Creation of Lexicons and Language Models for Automatic Broadcast News Transcription

Tvorba slovníků a jazykových modelů pro automatický přepis zpravodajských pořadů

Autoreferát disertační práce

Autor:	Ing. Dana Nejedlová
Studijní program:	P2612 Elektrotechnika a informatika
Studijní obor:	2612V045 Technická kybernetika
Státní doktorská zkouška:	Liberec, 10. březen, 2004
Pracoviště:	Katedra informatiky Hospodářská fakulta Technická univerzita v Liberci Hálkova 6, 461 17 Liberec
Školitel:	Prof. Ing. Jan Nouza, CSc.

Rozsah disertační práce a jejích příloh

Počet stran:	88
Počet obrázků:	13
Počet tabulek:	26
Počet vzorců:	49
Počet příloh:	1

© Dana Nejedlová, leden 2006

Abstract

Industrial civilization generates a large amount of audio & video data. At the same time the knowledge of the information content of the resulting data collections is usually very useful. The research teams of the most developed countries have already solved the problem of transcription of audio signal of human speech into text, but in every language some space for further improvements of speech recognition technology still remains.

This thesis deals with all linguistic aspects of the creation of continuous speech recognizer for the Czech language. It describes the process of the preparation of text corpus, vocabulary for the recognizer, phonetic transcription of the words in the vocabulary, language model, the process of tuning of the recognizer's parameters, and the collection of test speech database.

This database and its predecessors were used for testing various bigram language models, the influence of phonetic transcription and text normalization on the accuracy of recognition of broadcast news.

The speech recognition experiments were carried out with the use of the recognizer developed at SpeechLab, the Laboratory of Computer Speech Processing at the Faculty of Mechatronics and Interdisciplinary Engineering Studies of the Technical University of Liberec. This system is also described in this thesis.

Abstrakt

Průmyslová civilizace vytváří veliké množství audiovizuálních dat. Zároveň je zřejmé, že znalost informačního obsahu výsledných kolekcí dat je obvykle velmi užitečná. Výzkumné týmy z nejvyspělejších zemí již vyřešily problém přepisu zvukového signálu lidské řeči na text, ale pro každý jazyk stále zůstává nějaký prostor pro vylepšování technologie rozpoznávání řeči.

Tato práce se zabývá všemi lingvistickými aspekty tvorby rozpoznávače plynulé řeči pro český jazyk. Popisuje proces přípravy textového korpusu, slovníku pro rozpoznávač, fonetickou transkripci slov ve slovníku, jazykový model, proces ladění parametrů rozpoznávače a tvorbu testovací databáze promluv.

Tato databáze a její předchůdkyně byly využity pro testování různých bigramových jazykových modelů, vlivu fonetické transkripce a normalizace textu na přesnost rozpoznávání zpravodajských pořadů.

Experimenty s rozpoznáváním řeči byly prováděny na rozpoznávači vyvinutém v Laboratoři počítačového zpracování řeči SpeechLab Fakulty mechatroniky a mezioborových inženýrských studií Technické univerzity v Liberci. Tento systém je v práci rovněž popsán.

Obsah

1. Úvod

Přepis zpravodajských pořadů do textové podoby je v případě českého jazyka stále realizován výhradně lidmi. V anglicky mluvícím světě je však tato činnost předmětem zájmu již poměrně známého vědeckého oboru snažícího se tento proces co nejvíce automatizovat. Zmíněný vědecký obor se nazývá „Broadcast News Transcription", se známou zkratkou BNT, a je speciálním oborem širšího oboru zvaného „automatické rozpoznávání spojité řeči". Tento obor je zase speciálním oborem širšího oboru zvaného „automatické rozpoznávání řeči". To napovídá, že dosti záleží na tom, zda rozpoznávaná řeč je spojitá.

Opakem řeči spojité je řeč izolovaná. První úspěšné pokusy s rozpoznáváním izolované řeči, které spočívaly v tom, že člověk říká jednotlivá slova oddělená pauzou do počítače a počítač jemu známá slova zapisuje, byly realizovány v Bellových laboratořích v USA v 50. letech 20. století. Je pozoruhodné, že úspěšné pokusy s rozpoznáváním spojité řeči, kdy počítač umí sám rozpoznávanou promluvu rozdělit na jednotlivá slova, byly realizovány až přibližně o 20 let později.

