prací a úkonů, počínaje sběrem textových a akustických dat, přes tvorbu výslovnostních slovníků, jazykových a akustických modelů, až po finální podobu automaticky přepsaných textů.

- Prozkoumat a navrhnout možnosti nasazení metod strojového učení zejména v nejkritičtější části vývoje, kterou je získávání a anotace dat pro trénování akustických modelů pro jednotlivé jazyky. Zde se zaměřit především na využití tzv. lehce supervizovaného přístupu, v němž hraje úlohu supervizora vlastní vyvíjený ML-ASR systém.

- Navržené postupy aplikovat na všechny slovanské jazyky, a to s využitím jejich podobných rysů a metod založených na mezijazykovém transferu a adaptaci.

- Při vývoji a získávání dat pro učení se zaměřit na cílovou doménu budoucích aplikací, kterou bude (v souladu se zmíněnými projekty) zejména automatický přepis a monitoring televizních a rozhlasových stanic vysílajících v 13 národních slovanských jazycích. Pro tuto oblast také vytvořit sadu reálných testovacích dat použitelných pro objektivní vyhodnocení přesnosti přepisu a porovnání různých přístupů.

- Ověřit použitelnost navržených metod a postupů na několika dalších vybraných evropských neslovanských jazycích.

3 Lingvisticko-lexikální část multilingválního systému rozpoznávání řeči

Tato a následující kapitola popisují výzkum a vývoj metod, které byly použity při tvorbě systémů rozpoznávání řeči pro slovanské a následně i další jazyky. Popis je pro přehlednost rozdělen do dvou kapitol. Tato kapitola se zabývá tvorbou textového korpusu, slovníku a jazykového modelu. Součástí slovníku jsou ale také výslovnosti jednotlivých slov, což bude řešeno v následující kapitole, zabývající se akustickofonetickou částí ASR systémů.

3.1 Tvorba korpusu pro daný jazyk

Textový korpus sestává z velkého množství textů daného jazyka a následně je z něho vytvářen slovník a jazykový model. Kromě využívání již hotových korpusů může být takový korpus vytvořen shromážděním a zpracováním volně dostupných dat z internetu.

Pro tvorbu ideálního textového korpusu je zapotřebí nasbírat dostatečné množství textů z různorodých zdrojů, aby tak pokrýval co největší část zpracovávaného jazyka. Řádově se jedná minimálně o stovky MB, nejlépe jednotky GB. Po získání dostatečného množství dat je potřeba texty zpracovat do použitelné podoby, tj. normalizovat, vyfiltrovat nežádoucí elementy a naformátovat pro další použití.

V první řadě je potřeba najít vhodné a dostatečně objemné zdroje dat. Nejvýznamnějším zdrojem textových dat jsou většinou webové stránky novinových zpravodajských portálů či rozhlasu a televize, které mají velký rozsah témat a pokryjí tak velkou část slovní zásoby, ale především obsahují velké množství dat v podobě článků přidávaných každý den. V případě práce s cizími jazyky, kdy se bez znalosti prostředí těžko hledají vhodné weby, jsou dobrým zdrojem například speciální portály shromažďující odkazy na zpravodajské weby jako např. ABYZ News Links[1].

Texty z vybraných webů jsou hromadně staženy. K tomuto účelu existuje množství nástrojů, nicméně pro potřeby této práce byl vytvořen vlastní stahovací nástroj, který umožňuje přesně nastavovat všechny potřebné parametry pro optimální získávání vhodných dat pro tvorbu korpusu.

[1] http://www.abyznewslinks.com/

Stažené texty je potřeba nejprve zpracovat a normalizovat do jednotné podoby. Každý web může používat různé druhy kódování, různé typy zápisu znaků s diakritikou, interpunkce či dalších speciálních znaků týkajících se daného jazyka. Dále je potřeba odstranit texty, které nesplňují určitá kritéria jako je např. délka řetězce, počet slov, velký znak na začátku či interpunkční znaménko na konci. Tato kritéria z velké části zaručí, že jsou nakonec použity jen reálné věty a všechny nadpisy, popisky a další nežádoucí texty jsou odfiltrovány. Jako optimální délka textu byla ve většině případů vybrána hranice 10 znaků a alespoň tří až pěti slov (tokenů oddělených mezerou). Potřeba je z textů odstranit webové a e-mailové adresy, které by mohli být dekomponovány na jednotlivé části a zanášely by tak textový korpus. Na závěr je vhodné odstranit časté duplicity (např. věty opakující se v každém článku jako prohlášení a podobně), které by pak mohly významně zkreslovat slovní statistiku.

Tabulka 3.1: Rozdíly v abecedách východoslovanských jazyků

	RU	**UK**	**BE**
RU	-	э, ё, ы, ъ	и, ъ, щ
UK	ґ, є, і, ї, '	-	ґ, є, и, ї, щ
BE	і, ў, '	э, ё, ы, ў	-

Následně je potřeba vyfiltrovat texty v jiném než cílovém jazyce. Pro jazyky používající jiné abecedy to lze udělat jednoduše po použitých znaků. Pro jazyky využívající stejnou abecedu je možné použít filtrování například na základě nejčastějších klíčových slov pro daný jazyk. Případně je možné použít existující nástroje pro identifikaci jazyka jako například *fastText*[2] [25], který v současné době obsahuje modely pro 170 jazyků. V případě velmi blízkých jazyků, jako jsou například východoslovanské, byla navržena metoda identifikace podle unikátních znaků abecedy každého jazyka. Identifikace probíhá vždy mezi dvěma jazyky podle rozdílů v jejich abecedách. Rozdíly ve východoslovanských abecedách jsou zobrazeny v tabulce 3.1. Tato metoda dosáhla úspěšnosti 96,8 % na 500 testovacích větách.

Tabulka 3.2: Statistika těžení textových dat pro východoslovanské jazyky

Jazyk	**RU**	**UK**	**BE**
Počet zdrojů	12	28	11
Stažených dat	2,38 GB	4,39	2,56 GB
Dat po zpracování	998 MB	2,37 GB	814 MB
Dat po jazykové filtraci	998 MB	758 MB	280 MB

V Tabulce 3.2 jsou zobrazeny statistiky textových korpusů východoslovanských jazyků. Kolik webových serverů bylo použito, kolik dat bylo staženo a kolik následně zbylo po zpracování výše popsaným způsobem a jazykovém filtrování.

[2]https://fasttext.cc/

3.1.1 Vnitřní kódování jazyků

Jelikož se tato práce zabývala i vývojem systémů pro slovanské jazyky používající cyrilici, bylo pro usnadnění práce vhodné vytvořit interní abecedu a převodník. Díky tomu je možné používat při práci standardní klávesnici v latince a není potřeba řešit instalaci jiných abeced a fontů do nástrojů používaných k vývoji. Zároveň při potřebě ručních úprav není potřeba hledat speciální znaky, ale je možno vše zapisovat snadno dostupnými znaky na klávesnici.

Existují oficiální způsoby transliterace například pro ruskou azbuku do latinky (tzv. romanizace). Nicméně tyto převody jsou většinou jednosměrné a přesný převod zpět je problematický. Proto byla navržena vlastní interní abeceda pro cyrilici dle takových zásad, aby se převod co nejvíce přibližoval české výslovnosti, byl psatelný na české klávesnici, aby usnadňoval práci všem členům týmu, a zároveň aby umožňoval přesný převod jedna ku jedné do interní abecedy a zpět do cyrilice. V tabulce 3.3 je ukázka převodu pro zpracované jazyky používající cyrilici.

Tabulka 3.3: Ukázka převodu různých národních variant cyrilice do interní abecedy

RU	Член Общественной палаты России Ирина Волынец предложила изменить правила трудоустройства россиян по совместительству
	Člěn Obşěstvěnnoj palaty Rossii Irina Volyněc prědložila izměnit^ pravila trudoustrojstva rossiân po sovměstitěl^stvu
UK	Колишній небожитель політичного Олімпу Давид Жванія, публікує одне скандальне відео за іншим
	Kolyšnij nebožytel^ polityčnoho Olimpu Davyd Žvaniâ, publikuě odne skandal^ne video za inšym
BE	Мы паспяхова процідзейнічаем незаконнай міграцыі, наркатрафіку, кантрабандзе і гэтак далей
	My paspâxova procidzějničaěm nězakonnaj mihracyi, narkatrafiku, kantrabandzě i hetak daléj
BG	По-късно днес правителството ще се събере и официално ще приеме постановление за удължаване на извънредното положение до средата
	Po-kăsno dnes pravitelstvoto şe se săbere i oficialno şe prieme postanovlenie za udălžavane na izvănrednoto položenie do sredata
MK	Некаде има многу голем број на ученици, па ќе мора да се воведе учење во смени, а во други пак, има помали паралелки и нема да има потреба
	Nekade ima mnogu golem broj na učenici, pa ťe mora da se vovede učeňe vo smeni, a vo drugi pak, ima pomali paralelki i nema da ima potreba

3.2 Tvorba slovníku

Slovník je jednou ze stěžejních částí rozpoznávače. Obsahuje všechna slova, která mohou být rozpoznávačem rozpoznána. Ke každému slovu ve slovníku jsou přiřazeny přípustné výslovnosti (což bude probráno v následující kapitole zabývající se akusticko-fonetickou stránkou ASR).