Rozpoznávání spojité řeči je totiž o hodně složitější než rozpoznávání jednotlivě vyslovovaných slov. Pro rozpoznávání izolovaných slov bylo potřeba vyřešit tři následující úkoly:

1. Počítačová reprezentace zvuku.
2. Extrakce příznaků lidské řeči. Zvukový záznam s lidskou řečí obsahuje mnohem více informací, než je nutné mít pro rozpoznání řeči. Vybrání pouze těch parametrů, které popisují řeč umožní vytvořit dobré modely slov, které má rozpoznávač umět rozpoznat.
3. Algoritmizace porovnání posloupnosti příznaků v rozpoznávaném signálu s modely slov.

Pro rozpoznávání spojité řeči bylo potřeba vyřešit úkolů mnohem více:

1. Počítačová reprezentace zvuku.
2. Segmentace spojitého vstupního signálu na úseky dlouhé několik slov.
3. Extrakce příznaků lidské řeči.
4. Reprezentace slov pomocí takzvaných skrytých markovských modelů a vývoj algoritmů pro práci s nimi.
5. Segmentace akustického signálu slov na fonémy, čímž se získají data pro trénování skrytých markovských modelů.
6. Trénování skrytých markovských modelů reprezentujících fonémy.
7. Fonetické transkripce textové podoby slov.
8. Jazykové modelování.
9. Zpracování velkých textových korpusů.
10. Tvorba slovníku pro rozpoznávač.
11. Tvorba a ladění rozpoznávače.

Minimálně body 7 a 8 jsou závislé na jazyku, takže je zde prostor pro originální objevy laboratoří pracujících s různými jazyky.

Úspěšná realizace automatického rozpoznání spojité řeči bude značnou pomocí pro velký úkol dneška, kterým je získávání znalostí z multimediálních dat, která denně vznikají.

2. Systém pro automatický přepis zpravodajských pořadů vyvinutý v laboratoři SpeechLab

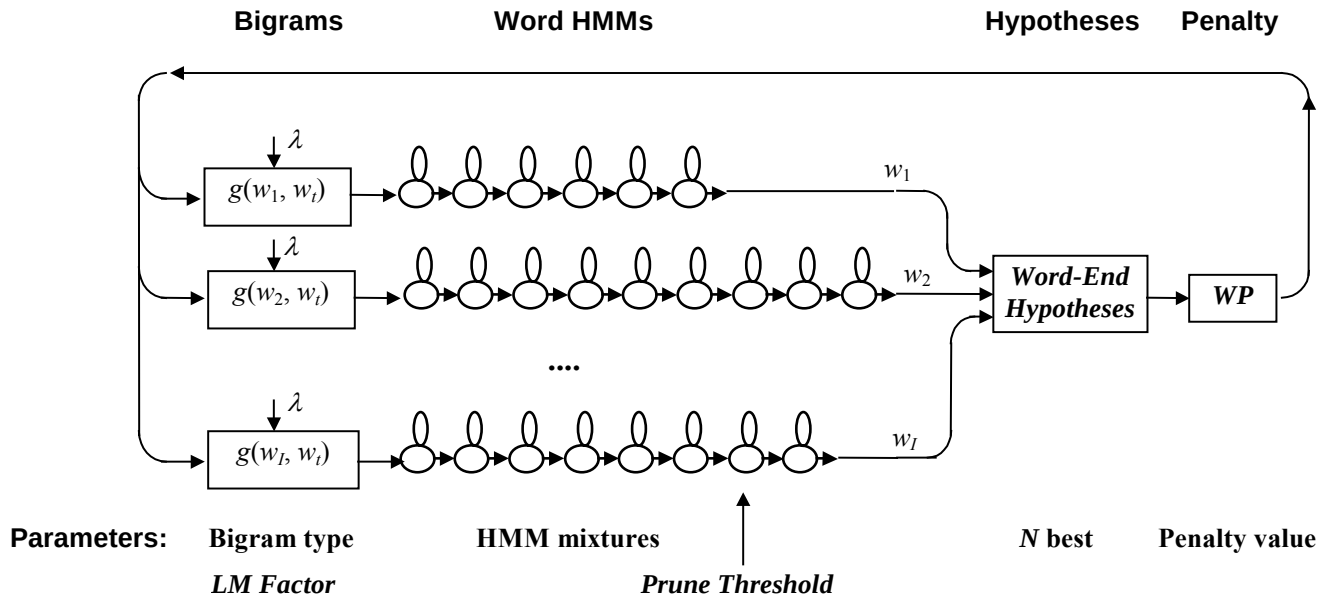
Výzkum popsaný v této práci byl proveden v české laboratoři pro zpracování řeči zvané SpeechLab. Tato laboratoř byla založena profesorem Janem Nouzou v roce 1994 a je součástí Fakulty mechatroniky Technické univerzity v Liberci.

Výzkum ve SpeechLabu začal rozpoznáváním izolovaných slov a krátkých frází. První systém pro rozpoznávání spojité řeči byl zde vyvinut v roce 1999 a jeho slovník čítal 200 slov. Tento systém byl postupně zdokonalován, aby pracoval rychleji a s většími slovníky, takže v roce 2005 rozpoznávač této laboratoře dokáže pracovat se slovníkem o 312 tisících slovech.

Úkolem rozpoznávače je najít takovou sekvenci slov, která maximalizuje pravděpodobnost (1).

$$P(w_1^*,...,w_N^*) = \underset{w,N}{Max} \sum_{n=1}^{N} (\lambda \cdot \ln g(w_{n-1}, w_n) + WP + \ln V(w_n)) \qquad (1)$$

Člen $g(w_{n-1}, w_n)$ v rovnici (1) reprezentuje jazykový model neboli bigramovou pravděpodobnost, že slovo w_n následuje za slovem w_{n-1}. Člen $V(w_n)$ v rovnici (1) je roven akustickému skóre slova w_n dosaženého jeho skrytým markovským modelem vyhodnoceným na sekvenci vektoru příznaků vypočtených ze vstupního signálu. Z důvodu zcela rozdílné povahy jazykového modelu $g(w_{n-1}, w_n)$ a akustického skóre $V(w_n)$ má člen pro jazykový model váhu zvanou λ nebo *LM Factor*. Některá naše i cizí pozorování ukazují, že rozpoznávač má tendenci preferovat kratší slova před delšími. Pro potlačení tohoto jevu jsme zavedli parametr zvaný *Word Insertion Penalty* nebo *WP*, který zhoršuje skóre každého posuzovaného slova. Rovnice (1) je řešena pomocí takzvaného Viterbiho algoritmu popsaného v literatuře [1] a [2] (str. 176 – 180) obohaceného o techniky prořezávání hypotéz. Schéma rozpoznávače ukazuje Obrázek č. 1.



Obrázek č. 1 Struktura rozpoznávání sekvence slov se všemi klíčovými parametry [3]

Vstupní signál je vzorkován pomocí 16 bitů s frekvencí 8 kHz a jsou z něj extrahovány 39-příznakové MFCC vektory pro každý frame. MFCC je zkratka pro „mel frequency cepstral coefficients“, česky „kepstrální koeficienty mel-frekvence“ vysvětlované například v [1] (str. 75 – 87) Všechna slova ve slovníku rozpoznávače jsou reprezentována zřetězenými modely fonémů. Jednotlivé modely fonémů korespondují s třístavovými skrytými markovskými modely trénovanými na několikahodinové anotované databázi promluv. Skryté markovské modely jsou 32-mixturové monofóny.