Slovník je vytvořen výběrem slov (řetězců oddělených mezerami) dle jejich četnosti v textovém korpusu. Počet slov vybraných do slovníku záleží na typu jazyka. Například u analytických jazyků, jako je angličtina nebo španělština, vystačí slovník o velikosti do sto tisíc slov pro pokrytí většiny slovní zásoby. U slovanských jazyků, které mají velkou míru skloňování a časování slov, se ukazuje jako nezbytné množství alespoň tři sta tisíc slov. Dostačující množství slov pro slovník je nalezeno pomocí míry množství slov mimo slovník OOV (tzv. Out-of-vocabulary). Ta se spočítá jako poměr počtu unikátních slov v korpusu, která nebyla vybrána do slovníku, k celkovému počtu unikátních slov. Přijatelná míra OOV je kolem 1-3 %.

Ne všechna vybraná slova jsou reálná slova z daného jazyka, především ta krátká vznikající ze zkratek či rozdělených slov. Je proto vhodné zkontrolovat všechna jedno, dvou a případně i třípísmenná slova, zdali opravdu v daném jazyce existují. Dále je vhodné detekovat a zkontrolovat další podezřelá slova, např. příliš dlouhá, obsahující znaky nepatřící do abecedy zpracovávaného jazyka, dlouhá slova, slova bez samohlásek, atd.

Dalším krokem, který může zlepšit rozpoznávání, je přidání kolokací, tedy velmi častých slovních spojení, kdy dvě či více slov jsou v korpusu spojena do jednoho tokenu a následně přidána do slovníku. Může se jednat o unikátní slovní spojení, kdy se samostatná slova téměř nevyskytují (např. Addis Abeba, Los Angeles, ad absudrum, de iure, ...). Další jsou častá spojení slov především s předložkami či spojkami, a to zejména jednofonémovými. Tím může být značně zlepšeno rozpoznávání, jelikož krátká slova jako předložky a spojky jsou často špatně rozpoznávána. Přidáním kolokací je možné zlepšit rozpoznávací skóre i v řádu procent, jak bylo ukázáno v [26]. Kolokace jsou opět vybírány podle statistiky četnosti v korpusu.

Tabulka 3.4 zobrazuje, jak velké slovníky byly vybrány z korpusů pro východoslovanské jazyky a procento zbývajících slov v korpusu, která nebyla přidána do slovníku.

Tabulka 3.4: Statistika vytvořených slovníků pro východoslovanské jazyky

Jazyk	Velikost korpusu	Počet slov	OOV
RU	998 MB	326 tis.	2,02 %
UK	758 MB	324 tis.	1,94 %
BE	280 MB	293 tis.	1,30 %

3.3 N-gramový jazykový model

Na závěr je pro slova ze slovníku a korpusu vypočítán n-gramový jazykový model. V této práci bylo využito bigramových modelů z důvodu použitého ASR systému. Nicméně vzhledem k tomu, že slovanské jazyky mají relativně volný slovosled a velkou slovní zásobu, byl by vyžadován mnohem větší korpus pro natrénování n-gramového modelu vyššího řádu. Zároveň by se razantně zvýšila výpočetní náročnost rozpoznávání a systémy by tak nemuselo být možné nasadit pro online rozpoznávání v reálném čase.

Pro vytvoření bigramového jazykového modelu jsou spočítány četnosti slov a dvojic slov v korpusu. Podmíněná pravděpodobnost pro každý bigram je poté spočítána podle vztahu 3.1, kde C je četnost výskytu slova či sekvence slov v korpusu.

$$P(w_i|w_{i-1}) = \frac{C(w_i, w_{i-1})}{C(w_i)} \tag{3.1}$$

Pro neviděné sekvence slov, které by měly četnost 0, je použito vyhlazování pomocí Witten-Bellova algoritmu [27].

Jak již bylo zmíněno v předchozí části, pro doplnění většího kontextu jsou přidány kolokace těch nejčastějších slovních spojení, což částečně supluje n-gramový model vyššího řádu. Z tabulky 3.5, která zobrazuje statistiky ruského korpusu a jazykového modelu, lze vidět, že doplněné kolokace jsou obsaženy přibližně v 10 % všech bigramů.

Tabulka 3.5: Statistika ruského korpusu a jazykového modelu

Velikost korpusu	998 MB
Celkový počet slov	149 352 988
Počet unikátních slov	1 594 109
Slov ve slovníku	326 324
Míra OOV	2,02 %
Celkový počet bigramů	144 233 687
Počet bigramů pro slova ve slovníku	127 231 285
Počet unikátních bigramů	30 864 417
Počet unikátních bigramů pro slova ve slovníku	28 589 865
Počet kolokací	1328
Počet bigramů obsahujících kolokaci	13 589 189
Počet unikátních bigramů s kolokací	3 566 667

4 Akusticko-fonetická část multilingválního systému rozpoznávání řeči

V této části jsou řešeny ty moduly a nástroje, které souvisí s převodem akustického signálu řeči na její textový přepis, tj. sestavení foneticko-akustického inventáře, vytvoření výslovnostní části slovníku a s tím související vývoj grafémově-fonémového převodníku a zejména vývoj akustického modelu spojený s automatizovaným vytvářením trénovací databáze pro tento model.

4.1 Foneticko-akustický inventář

Cílem je sestavit optimalizovaný inventář fonémů pro konkrétní jazyk (a doplnit ho o jazykově nezávislý inventář nejčastěji se vyskytujících neřečových zvuků). Při sestavování inventáře se řídíme následujícími kritérii:

- Jako základ pro tvorbu fonetického inventáře slouží fonetické studie a tabulky.
- Snaha o sdílení zvukově blízkých fonémů mezi jazyky, čehož může být využito při tvorbě multilinguálních modelů, které mohou být následně využity v počátečních fázích vývoje akustického modelu pro nový jazyk.
- Snaha o usnadnění automatického převodu mezi ortografickou a fonetickou podobou slova tak, aby bylo možné co nejjednodušeji pracovat s výslovnostmi, zejména při manuálních úpravách, a to i při minimální znalosti daného jazyka.
- Snaha o minimalizaci počtu fonémů, zejména sdružováním foneticky blízkých fonémů a alofonů, a to s ohledem na schopnosti akustického modelu „naučit“ se a v rámci pravděpodobnostních parametrů pokrýt různé fonémové alternativy. Může být navrženo několik hypotéz, které jsou průběžně testovány a nakonec vybrána ta nejvhodnější.

Fonetické studie obecně používají pro zápis fonémů mezinárodní fonetickou abecedu[1] (IPA). Ta nicméně není příliš vhodná pro práci na standardní klávesnici ani pro strojové zpracování. Pro zjednodušení práce je vhodnější varianta, kdy každý foném má svůj vlastní znak, a navíc tyto znaky jsou zvoleny tak, aby je bylo možné psát na standardní klávesnici (v našem případě na české).

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

Proto byla pro každý zpracovaný jazyk navržena vlastní fonetická abeceda, která pro co největší zefektivnění práce využívá přístup navržený v [28] pro českou fonetickou abecedu. Ten se řídí následujícími zásadami:

- Fonémy jsou označovány pouze jedním znakem pro snadnou čitelnost a zamezení nejednoznačností.
- Rozlišuje se mezi velkými a malými znaky.
- K označení fonému je využit znak, který se s daným fonémem nejčastěji pojí (při práci v českém týmu je tedy zohledněna česká výslovnost).
- Znak by měl být zapsatelný na české klávesnici, pro usnadnění rychlého zápisu a oprav.

Obecně je v první fázi vývoje výhodnější vybrat fonetickou abecedu spíše větší, aby pokryla co nejvíce fonémů. Po prvních experimentech lze analyzovat výstupy rozpoznávače řeči a na jejich základě rozhodnout o redukci abecedy. Tabulka 4.1 zobrazuje ukázku fonetické transkripce ruského textu s mezikrokem převodu do interní abecedy.

Tabulka 4.1: Ukázka fonetické transkripce ruštiny

Původní text	В связи с реорганизацией совхозов в тысяча девятьсот девяносто третьем году возглавил крестьянско-фермерское хозяйство в Новосергиевском районе
Převod do interní abecedy	V svâzi s rĕorganizaciĕj sovhozov v tysâča dĕvât^sot dĕvânosto trĕt^ĕm godu vozglavil krĕst^ânsko-fĕrmĕrskoĕ hozâjstvo v Novosĕrgiĕvskom rajoně
Vlastní fonetická transkripce	f sVázi s Reorgaňizácijej sofxózof f týSača ďeVáťsod ďeVánosto tRéťjem gódu vozglávil kResťjánskoFérMerskoje xoZájstvo v novoSérgijefskom rajóňe

4.2 Vytváření výslovnostní části slovníku

Pro vytvoření výslovnosti, tedy převod z ortografické podoby slova na fonetickou, využíváme tzv. G2P převodníku (Grapheme-to-phoneme). Ten může být různých typů od systému využívající předem vytvořená produkční pravidla až po systémy využívající metod strojového učení. Produkčních pravidel je využíváno zejména v počátečních fázích vývoje, kdy nejsou k dispozici žádná data, na kterých by bylo možné trénovat.