Akustické modely všech slov jsou vyhodnocovány paralelně s předpokladem, že libovolné slovo může začít v libovolném framu. Frame je 10-milisekundový segment akustického vstupního signálu. Každý frame je převeden do akustického příznakového vektoru během procesu zvaného parametrizace. V každém framu jsou méně pravděpodobné cesty ve Viterbiho algoritmu odříznuty, aby se urychlil výpočet. To je ovlivňováno parametrem zvaným *Prune Threshold*. Slova, jejichž skóre vypočtené Viterbiho algoritmem dělené nejlepším dosud nalezeným skóre je nižší než *Prune Threshold*, jsou dočasně odstraněny z výpočtů. Nejvyšší skóre v konečných stavech skrytých markovských modelů patří nejpravděpodobnějším slovům, která končí v daném framu. Pro další výpočty je vybrán jen omezený počet nejlepších slovních kandidátů. Tento počet slov je určen parametrem zvaným *Number of Word-End Hypotheses*. Skóre každého slovního kandidáta je penalizováno konstantou zvanou *WP* nebo *Word Insertion Penalty*. V dalších výpočtech jsou jako následující slovo brány v úvahu hypotézy o všech možných slovech ve slovníku rozpoznávače. Počáteční skóre nového slova je rovno skóre slova, které bylo předtím ukončeno a ke kterému je připočtena bigramová pravděpodobnost $\lambda \cdot \ln g(w_{n-1}, w_n)$ nového slova za podmínky, že mu předchází předtím ukončené slovo.

Ta část výše popsaného rozpoznávače, která má nejblíže tématu této práce, je bigramový jazykový model. Když má slovník rozpoznávače 312 000 slov, bigramový jazykový model je tabulka pravděpodobností posloupností všech možných slovních párů z tohoto slovníku, takže tato tabulka má počet hodnot rovný druhé mocnině čísla 312 000. Když je každá hodnota v počítači reprezentována 4 byty, celý jazykový model zabírá 363 GB paměti. Vzhledem k hardwarovým omezením současné výpočetní techniky je nezbytné tuto tabulku komprimovat.

Popis komprese jazykového modelu je v článku [4]. Komprese využívá fakt, že mnoho hodnot ve vyhlazeném jazykovém modelu je stejných. Tabulka bigramového modelu je rozdělena do vektorů $\mathbf{h}(w_{n-1})$ hodnot podmíněných pravděpodobností slovních párů, které sdílí stejné první slovo. Tyto vektory mohou být efektivně komprimovány, protože obsahují menší nebo větší skupiny stejných hodnot. Výsledkem je, že jazykový model 312 tisícového slovníku zabírá pouze 251 MB operační paměti [5]. Hodnoty ve vektorech $\mathbf{h}(w_{n-1})$ jsou navíc uspořádány nikoliv podle svého pořadí ve slovníku ale podle svých hodnot od nejvyšší do nejnižší. Toto uspořádání umožňuje další výpočetní úsporu. V každém framu je vyhodnocováno pouze tolik vektorů $\mathbf{h}(w_{n-1})$, kolik je hodnota parametru *Number of Word-End Hypotheses*. A z tohoto malého množství vektorů je využito pouze omezené množství nejvyšších hodnot, protože méně pravděpodobná následující slova jsou z výpočtů odstraněna díky parametru *Prune Threshold*.

3. Stav výzkumu automatického přepisu zpravodajských pořadů ve světě

Výzkum různých laboratoří se liší hlavně dvěma faktory: jazykem řeči, která je automaticky rozpoznávána, a tím, zda-li je používán vlastní nebo veřejně dostupný rozpoznávač.

Hlavní vlastnost jazyka, která ovlivňuje úspěšnost rozpoznávání, je takzvané pokrytí textu slovníkem. Rozpoznávání v jazycích, které s určitým počtem nejfrekventovanějších slov pokryjí co největší procento slov v předem neznámém textu, je ve velké výhodě. Takovým jazykem je angličtina. Když se k tomu ještě přičte fakt, že angličtina je jazykem zemí s nejrozvinutějším vědeckým výzkumem, není divu, že rozpoznávání řeči v anglickém jazyce má již mnoho úspěšných komerčních aplikací. Výsledky výzkumu rozpoznávání angličtiny jsou využívány i v laboratořích zabývajících se rozpoznáváním jiných jazyků, ale rozdílnost jazyků často vede k tomu, že pro jiné jazyky je nutné zvolit jiné metody.

Jedním z nejvážnějších problémů způsobených odlišností angličtiny od ostatních jazyků je množství slov nutné k dosažení určitého pokrytí. Zatímco slovník 60 tisíc nejfrekventovanějších anglických slov pokryje 99 % textu, v češtině je to 92 %, jak uvádí práce [6] na straně 66. Výsledkem výzkumu rozpoznávání angličtiny je i několik rozpoznávačů, které jsou veřejně dostupné. Některé výzkumné týmy staví svůj výzkum na nich, protože vývoj vlastního rozpoznávače je velmi náročný. Pokud je předmětem zájmu těchto týmů jazyk s mnohem menším pokrytím, než má angličtina, musí tyto týmy překonávat mnohé možná nepřekonatelné problémy, protože velikost slovníku veřejně dostupných rozpoznávačů je dimenzována pouze pro potřeby angličtiny na přibližně 60 tisíc slov.

Výsledkem je však mnoho originálních přístupů k řešení, které byly vymyšleny při rozpoznávání jiných jazyků, než je angličtina. Mezi tyto přístupy patří například rozdělování nebo naopak spojování lexikálních jednotek rozpoznávače, různé způsoby vyhlazování jazykového modelu a jeho přizpůsobování tématu promluv nebo také používání jazykového modelu založeného na slovních gramatických kategoriích.

Výzkumy rozpoznávání angličtiny a jiných světových jazyků, které nemají příliš vážné problémy s pokrytím, mají za výsledek některá doporučení pro úspěšnou realizaci praktických aplikací. Je to například stanovení maximální chybovosti a doby odezvy rozpoznávače, která nesmí být překročena, chceme-li realizovat systém pro titulkování živých zpravodajských pořadů. Zajímavá je také metoda vylepšení práce rozpoznávače tím, že mu zpravodajský pořad přeříkává speciální osoba, která má znalosti jeho slovníku a ví, jak mají vypadat dobře formulované titulky.

Porovnáme-li češtinu s ostatními jazyky, které jsou předmětem zájmu automatického rozpoznávání řeči, vidíme, že čeština jako vysoce ohebný jazyk má největší problémy s pokrytím. Chceme-li někdy dosáhnout úspěšnosti rozpoznávání češtiny srovnatelné s rozpoznáváním angličtiny s 60-tisícovým slovníkem, musíme použít slovník o několika stech tisících slovních tvarů. Práce [6] řeší tento problém tím, že slova rozkládá na jejich kmeny a koncovky. Tak lze dosáhnout vyššího pokrytí se stejným slovníkem i vyšší přesnosti rozpoznávání. Nevýhodou tohoto přístupu však je, že rozložení slov na části zkracuje úseky textu popsané jazykovým modelem. Trigramový model z rozložených slov nemůže popsat ani pravděpodobnost dvou následujících celých slov, pouze buďto začátek slova, jeho koncovku a začátek druhého slova nebo koncovku prvního slova a celé následující slovo. Hlavním výsledkem práce [6] je zjištění, že pro dosažení úspěšnosti rozpoznávání češtiny srovnatelné s angličtinou bude patrně nutné zvětšit velikost slovníku rozpoznávače.

4. Cíle disertační práce

Výzkum popsaný v této práci je založen na předpokladu, že slova jsou přirozené lingvistické jednotky nesoucí sémantickou, syntaktickou a gramatickou informaci zakódovanou do posloupnosti fonémů. Všechny tyto čtyři vlastnosti (sémantická, syntaktická, morfologická a fonologická) spolu úzce souvisí a mohou být reprezentovány lépe na úrovni celých slov než na úrovni slovních částí (morfémů) nebo slovních gramatických kategorií [7]. Abychom dosáhli dobrých výsledků pomocí tohoto způsobu reprezentace, musíme splnit tyto cíle:

1. Příprava textového korpusu
2. Sestavení slovníku obsahujícího několik stovek tisíců slov
3. Fonetická transkripce slov ve slovníku
4. Výpočet různých bigramových jazykových modelů
5. Příprava testovací databáze promluv
6. Testování slovníku, jazykového modelu a parametrů rozpoznávače na databázi promluv
7. Vytvoření kritérií pro měření kvality přepisu

5. Řešení

5.1. Textový korpus

Účelem textového korpusu je získání informace o frekvencích jednotlivých slov a jejich řetězců. Ze seznamu nejfrekventovanějších slov se sestavuje slovník rozpoznávače. Z frekvencí řetězců slov se počítá *n*-gramový jazykový model.

Korpus si sbírají členové SpeechLabu sami. Hlavním zdrojem korpusu jsou různé internetové noviny. V roce 2005 měl korpus velikost 2,6 GB prostého (neformátovaného) textu a stále se zvětšuje.