Slovanské jazyky jsou tzv. ortograficky mělké, tedy rozdíl mezi výslovností a psanou podobou je relativně malý, a pro fonetickou transkripci tedy stačí menší množství pravidel. Mohou tak být využita produkční pravidla využívající okolní kontext fonému, kde fonetická transkripce tedy probíhá tak, že je procházen ortografický tvar slova znak po znaku, je kontrolováno jeho okolí a podle toho je použito vhodné produkční pravidlo. Pravidla jsou definována ve tvaru:

$$A \rightarrow B/C\_D \tag{4.1}$$

To znamená, že pokud řetězci A předchází řetězec C a je následován řetězcem D, je přepsán na řetězec B. Pro usnadnění zápisu mohou řetězce C a D obsahovat i zástupné skupiny (např. pro samohlásky, znělé souhlásky, atd.), pro které jsou vygenerovány všechny možné variace při porovnávání pravidla. Použitý systém byl vyvinut na základě systému popsaného v [9], kde lze nalézt podrobný popis generování fonetické transkripce pro češtinu.

Příklad několika pravidel pro běloruštinu vypadá následovně:

```
šsâ => Sa / _
t^sâ => ťSa / _
t^sě => ťSe / _
t^sô => ťSo / _
t^si => ťSi / _
t^sú => ťSu / _
t^ => ď / _<W,WW,QW>Q
t^ => ť / _<Q,QQ,WQ>W
t^ => ď / _Q
t^ => ť / _
```

V pravidlech je využito zástupných symbolů jako W pro neznělé souhlásky a Q pro znělé souhlásky. Tyto skupiny jsou vždy vyjmenovány dopředu.

Slova mohou mít i případně alternativní výslovnosti. V naprosté většině případů se jedná o spodobu znělosti na konci slov. Zbylé případy jsou zavedené výslovnosti odlišné od základních pravidel, většinou z historických důvodů. Pro takové případy, pokud se dají popsat novou sadou pravidel, se vygeneruje další výslovnost používající tato upravená pravidla. Pokud se jedná o nesystémové výjimky, je vhodné najít způsob, jak tyto výjimky detekovat ve slovníku a následně pro ně vygenerovat výslovnost. Na závěr může být přistoupeno k ruční tvorbě výslovností pro konkrétní specifická slova.

V pozdějších fázích vývoje, kdy už je k dispozici dostatek dat pro trénování, je pro tvorbu výslovností nasazen G2P systém využívající neuronové sítě. Tím se zabývají další členové týmu v pozdějších fázích, kdy je ASR systém již nasazován do praxe a není to tak součástí této práce.

4.3 Vytváření databáze trénovacích nahrávek

Z trénovacích nahrávek je trénován akustický model. Tvorba akustického modelu je tou nejobtížnější částí při vývoji celého systému rozpoznávání řeči. Pro vytvoření akustického modelu použitelného pro rozpoznávání spojité řeči je zapotřebí alespoň několik hodin nahrávek řeči společně s jejich přesnými fonetickými přepisy. Nahrávky musí pocházet od více mluvčích a měly by být dostatečně fonémově rozmanité. Fonetické přepisy by zároveň měly obsahovat i anotaci neřečových zvuků. Hlavním cílem této kapitoly je popsat způsob automatického vytěžování trénovacích dat z volně dostupných zdrojů, které jsou k dispozici na internetu.

4.3.1 Vlastní systém vytváření trénovacích dat

Základní myšlenkou navrhovaného přístupu tvorby vlastních trénovacích dat je nalézt na internetu co největší množství audio nebo video záznamů, ke kterým je připojen nějaký text, jenž by měl obsahovat promluvy v nahrávkách. Nahrávky jsou přepsány existujícím systémem a tyto přepisy jsou následně porovnány s připojeným textem (dále nazývaný referenční).

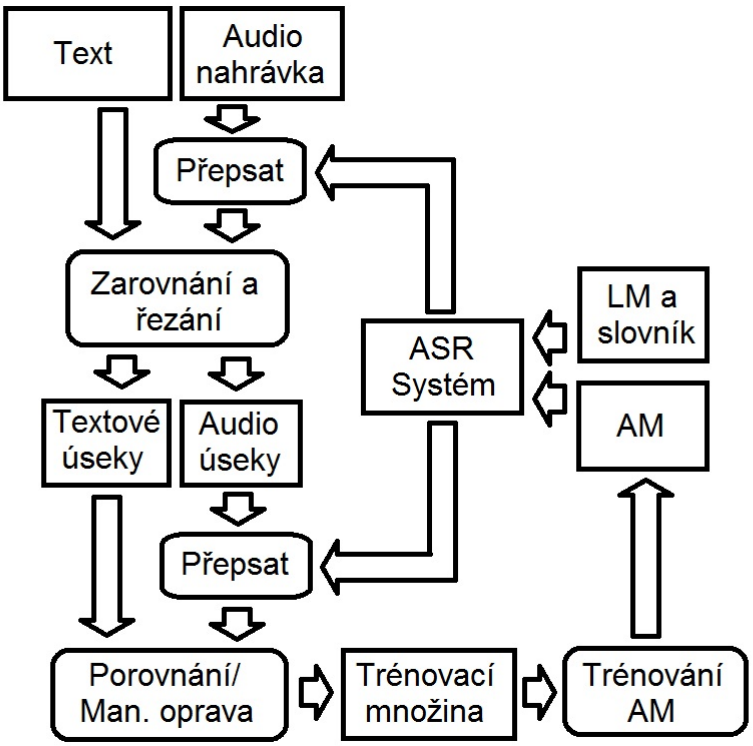
Během porovnávání jsou hledány úseky, kde se shoduje automatický přepis s referenčním textem. Tyto segmenty jsou vyříznuty a dále použity. Pokud se výstup rozpoznávače plně shoduje s textem, je segment přidán do trénovací množiny. Pokud se shoduje jen částečně (např. více než 80 %), může být zkontrolován a opraven manuálně za využití speciálního nástroje, nebo být znovu rozpoznán s novým, lepším modelem, kdy má šanci být správně rozpoznán.

Po získání určitého množství nových trénovacích dat může být natrénován nový akustický model a celý proces se opakuje se zbylými nahrávkami, nebo pouze s těmi krátkými úseky, které neměly 100% shodu a nebyly tak přidány do trénovací množiny. Celý proces zpracování dat a tvorby akustického modelu je zachycen na obrázku 4.1. Jednotlivé kroky jsou detailněji popsány dále.

Zdroje dat

Nejcennějším zdrojem dat jsou webové stránky zpravodajských či televizních a rádiových stanic, které mají volně dostupné zpravodajské i jiné pořady společně s titulky či doslovnými přepisy (v tom nejlepším případě), anebo alespoň doplněné nějakým textem, který popisuje či cituje, o čem se mluví v pořadu. Druhým velmi dobrým zdrojem jsou často parlamentní archivy se záznamy ze zasedání. Dále je možno nalézt a použít volně dostupné audioknihy, internetové lekce daného jazyka a různé další zdroje. Hledání vhodných zpravodajských zdrojů v cizím jazyce je obtížné bez znalosti tohoto jazyka a bez znalosti prostředí. Vhodným zdrojem je tak již zmíněný server ABYZ News Links[2] shromažďující odkazy na zpravodajské weby.

[2] http://www.abyznewslinks.com/



Obrázek 4.1: Schéma iterativního těžení akustických dat

Zpracování dat

Když jsou k dispozici data, přichází na řadu samotný proces těžení dat znázorněný na obrázku 4.1. Pro zjednodušení nejdříve uvažujme, že již je k dispozici fungující systém, tedy existuje výslovnostní slovník, jazykový model a akustický model pro zpracovávaný jazyk. (Následující podkapitola se zabývá případem, kdy se začíná s novým jazykem od úplného začátku a nejsou tudíž k dispozici žádná akustická data). Za využití tohoto systému jsou postupně rozpoznávány jednotlivé nahrávky a výstup rozpoznávače je vždy porovnán s referenčním textem. K porovnání je využit následující postup z [29], kde je využito metody hledání nejmenší vzdálenosti (Minimum edit distance) pro zarovnání obou textů.

K zarovnání dochází na lokální úrovni nejvíce se shodujících částí obou textů. Metoda hledá optimální zarovnání slov referenčního textu r_j o počtu slov J a slov ve výstupu rozpoznávače w_i o počtu slov I. Počty slov I a J se mohou výrazně lišit, podle toho, jak moc odpovídá referenční text promluvě v nahrávce. Tato úloha je řešena pomocí dynamického programování za využití přiřazovací matice A, ve které je hledáno optimální řešení. Proces začíná inicializací prvního řádku a sloupce matice A:

$$A(i,0) = P_D \cdot (i-1), 1 \leq i \leq I; A(0,j) = P_I \cdot (i-1), 1 \leq i \leq J \qquad (4.2)$$

Následně jsou rekurzivně dopočítány zbývající hodnoty matice dle vztahu:

$$A(i,j) = min[A(i-1,j-1)+d(r_i,w_i)-b_{i-1,j-1}; A(i,j-1)+P_I; A(i-1,j)+P_D] \quad (4.3)$$

kde

$$d(r_i,w_j) = \begin{cases} 0, & \text{pokud } r_i = w_j \\ P_S, & \text{pokud } r_i \neq w_j \end{cases} \quad (4.4)$$

a

$$b_{i-1,j-1} = \begin{cases} 0, & \text{pokud } r_i \neq w_j \\ P_S, & \text{pokud } r_i = w_j \end{cases} \quad (4.5)$$

Hodnoty P_D, P_I, a P_S jsou hodnoty pro penalizaci v případě delecí, inzercí a substitucí slov. Obvyklou hodnotou penalizace je 1. Hodnota $b_{i,j}$ napomáhá k vyhledávání nepřerušených sekvencí přidáním bonusu v (4.3). Po dopočítání matice A je nalezeno nejlepší zarovnání zpětným průchodem matice z posledního bodu (I,J) do počátku (1,1) po nejmenších hodnotách. Každé slovo je při zpětném průchodu označeno jako shoda, delece, inzerce nebo substituce.