Původní formát textu je HTML, takže první operace, kterou je třeba udělat, je odstranění HTML značek. Výsledný prostý text se musí vyčistit. Čištění korpusu lze shrnout do následujících bodů [5]:

1. Každá věta je ve finálním korpusu umístěna na jeden řádek. Identifikace vět je automatická a algoritmus, který byl pro tento účel vyvinut obsahuje mnoho pravidel říkajících, která tečka skutečně označuje konec věty.
2. Jednotlivá slova v závorkách jsou potom vymazána. Taková slova jsou totiž většinou z hlediska jazykového modelování nezajímavé zkratky.
3. Opakující se záhlaví, zápatí a formátovací znaky jsou vymazány.
4. Nesklonné zkratky jsou přepsány na celá slova.
5. Výrazy typu *x-letý*, kde *x* je psáno číslicemi, jsou přepsány na celá slova.
6. Každé slovo je převedeno na malá písmenka a každý interpunkční znak je obklopen mezerou, aby bylo možné je počítat a provádět analýzu jejich sousedů.
7. Čísla znamenající hodiny a datumy a některá další čísla jsou přepsány do jejich mluvené podoby. Řadové číslovky, před kterými je předložka, jsou přepsány do jejich gramaticky správného pádu a rodu s pomocí českého morfologického analyzátoru.

8. Slova jsou přepsána do jejich standardní ortografické podoby. Protože mnoho slov, obzvláště těch cizího původu, má alternativní způsoby zápisu, sjednotili jsme jejich ortografii na nejčastější varianty. To trochu zmenšilo slovník rozpoznávače a zvětšilo jeho pokrytí normalizovaného textu. Ručně jsme nalezli 35 000 slovních tvarů, které by měly být přepsány. Tato přepisovací pravidla jsou také aplikována na referenční transkripce promluv určených pro testování rozpoznávače. [7] Tento proces se obvykle nazývá normalizace ortografie.
9. Kolokace, neboli fráze slov, které se často vyskytují vedle sebe, jsou spojeny speciálním znakem, aby s nimi bylo při počítání jazykového modelu zacházeno jako s jedním slovem. V současné době máme v našem slovníku 1 700 kolokací.

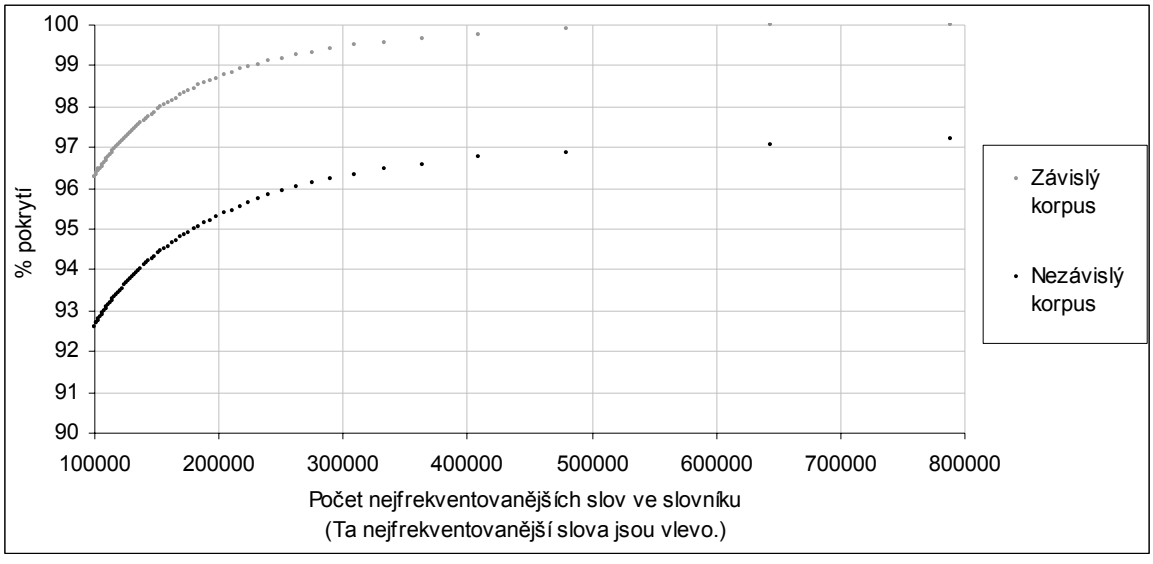
5.2. Slovník

5.2.1. Výběr slov

Spolu s tím, jak jsme zvětšovali slovník našeho rozpoznávače, zabývali jsme se i metodami výběru slov do slovníku a teoretickým odhadem pokrytí velkých slovníků.