Jakmile jsou texty zarovnány, je dále použit algoritmus hledající jakékoliv souvislé segmenty, které se do určité míry shodují a jsou ohraničeny tichem nebo některým z hluků. Shoda nemusí být 100%, rozpoznaný text nemusí být zcela odpovídat textu referenčnímu. Při vybrání i těchto segmentů je šance, že budou správně rozpoznány v následujících iteracích s lepším akustickým modelem. Tyto vybrané segmenty jsou následně vyříznuty z původní audio nahrávky i z textu.

Když jsou z původní nahrávky vyřezány shodující se segmenty ve formě audio nahrávek a textu, nahrávky jsou znovu zpracovány rozpoznávačem a opět porovnány s jejich referenčním texty. K vyhodnocení je použita míra WER.

Toto porovnání je provedeno pro všechny vyřezané segmenty. Ty, které mají hodnotu WER nulovou, jsou přidány do trénovací množiny pro trénování nového akustického modelu. Tento přístup zaručuje dostatečnou úroveň kontroly a je víceméně zaručeno, že fonetické transkripce vytvořené tímto procesem skutečně odpovídají tomu, co bylo řečeno v nahrávce.

Tabulka 4.2: Ukázka výstupu rozpoznávače řeči aplikovaného na ukrajinštinu

Ort:	porâdok	kraîny	na	vykonannâ	social^nyx	iniciatyv	prezydenta
Phon:	poRadok	krajiny	na	vykonaňa	sociaLnyX	iňiciatyf	prezydenta
Start:	101	137	148	196	258	318	376
Stop:	137	148	196	258	318	376	426

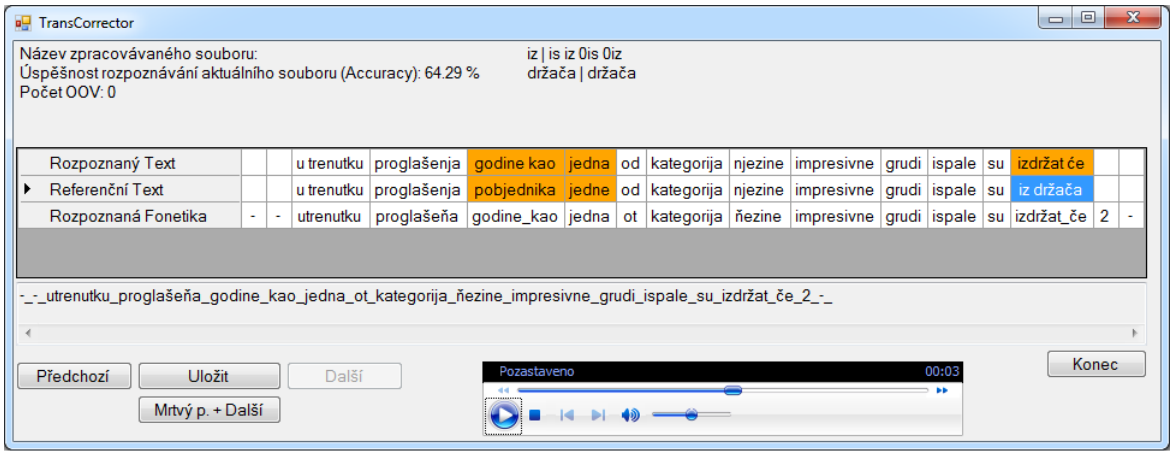
Manuální kontrola a editace

Pomocí speciálního nástroje (na obrázku 4.2) je možno zobrazit jednotlivé segmenty a jejich referenční a rozpoznaný text v ortografické a fonetické podobě.

Oranžově jsou zobrazeny části, kde se referenční text s výstupem rozpoznávače neshoduje. K tomu si lze poslechnout nahrávku a podle ní odhadnout správnou variantu a manuálně opravit neshodující se části. Chyba může být jak ve výstupu rozpoznávače, tak ale i v referenčním textu, který neodpovídá přesně promluvě v nahrávce. Lze zvolit, jaké úseky se budou zobrazovat, podle procent shody referenčního textu a výstupu rozpoznávače. Je tedy tak možné kontrolovat a opravovat jen ty úseky, které mají chybu jen v jednom či dvou slovech.

Tento krok manuální editace není nezbytný, ale slouží zaprvé k urychlení práce v počáteční fázi vývoje, kdy akustický model není ještě příliš kvalitní a výtěžnost dat je nízká. Zadruhé lze pomocí tohoto kroku snadno odhalit chyby ve výslovnosti nebo ortograficky chybná slova, která se dostala do slovníku.

Práce s tímto nástrojem je velice efektivní, neboť anotátor se může zaměřit především na barevně vyznačené úseky. Rozhodnutí, zde je správné rozpoznané nebo referenční slovo lze učinit jediným kliknutím myši. Pro každou nahrávku tak často stačí pouze dva až tři jednoduché úkony.



Obrázek 4.2: TransCorrector - nástroj na kontrolu a opravu fonetických přepisů

Trénování nového modelu

Posledním krokem celého algoritmu je natrénování nového akustického modelu z vytěžených segmentů. Nový model může být použit zaprvé pro zpracování zbývajících segmentů. Tím, že byl model trénován na segmentech z jedné nahrávky, se adaptoval na akustické podmínky v této nahrávce a šance úspěšného rozpoznání a vytěžení dalších segmentů se tím zvyšuje. Zadruhé je nový model použit pro zpracování dosud nepoužitých nahrávek. Formální zápis procesu pro zpracování dat a trénování akustického modelu je zapsán následujícím pseudokódem:

`Iterativní přetrénování:`

1. `Pro každý dokument`
    - `Proveď segmentaci shodných úseků`
2. `Pro každý segment`
    - `Přepiš segment`
    - `Porovnej a přidej do trénovací sady/(Oprav manuálně)`
3. `Přetrénuj akustický model`
4. `Opakuj krok 1. nebo 2.`

4.3.2 Využití multilingválního akustického modelu pro tvorbu trénovacích dat

V případě, že se začíná s vývojem systému pro nový jazyk a nejsou zatím k dispozici žádná akustická data pro tento jazyk, je zde možnost využít v těžícím schématu akustický model pro jiný jazyk (tzv. jazyk zdrojový), případně i více jazyků. V anglicky psané literatuře se tento přístup většinou označuje jako bootstrapping.

Z dostupných jazyků se snažíme vybrat jazyk foneticky nejbližší. Čím jsou si vybrané jazyky foneticky bližší, tím efektivnější je celý proces. Aby systém mohl využívat akustický model pro jiný jazyk, je nutné namapovat fonetickou abecedu cílového jazyka na abecedu jazyka zdrojového a změnit podle toho výslovnosti ve slovníku. Jazykový model zůstává původní pro cílový jazyk.

Akustický model pro čistě zdrojový jazyk je použit pouze v první iteraci těžícího algoritmu. Ve chvíli, kdy je získáno dostatečné množství trénovacích dat (řádově to mohou být desítky minut až jednotky hodin), je natrénován nový akustický model smícháním trénovacích dat pro zdrojový jazyk a nově získaných dat pro cílový jazyk. Tímto procesem se akustický model postupně adaptuje na cílový jazyk a výtěžnost nových dat se zvyšuje.

Ve chvíli, kdy je získáno dostatečné množství trénovacích dat pro cílový jazyk (většinou alespoň 5 hodin), může být zdrojový jazyk odstraněn z trénování, všechna získaná data převedena zpět do fonetické abecedy cílového jazyka a natrénován akustický model pouze pro cílový jazyk, který je dále využíván pro získávání dalších dat. Tento proces využití multilingválního systému je zapsán následovně:

`Multilingvální trénování:`

1. `Namapuj fonémy na zdrojový jazyk`
2. `Přidej zdrojový jazyk do trénovací množiny`
3. `Iterativní přetrénování`
4. `Odstraň zdrojový jazyk z trénovací množiny`
5. `Přemapuj fonémy zpět`
6. `Přetrénuj`

Výběr vhodného zdrojového jazyka

Při volbě zdrojového jazyka jsme v první řadě omezeni na výběr z jazyků, které jsou k dispozici. Ne vždy to musejí být ty nejbližší jazyky, a proto je potřeba vybrat ten nejvhodnější, či nějakou kombinaci jazyků. Dalším kritériem výběru jsou také možnosti mapovaní fonetických abeced.

Ne vždy může být toto mapování jednoduché, jelikož jednotlivé jazyky většinou mají unikátní fonémy, které v ostatních jazycích nejsou (jako např. české /*ř*/ nebo ukrajinské měkké /*c*/). Často se tedy mapuje foném jedné abecedy na nějaký nejbližší foném druhé abecedy či na více fonémů.

Pro výběr nejvhodnějšího jazyka (jazyků) je nutné udělat úvodní test. K tomu jsou potřeba alespoň nějaká testovací sada v cílovém jazyce a připravený slovník a jazykový model. Následně se slovník namapuje na fonetické abecedy testovaných jazyků či jejich kombinací a provede se testování. Následuje ukázka takového postupu pro výběr vhodného zdrojového jazyka pro ukrajinštinu.

Příklad výběru zdrojového jazyka pro vývoj ukrajinštiny

Pro výběr zdrojového jazyka pro ukrajinštinu bylo uvažováno mezi českým, polským a ruským akustickým modelem, které byly v tu chvíli k dispozici.

Přestože se může ukrajinština jevit jako velmi blízká ruštině, v některých ohledech se liší (např. nemá redukci nepřízvučných samohlásek nebo tak silnou palatalizaci souhlásek jako ruština) a podle některých zdrojů má foneticky blíž k polštině (zejména na západě Ukrajiny). Tabulky 4.3 ukazuje příklad, jak byla ukrajinská fonetická sada mapována na ruskou.

Tabulka 4.3: Mapování ukrajinské fonetické sady na ruskou

UK	a	e	i	o	u	X	ć	Ć
RU	á	é	í	ó	ú	x	cj	Cj

Pro testování byly zvoleny tři různé sady. První sada byla vytvořena nahráváním několika rodilých mluvčích, jako druhá byla využita databáze VoxForge pro ukrajinštinu a jako třetí sada bylo využito několik krátkých zpravodajských pořadů ze stanice 5UA, které byly manuálně přepsány. Další informace jsou v Tabulce 4.4.

Tabulka 4.4: Testovací sady pro ukrajinštinu

Testovací sada	**Velikost**	**Mluvčích**
Studiové nahrávky	57 min	5
VoxForge	40 min	9
Zpravodajství 5UA	53 min	-

Na těchto datech byly otestovány všechny tři systémy a vyhodnoceny pomocí vzorce WER. Akustický model pro každý jazyk byl natrénován na stejném množství dat za využití databáze GlobalPhone, která byla k dispozici. Díky tomu byly všechny modely vytvořeny na stejném typu dat a byly tak zajištěny rovnocenné podmínky pro porovnání. V Tabulce 4.5 jsou zobrazeny výsledky testu. Nejlepšího skóre dosáhl ruský model, který byl tedy následně vybrán jako zdrojový jazyk pro vývoj ukrajinštiny.

Tabulka 4.5: Výsledky multilingválního testu pro výběr vhodného jazyka pro vývoj ukrajinštiny

AM	**Velikost**	**WER** [%]		
		Studiové nahrávky	**VoxForge**	**Zprav. 5UA**
RU	11 hod.	40,3	78,3	59,5
CZ	11 hod.	65,8	92,1	74,8
PL	11 hod.	63,0	91,4	79,7

4.3.3 Nesupervizovaný přístup tvorby trénovacích dat

V případě, že nejsou k dispozici žádné nahrávky s textem, které by mohly být zpracovány, přichází na řadu možnost tvorby trénovacích dat pouze z nahrávek bez textu pomocí nesupervizovaného přístupu. K tomu je zapotřebí mít již existující akustický model pro daný jazyk s dostatečným množství trénovacích dat.

Trénovací data jsou rozdělena do různých skupin a z nich jsou natrénovány různé akustické modely. Rozdělení probíhá nejlépe podle zdrojů, ze kterých byly vytvořeny, aby každý model byl natrénován na rozdílných typech dat a dospělo se k určité objektivnosti srovnávání. Případně mohou být pro každý model nastaveny i jiné parametry systému.

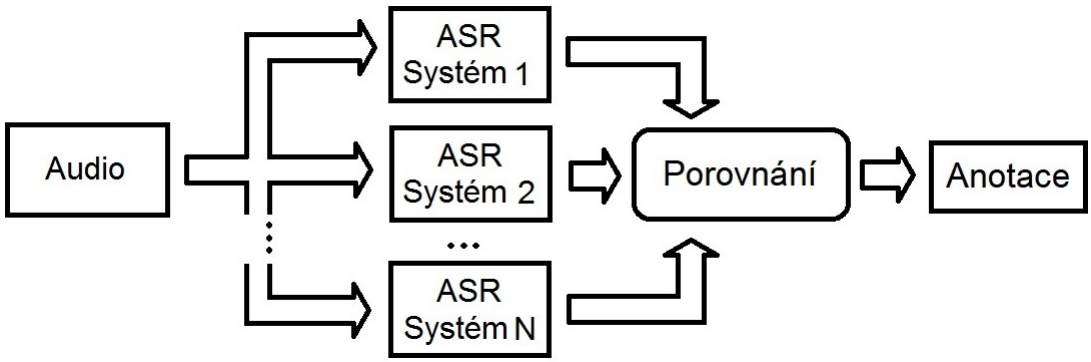
Zpracovávané nahrávky jsou následně rozděleny na krátké segmenty v místech, kde je ticho či nějaký hluk, a každý segment je následně rozpoznán pomocí všech vytvořených akustických modelů. Pokud se všechny přepisy shodují, je segment považován za správně rozpoznaný a i se svým automaticky vytvořeným fonetickým přepisem přidán do trénovací sady. K tomuto účelu mohou být použity i akustické modely z jiných jazyků, čímž se ale může snížit efektivita.

Formálně je proces zapsán následovně:

`Nesupervizované trénování:`

1. `Pro každý dokument`
    - `Proveď segmentaci`
2. `Pro každý segment`
    (a) `Pro každý rozpoznávač`
        - `Přepiš segment`
    (b) `Pokud se všechny výstupy shodují`
        - `Přidej do trénovací sady`
3. `Přetrénuj model`

Tento přístup byl aplikován na zpracování 120 polských televizních pořadů o délce 30 minut. Pro zpracování byly využity 4 rozpoznávače natrénované na různých datech. Celkem bylo tímto přístupem vytěženo 16,4 hodin. Pro dodatečné vyhodnocení byla manuálně zkontrolována náhodná podmnožina dat a bylo zjištěno, že 9 z 10 segmentů bylo správně přepsáno a zbývající segmenty obsahovaly pouze marginální chyby ve výslovnosti. Nesupervizované těžící schéma je zobrazeno na obrázku 4.3.



Obrázek 4.3: Nesupervizované těžení akustických dat

4.4 Trénování akustického modelu

Jak již bylo zmíněno v kapitole 1.2, pro účely této práce byl využit systém rozpoznávání řeči vyvinutý na Ústavu informačních technologií a elektroniky Fakulty mechatroniky TUL [30]. Tento systém byl vyvinut pro češtinu, ale díky jeho modulární struktuře je možné ho jednoduše adaptovat na nové jazyky. V dnešní době je schopen pracovat se slovníkem velikosti půl milionu slov v reálném čase. Dále umožňuje práci jak s GMM, tak s DNN modely.

GMM jsou trénovány jako 32-složkové gaussovské směsi a jako příznakové vektory vstupních segmentů jsou využity 39-dimenzionální MFCC (Mel-Frequency Cepstral Coefficients). Vstupní signál je segmentován po 25 ms s překryvem 10 ms. Pro trénování je použit nástroj HTK Speech Recognition Toolkit.[3] GMM jsou využívány pro zarovnání dat pro trénování DNN.

V případě DNN je využita standardní architektura s pěti skrytými vrstvami v konfiguraci 1024-768-768-512-512 a aktivační funkcí ReLU. Vstupní signál je parametrizován pomocí 39-dimenzionálních Log filter banks. Segmentace vstupu je stejná jako v případě GMM. Pro dodání kontextu je každý frame doplněn pěti předchozími a pěti následujícími framy. Pro trénování je využita knihovna Torch.[4]

[3] http://htk.eng.cam.ac.uk/
[4] http://torch.ch

4.5 Výsledky aplikace popsaných metod a postupů na východoslovanské jazyky

Na závěr jsou zde pro ukázku aplikace popsaných metod popsány některé vybrané kroky z vývoje akustických modelů pro všechny tři východoslovanské jazyky. Údaje ohledně korpusu, slovníku a jazykového modelu jsou popsány v kapitole 3 zabývající se textovou částí.

Jako první byl vývoj zahájen na ruštině za pomoci českého akustického modelu. Ruština byla následně použita pro vývoj ukrajinštiny a ta následně pro běloruštinu. Hlavním cílem bylo vytvořit systém určený především pro monitoring médií, zpravodajské pořady byly proto nejvhodnějšími daty. Krom toho byly zpracovány a použity i záznamy z ruského parlamentu a ruská řečová databáze GlobalPhone.

4.5.1 Ruština

Jako první byla zpracována ruština za využití českého akustického modelu. Pro začátek byl k dispozici ruský řečový korpus GlobalPhone. Ten ale neobsahuje fonetické transkripce, a proto bylo nutné je vytvořit s pomocí českého akustického modelu.

Tabulka 4.6: Vývoj AM pro ruštinu na databázi GlobaPhone

AM	**Velikost**	**WER** [%]
CZ	10 hod.	58,3
RU	5 hod.	30,7
RU	10 hod.	21,6
RU	23 hod.	18,2

Ruská fonetická abeceda byla namapována na českou a podle toho byly upraveny výslovnosti ve slovníku. Následně byl aplikován popsaný algoritmus automatického těžení dat dohromady s manuální kontrolou a korekcí přepisů pro kontrolu a urychlení procesu. Po vytěžení prvních pěti hodin byl systém převeden na čistě ruský a těžení pokračovalo. Postupný proces vývoje je zobrazen v tabulce 4.6.

Pro vyhodnocování tohoto procesu bylo z GlobalPhone odebráno 10 mluvčích, jejichž přepisy byly manuálně zkontrolovány a následně využity pro testování. Údaje o testovací a výsledné trénovací sadě jsou v tabulce 4.7.

Následně bylo pokračováno ve zpracování dalších zdrojů z ruských zpravodajských webů a z ruského parlamentu. Celkové množství získaných a zpracovaných dat shrnuje tabulka 4.9.

Tabulka 4.7: Statistika ruské trénovací a testovací sady využívající databáze Globalphone

	Trénovací sada	**Testovací sada**
Počet mluvčích	105	10
Počet nahrávek	10644	1202
Délka	23 hod.	2,4 hod.

4.5.2 Ukrajinština

Ruský model byl následně využit pro vývoj ukrajinštiny. Zde nebyla k dispozici žádná data pro začátek, pouze testovací sady využité pro výběr jazyka. Ty byly využity společně s ruskými daty k vytěžení prvních pěti hodin. Následně byly odebrány z trénovací sady a dále využívány pouze pro testování průběhu vývoje.

Ukrajinská fonetická sada byla navržena jako podmnožina ruské fonetické sady s rozdílem ukrajinského "měkkého" /*c*/, které bylo namapováno na kombinaci standardního "tvrdé" ruského /*c*/ a /*j*/.

Zpracovávány zde byly opět zpravodajské pořady a celkové množství použitých a zpracovaných dat je uvedeno v tabulce 4.9. Postupný průběh vývoje ukrajinského modelu je zobrazen v tabulce 4.8.

Tabulka 4.8: Vývoj AM pro ukrajinštinu

AM	**Velikost**	**WER [%]**		
		Studiové nahrávky	**VoxForge**	**Zprav. 5UA**
RU	25 hod.	40,3	78,3	59,5
UK+RU	30 hod.	28,3	69,6	30,0
UK+RU	42 hod.	26,3	67,9	25,5
UK	8 hod.	29,7	68,7	28,2
UK	22 hod.	26,1	58,5	22,3
UK	40 hod.	26,7	57,3	19,7

4.5.3 Běloruština

Pro vývoj běloruštiny byla popsaným způsobem vybrána ukrajinština jakožto nejvhodnější zdrojový jazyk a vývoj probíhal obdobným způsobem, jako u předchozích jazyků. Nicméně u běloruštiny bylo zapotřebí odfiltrovat nahrávky v ruštině a ukrajinštině, které se zde objevovaly ve velkém množství. Tím se objem celkově vytěžených dat snížil na 16 hodin, jak ukazuje tabulka 4.9 společně s dalšími údaji.

Tabulka 4.9: Statistika zpracování akustických dat pro východoslovanské jazyky

Jazyk	**RU**	**UK**	**BE**
Zdrojů	7	6	3
Stažených dat	4627 hod.	4015 hod.	955 hod.
Celkem vyřezaných úseků	192 hod.	161 hod.	48 hod.
Celkem vytěžených dat	58,3 hod.	48 hod.	16,4 hod.

4.5.4 Vyhodnocení

Z tabulky 4.9 zobrazující statistiku akustických dat lze vyčíst, že bylo staženo obrovské množství dat, ale jen zlomek byl nakonec vytěžen. Důvodů je k tomu několik.

Hlavní příčinou byla použitá data. Většina dat byly automaticky stažené pořady z webových stránek spolu s textem, který se u nich vyskytoval. Ten ale ve většině případů vůbec neodpovídal promluvě v nahrávkách a byl většinou jen slovní popis obsahu nahrávky, nebo jakýkoliv další možný text k danému tématu.

Bylo tak potřeba zpracovat ohromné množství dat, aby byly nalezeny alespoň některé úseky, kde se text shoduje s promluvou v nahrávce. Druhým důvodem, proč byla výtěžnost tak malá, je také fakt, že při tvorbě přepisů a jejich porovnávání s textem není systém 100% úspěšný. Tedy ne vše přepíše správně, a tudíž nemusí být nalezeny úplně všechny shodující se části.

Další důvod je také ten, že stažená data obsahují různé znělky, hudbu, nebo to jsou jen videa bez promluvy a tak i přes to, že byla zpracována, nemohlo z nich být vytěženo nic.

Na druhou stranu tento přístup zaručuje vytvoření trénovacích dat s velmi přesnými fonetickými přepisy. Jak je důležitá přesnost trénovacích dat, bylo ověřeno v dodatečných experimentech.

5 Souhrnné výsledky dokumentující vývoj ASR systémů pro slovanské jazyky

Pro závěrečné otestování vyvinutých systémů byly vytvořeny standardizované testovací sady. Cílem bylo systémy otestovat na reálných datech z televizního a rozhlasového vysílání, aby bylo možné posoudit jejich použitelnost v reálném provozu při monitorování médií. Systémy byly vytvořeny pro všech 13 slovanských jazyků. Nicméně pro bosenštinu a černohorštinu byl kvůli nedostatku akustických dat použit srbochorvatský akustický model. Rovněž se u těchto jazyků nepodařilo sestavit standardizovaná testovací data, jelikož nebyl nalezen rodilý mluvčí pro jejich validaci. Proto bylo výsledné testování provedeno pouze na 11 slovanských jazycích.

Je nutné dodat, že systém pro češtinu byl na pracovišti vyvíjen od 90. let minulého století a již byl úspěšně nasazen v mnoha výzkumných i komerčních aplikacích. Na jeho základě byly pak vytvořeny systémy pro slovenštinu a polštinu a následně započal vývoj systémů pro východoslovanské a jihoslovanské jazyky, kterého jsem se už v rámci týmu zúčastnil.

V této kapitole je popsán výběr a tvorba testovacích dat, následuje popis a statistiky vytvořených systémů a jejich výsledky dosažené na testovacích datech.

Tabulka 5.1: Statistika testovacích sad pro slovanské jazyky

Jazyk	**Kód**	**Délka [min]**	**Počet slov**
Čeština	CZ	95	13494
Slovenština	SK	92	12365
Polština	PL	105	14742
Slovinština	SL	109	14943
Chorvatština	HR	104	15319
Srbština	SR	89	12791
Makedonština	MK	94	12916
Bulharština	BG	100	15197
Ruština	RU	93	12277
Ukrajinština	UK	75	9440
Běloruština	BE	82	11716

5.1 Standardizovaná testovací sada

Pro testování byly vybrány zpravodajské pořady z hlavních televizních a rozhlasových stanic v jednotlivých zemích. Pro každý jazyk byly zvoleny 3 pořady z alespoň dvou různých stanic v celkové délce okolo 90 minut. Testovací data byla vytvořena z dat odvysílaných několik měsíců (v několika případech roků) později po natrénování akustických a jazykových modelů, aby byla zajištěna skutečná nezávislost experimentů.

Jedná se o kompletní pořady (od otevírací až po závěrečnou znělku) obsahující všechny typy zvuků a promluv běžně se vyskytujících v těchto zprávách - tj. čistá řeč ve studiu, řeč s hudbou či hlukem na pozadí, spontánní řeč lidí na ulici, dabovaná řeč s původní promluvou na pozadí a podobně. Referenční texty byly vytvořeny a zkontrolovány rodilými mluvčími pro všech 11 testovaných jazyků.

Zároveň byly v referenčních textech označeny úseky v jiném než cílovém jazyce. U některých pořadů se může jednat o pasáže v cizím jazyce s přidanými titulky. U dalších, především u východoslovanských zemí, které se vyznačují silně bilinguálním prostředím, se často objevuje používání více jazyků v rámci jednoho pořadu. Konkrétně je to především užívání ruštiny v ukrajinských a běloruských pořadech, ale v jednom běloruském pořadu se objevil i rozhovor v polštině.

Tabulka 5.1 zobrazuje statistiky testovacích sad pro jednotlivé jazyky společně s použitým jazykovým kódem dle ISO 639-1. Testovací sady byly zároveň zpřístupněny a jsou veřejně dostupné.[1]

Tabulka 5.2: Charakteristiky vytvořených modulů pro slovanské jazyky

Jazyk	**Autorův podíl**	**Velikost korpusu [GB]**	**Velikost slovníku [tis.]**	**Počet fonémů**	**Velikost trénovací sady [h]**	**Počáteční jazyk**
CZ	0	6,2	388	41	1050	-
SK	0	2,9	302	41	118	CZ
PL	1	3,00	303	36	58	CZ
SL	1	0,91	300	32	42	HR
HR	1	1,10	304	32	45	CZ
SR	1	1,23	307	32	40	CZ
MK	1	0,83	265	33	40	BG
BG	1	0,98	283	33	41	HR
RU	1	0,98	326	53	58	CZ
UK	2	0,75	324	39	48	RU
BE	2	0,28	293	36	16	UK

[1] https://owncloud.cesnet.cz/index.php/s/qLTs9K5LAeqIZAV

5.2 Charakteristiky vytvořených modulů

V tabulce 5.2 jsou vypsány výsledné charakteristiky všech modulů systému pro jednotlivé jazyky. Jako první je uvedeno v jaké míře jsem se na kterém jazyce osobně podílel v průběhu jeho vývoje. Použito je kódování 0 - nepodílel, 1 - ve spolupráci se školitelem, 2 - samostatně.

Dále je uvedena velikost textového korpusu, ze kterého byl trénován jazykový model, velikost použitého slovníku, rozsah fonetické sady a množství akustických dat použitých pro trénování akustického modelu.

Nakonec je uveden počáteční jazyk, který byl využit v počátečních fázích vývoje akustického modelu daného jazyka (bootstrapping). Jelikož byly systémy pro jednotlivé jazyky vyvíjeny průběžně, vždy bylo vybíráno pouze z dostupných a již dobře fungujících systémů, a tak tedy ve většině případů byla použita čeština.

Trénovací data pro západoslovanské jazyky, a to především pro češtinu a slovenštinu, výrazně převyšují množství dat pro ostatní jazyky, jelikož byly vyvíjeny mnohem delší dobu a jsou již komerčně nasazeny v mnoha aplikacích. Je však třeba říci, že v testech byly použity pouze vývojové verze systémů, nikoliv aktuální produkční verze, které se liší zejména novými typy neuronových sítí využitých v akustických modelech a rovněž pravidelně aktualizovanými slovníky a jazykovými modely. Tyto modifikace jsou již řešeny jinými členy týmu.

Tabulka 5.3: Výsledky rozpoznávání na vytvořených testovacích sadách

Jazyk	**OOV** [%]	**OOL** [**min**]	**WER GMM** [%]	**WER DNN** [%]
CZ	0,87	2,6	24,01	14,72
SK	1,37	1,2	28,05	18,42
PL	0,92	2,3	25,91	20,80
SL	0,68	4,0	23,84	16,16
HR	0,99	0,7	27,11	20,07
SR	0,41	0,3	26,25	18,90
MK	0,52	1,6	26,43	14,54
BG	0,61	0,1	27,66	20,86
RU	2,18	3,7	33,76	22,08
UK	2,75	8,9	36,32	30,15
BE	3,12	15,7	41,83	35,95

5.3 Výsledky rozpoznávání na vytvořených testovacích sadách

Na testovacích sadách byly následně ověřeny všechny systémy. Tabulka 5.3 zobrazuje kromě dosažených výsledků i míru OOV referenčních textů a délky úseků v jiném než cílovém jazyce. Úseky označené jako OOL byly vyřazeny z vyhodnocení pro zachování objektivity.

Výsledné hodnoty WER jsou uvedeny jak pro GMM tak pro DNN modely z důvodu, že v průběhu vývoje byly využívány oba typy modelů. Nicméně DNN modely vždy dosahují vyšší úspěšnosti rozpoznávání a jsou využívány ve finálních verzích systémů.

Z výsledků je patrné, že nejlepších výsledků dosahuje samozřejmě čeština, která se může opřít o mnohem více trénovacích dat a rovněž precizněji připravený slovník. Většina dalších jazyků se však vešla pod hranici 20 % WER, což již znamená poměrně dobře použitelné přepisy pro účely analýz a monitoringu, a také jako základ pro případnou efektivní ruční editaci. Velice dobré výsledky jsou vidět u makedonštiny a slovinštiny, což do jisté míry souvisí i s akustickou kvalitou dat, výslovností mluvčích, obsahem, apod. Tam, kde byl větší podíl promluv profesionálních řečníků snímaných ve studiu, mohl systém dosáhnout mnohem lepších výsledků než tam, kde byly ve větší míře využity spontánní rozhovory z ulice či jinak rušného prostředí.

Dále je vidět, že hodnoty WER u východoslovanských jazyků byly obecně vyšší, což je dáno jak větší složitostí těchto jazyků (oproti prvním dvěma skupinám), tak i vyšší výslovnostní variabilitou způsobenou regionálními odlišnostmi a bilingvismem v těchto zemích. Vyšší je i míra OOV, a to i při větším počtu slov ve slovníku. Nejhoršího skóre dosáhla běloruština, což bylo způsobeno především nízkým množstvím trénovacích dat, které se podařilo získat procesem automatického těžení (důvody jsou podrobněji vysvětleny v předchozí kapitole).

Závěr

V disertační práci jsem se zaměřil na řešení teoretických a praktických otázek spojených s efektivním vývojem multilingválních systémů automatického rozpoznávání řeči. Cílem bylo navrhnout, implementovat a na reálných datech ověřit postupy umožňující relativně rychle a s co nejmenšími náklady adaptovat existující systém pro nové jazyky.

Původní zadání počítalo se zaměřením na slovanské jazyky, u nichž bylo možné využít řadu společných lingvistických i fonetických rysů, nicméně se ukázalo, že navržený postup je dobře použitelný i pro jazyky z jiných jazykových skupin.

Základem celého přístupu je práce s textovými a akustickými řečovými daty, přičemž se ukazuje, že dobrých a v praxi použitelných výsledků lze dosáhnout s daty, která jsou veřejně přístupná na internetu, a s využitím vhodných metod strojového učení.

Navržený postup lze stručně popsat následujícím schématem:

1. Vytvoření dostatečně rozsáhlého a reprezentativního korpusu textů ze zdrojů přístupných na internetu a jeho následná úprava.

2. Vytvoření slovníku ze slov nejčastěji se vyskytujících v korpusu.

3. Vytvoření jazykového modelu pro daný slovník a korpus.

4. Definice fonetického inventáře pro daný jazyk.

5. Vygenerování výslovností pro všechna slova ve slovníku.

6. Shromáždění co největšího množství řečových nahrávek spolu s doprovodnými textovými daty, která více či méně odrážejí mluvený obsah nahrávek.

7. Iterativní proces výběru těch nahrávek, v nichž text odpovídá mluvenému obsahu. Shodu mluveného a textového obsahu určuje samotný vyvíjený rozpoznávací systém. Ten v počáteční fázi využívá akustický model jiného (již zvládnutého) jazyka a postupně ho adaptuje na nový jazyk na základě fonetických přepisů vytvořených systémem.

8. Při dostatečném množství trénovacích dat v cílovém jazyce pokračuje iterativní proces už s vlastním akustickým modelem, přičemž lze využít i nesupervizovaného přístupu, kdy se trénovací sada postupně rozšiřuje na základě vyhodnocování shody mezi různě nakonfigurovanými rozpoznávacími systémy.

9. Uvedený postup může běžet téměř automaticky. Je však vhodné doplnit ho o řízenou kontrolu těch nahrávek, ve kterých se referenční a rozpoznaný text liší v malém počtu (jednoho až dvou) slov, u nichž lze pomocí vhodně navrženého nástroje snadno (i pro laika) odhalit a opravit zdroj chyby.

10. Experimenty provedené na více než deseti různých jazycích ukazují, že popsaným postupem lze vytvořit funkční rozpoznávací systém pracující s reálnými daty s chybovostí nižší než 30 %, a který lze využít jak pro demonstrační účely (např. pro budoucího klienta), tak jako základ pro vývoj skutečné komerční aplikace.

Výše uvedené schéma se postupně vyvíjelo a optimalizovalo s každým dalším zpracovávaným jazykem. Pro často se opakující činnosti byly vyvíjeny nástroje, kterými jsem se snažil tyto činnosti co nejvíce zautomatizovat a převést je na programy či skripty s volitelnými parametry specifickými pro každý jazyk. Díky tomuto postupu a také díky modelům pro již zpracované jazyky je nyní možné uskutečnit základní vývoj systému pro další jazyk v průběhu cca 3-6 měsíců a dosáhnout přesnosti na úrovni výsledků zmíněných v poslední kapitole.

Shrnutí přínosů práce k rozvoji vědního oboru

Vědecké přínosy práce jsou shrnuty v následujících bodech:

- Byla navržena a prakticky prověřena metodologie procesu vývoje jazykově závislých modelů systému rozpoznávání řeči za využití co největší míry automatizace procesu především v časově náročných a opakujících se úlohách.

- V rámci metodologie byly vytvořeny jednotné zásady pro zpracování a tvorbu dat týkající se především způsobu značení, kódování či formátování dat.

- Pro jednotlivé dílčí kroky vývoje byly vytvořeny efektivní nástroje dodržující stanovené zásady, které mohou být využity jak pro tvorbu systémů pro další jazyky, tak i pro jiné úlohy v oblasti zpracování řečového signálu.

- Při práci s jazyky používající odlišné abecedy byl navržen způsob efektivního převodu mezi abecedami pro usnadnění práce lidí neznalých dané abecedy.

- V mnoha krocích vývoje bylo využito různých metod strojového učení, především tzv. lehce supervizovaného, které byly adaptovány pro konkrétní účely.

- Navržené metody byly aplikovány na všechny národní slovanské jazyky a tedy i ty, pro které zatím podle dostupné literatury žádné systémy pro rozpoznávání spojité řeči neexistovaly, jako běloruština, bosenština, černohorština či makedonština.

- Dále byly metody úspěšně aplikovány i na jiné neslovanské jazyky, a to i jazyky s nedostatečnými zdroji jako lotyština či albánština, kde podle dostupných zdrojů existuje pouze jeden systém rozpoznávání řeči v případě lotyštiny a pravděpodobně žádný v případě albánštiny.

- Zároveň byly ve spolupráci s rodilými mluvčími vytvořeny unifikované testovací sady z televizních a rozhlasových zpravodajských pořadů pro 11 slovanských jazyků (tedy téměř všech národních jazyků kromě bosenštiny a černohorštiny), na kterých byly testovány finální systémy. Data byla veřejně zpřístupněna a využita nejen námi, ale i zahraničními týmy, například v [31].

Shrnutí přínosů práce pro praxi

Vytvořené systémy byly postupně nasazovány partnerskou firmou Newton technologies, a.s., do praxe v rámci společných projektů "MULTILINMEDIA" a "DeepSpot" pro včasné upozorňování a monitorování médií nebo jako další komerční aplikace například pro automatické titulkování pořadů programem Beey či pro diktovací software Newton Dictate použitelný jak pro obecné diktování, tak i například pro soudy či speciální lékařská oddělení.

Tyto produkty jsou již komerčně nasazeny kromě České republiky také na Slovensku, v Polsku, ve Slovinsku, v Chorvatsku a v Srbsku a jsou podle potřeb neustále aktualizovány novými daty, která jsou zpracovávána popsanou metodologií.

Navazující a budoucí práce

Nástroje, data a metodologie vytvořené během této práce už průběžně jsou nebo budou využity v dalších projektech a výzkumech řešených na pracovišti.

Celý postup již byl úspěšně aplikován i dalšími členy týmu pro tvorbu nových systémů pro další evropské jazyky. V plánu je aplikace i pro další jazyky jako například švédština a norština v rámci nově zahajovaného mezinárodního projektu.

Data a znalosti získané v rámci této práce jsou využívány rovněž v rámci výzkumu a vývoje systémů pro identifikaci jazyka, především pro slovanské jazyky [32].

V neposlední řadě všechna vytvořená řečová data slouží také při vývoji společného multilingválního mnohavrstevného akustického modelu typu DNN, kterým se zabývají další členové týmu.

Literatura

[1] Huang, Xuedong, et al., „Spoken language processing: A guide to theory, algorithm, and system development“, Prentice hall PTR, 2001.

[2] Juang, Biing-Hwang, and Lawrence R. Rabiner. "Automatic speech recognition–a brief history of the technology development." Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara 1 (2005): 67.

[3] O'Shaughnessy, Douglas. "Automatic speech recognition: History, methods and challenges." Pattern Recognition 41.10 (2008): 2965-2979.

[4] Čermák, František. Jazyk a jazykověda. Karolinum Press, 2011.

[5] Sgall, Petr. Jazyk, mluvení, psaní. Karolinum Press, 2011.

[6] Pala, Karel, and Klára Osolsobě. Základy počítačové lingvistiky. Masarykova univerzita, 1992.

[7] Ashby, Michael, and Maidment, John, „Úvod do obecné fonetiky“, Charles University in Prague, Karolinum Press, 2015.

[8] Volín, J., „Fonetika a fonologie“, MSoČ 1, 2010, 43–45.

[9] Nouza, J., Koldovský, Z., Vích, R., „Řeč a počítač: principy hlasové komunikace, úlohy, metody a aplikace: sborník článků“, Liberec: Technická univerzita v Liberci, 2009. ISBN 978-80-7372-548-8.

[10] Gauvain, J-L., and Lori Lamel. "Large-vocabulary continuous speech recognition: advances and applications." Proceedings of the IEEE 88.8 (2000): 1181-1200.

[11] Schultz, Tanja, Martin Westphal, and Alex Waibel. "The globalphone project: Multilingual lvcsr with janus-3." Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop. 1997.

[12] Young, Steve J., and Sj Young. "The HTK hidden Markov model toolkit: Design and philosophy." (1993): 69.

[13] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

[14] Kucera, Karel. "The Czech National Corpus: principles, design, and results." Literary and linguistic computing 17.2 (2002): 245-257.

[15] Švec, Jan, et al. "Web text data mining for building large scale language modelling corpus." International Conference on Text, Speech and Dialogue. Springer, Berlin, Heidelberg, 2011.

[16] Schultz, Tanja. "SPICE-An Interactive Toolkit for Rapid Portability of Speech Processing Systems to new Languages." Multilingual Speech and Language Processing. 2006.

[17] Sharada, C. S., Vijaya, C., "Speech Recognition Using Monophone and Triphone Based Continuous Density Hidden Markov Models", International journal of research and scientific innovation 2015, pp. 30-35.

[18] Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2015.

[19] Juzová, Markéta, Daniel Tihelka, and Jakub Vít. "Unified Language-Independent DNN-Based G2P Converter." INTERSPEECH. 2019.

[20] Parent, G., Eskenazi, M., „Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data", In Proceedings of IEEE Workshop on Spoken Language Technology, pp- 312– 317, Berkeley, California, 2010.

[21] Braunschweiler, N., Gales, M., Buchholz, S., „Lightly supervisedrecognition for automatic alignment of large coherent speech recordings", in Proc. Interspeech, Makuhari, Japan, Sept. 2010, pp. 2222–2225.

[22] Meng, M., Wang, S., Liang, J., Ding, P., Xu, B., „Full utilization of closed-captions in broadcast news recognition", in Proc. IS-CSLP, Kent Ridge, Singapore, 2006.

[23] Davel, M. H., van Heerden, C., Kleynhans, N., Barnard, E., „Efficient harvesting of Internet audio for resource-scarce ASR", in Proc. Interspeech, 2011, pp. 3153-3156.

[24] Loof, J., Gollan, C., Ney, H., „Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System", Interspeech 2009. Brighton, UK. 2009.

[25] Grave, Edouard, et al. "Learning word vectors for 157 languages." arXiv preprint arXiv:1802.06893 (2018).

[26] Kolorenč, Jan. Tvorba a adaptace lingvistické vrstvy pro systém rozpoznávání mluvené češtiny. Diss. Technická Univerzita v Liberci, 2007.

[27] Witten, Ian H., Bell, Timothy C., "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression", IEEE Transactions on Information Theory, 37(4):1085–1094, 1991.

[28] Nouza, J., Psutka, J., Uhlíř, J., „Phonetic Alphabet for Speech Recognition of Czech", In Radio Engineering, vol. 6, no. 4, 1997, pp. 16-20.

[29] Nouza, J., Červa, P., Kuchařová, M., „Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages". Radioengineering, 22(3), 2013.

[30] Nouza, J., „A Czech Large Vocabulary Recognition System for Real-Time Applications", In Text, Speech and Dialogue (eds. Sojka, Kopecek, Pala) Springer-Verlag, Heidelberg, 2000, pp. 217-222.

[31] Abdullah, Badr, et al. "Cross-Domain Adaptation of Spoken Language Identification for Related Languages: The Curious Case of Slavic Languages." arXiv preprint arXiv:2008.00545 (2020).

[32] Matějů, L., Červa, P., Žďánský, J. a Šafařík, R. Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 1. vyd. Indie: ISCA, 2018 S. 1803 – 1807. ISSN: 2308-457X.

Autorovy publikace

1. Šafařík, R., Nouza, J., „Methods for rapid development of Automatic Speech Recognition“, ECMSM 2015, Liberec, 2015.

2. Nouza, J., Červa, P., Šafařík, R., „Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources“, In LTC'15, Poznań, Poland, November 2015.

3. Šafařík, R., Matějů, L., „Impact of Phonetic Annotation Precision on Automatic Speech Recognition Systems“, In Proc. TSP 2016, Vienna, Austria, pp. 311-314, June 2016.

4. Nouza, J., Šafařík, R., Červa, P., „ASR for South Slavic Languages Developed in Almost Automated Way“, In Proc. INTERSPEECH 2016, San Francisco, USA, September 2016.

5. Boháč, M., Matějů, L., Rott, M., Šafařík, R., „Automatic Syllabification and Syllable Timing of Automatically Recognized Speech - for Czech“, In Proc. TSD 2016, Brno, Czech Republic, pp. 540-547, September 2016.

6. Šafařík, R., Matějů, L., „The Impact of Inaccurate Phonetic Annotations on Speech Recognition Performance“, In Proc. TSD 2017, Prague, Czech republic, pp. 402-411, August 2017.

7. Šafařík, R., Nouza, J., „Unified Approach to Development of ASR Systems for East Slavic Languages“, In Proc. SLSP 2017, Le Mans, France, pp. 193-203, October 2017.

8. Nouza, J., Šafařík, R., „Parliament archives used for automatic training of multi-lingual automatic speech recognition systems“, In Proc. TSD 2017, Prague, Czech republic, pp. 174-183, August 2017.

9. Nouza, J., Červa, P., Šafařík, R., „Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources“, In Proc. Lecture Notes in Artificial Intelligence (LNAI), 2017.

10. Šafařík, R., Matějů, L.: „Automatic Development of ASR System for an Under-Resourced Language“, In proc. of 41st International Conference on Telecommunications and Signal Processing, TSP 2018, Athens, Greece, pp. 1-4, July 2018.

11. Šafařík, R., Matějů, L., Weingartová, L.:„The Influence of Errors in Phonetic Annotations on Performance of Speech Recognition System“, In proc. of 21st International Conference on Text, Speech and Dialogue, TSD 2018, Brno, Czech republic, pp. 419-427, September 2018.

12. Mateju, L., Cerva, P., Zdansky J., and Safarik R.: „Using Deep Neural Networks for Identification of Slavic Languages from Acoustic Signal“, Interspeech 2018, Hyderabad, India, pp. 1803-1807, September 2018.