Nejdříve jsme porovnali metody výběru slov pomocí statistické analýzy korpusu a pomocí generování slovních tvarů z databáze slovních kmenů a koncovek. Zjistili jsme, že analýza korpusu je sice pracná, protože mnoho slov musí kontrolovat člověk, ale jejím výsledkem je slovník, který má výrazně vyšší pokrytí nezávislého textu než slovník vzniklý generováním slovních tvarů.

Sloučením množin slov vzniklých oběma metodami výběru vznikl poměrně dobře anotovaný slovník 800 tisíc slovních tvarů. Tento slovník jsme využili k analýze pokrytí. Výsledek této analýzy nám pomáhá odhadnout, jak velký slovník máme pro rozpoznávač sestavit, chceme-li, aby měl určité pokrytí. Obrázek č. 2 ukazuje křivku pokrytí pro slovníky s počtem nejfrekventovanějších slov mezi 100 000 a 800 000.



Obrázek č. 2 Pokrytí závislého a nezávislého korpusu [8]

Pokud v angličtině je možné, aby 60-tisícový slovník pokryl 99 % textu, jak se uvádí v kapitole č. 3, tak v češtině ani 800-tisícový slovník takového pokrytí nedosahuje.

Analýza pokrytí nás utvrdila v tom, že je nutné stále rozšiřovat slovník rozpoznávače. Nejnovější verze slovníku pro rozpoznávač spojité řeči z roku 2005 má 312 tisíc slov.

5.2.2. Slučování a rozdělování slov

Článek [7] uvádí, že když do slovníku o 310 tisících slovech přidáme 1 708 nových slov tvořených zřetězeními slov, která ve slovníku už jsou, stoupne přesnost rozpoznání ze 75,5 % na 78,2 %. Tato zřetězení se odborně nazývají kolokace a musí být podchyceny již ve fázi čištění korpusu, viz bod 9 v kapitole č. 5.1.

Udělali jsme také pokus zjišťující, zda-li naopak rozdělení některých slov také nezlepší rozpoznávání. Motivací k tomu byla úvaha, že některá česká slova mají velmi časté předpony *ne-*, *nej-* a *nejne-* a mnoho přídavných jmen obsahuje číslovky, například *dvacetiletý*. Rozložením těchto slov by byl zredukován slovník, který by současně dosahoval vyššího pokrytí nezávislých textů. Výsledkem je zjištění, že dekompozice slov rozpoznání většinou nezlepšuje. Souhrnná přesnost rozpoznávání na databázi televizních i rozhlasových zpráv popsaných v kapitole č. 5.5 je 80,45 % s celými slovy a 79,76 % s dekomponovanými slovy i když procento slov rozpoznávaného testu nenalezených ve slovníku rozpoznávače kleslo z 1,35 % na 1,30 %, když byla slova rozložena (dekomponována).

5.2.3. Závislost přesnosti rozpoznávání na velikosti slovníku

Tabulka č. 1 ukazuje závislost přesnosti rozpoznávání (Accuracy) na počtu slov ve slovníku. „Rank“ je prahová četnost slov v trénovacím korpusu, kterou musela všechna slova v daném slovníku přesáhnout. „OOV“ je procento slov v rozpoznávaném textu, která nebyla nalezena ve slovníku. Všechny menší slovníky byly podmnožinou slovníku „Lex312k“. Testovací databáze promluv jsou 3 kompletní televizní zpravodajské pořady, viz kapitola č. 5.5.

Tabulka č. 1 Rozpoznávací skóre dosažená se slovníky různých velikostí [5]

Jméno slovníku	Počet slov	Rank	OOV [%]	Accuracy [%]
Lex64k	64 620	300	5,17	70,96
Lex102k	102 228	140	3,31	73,75
Lex149k	148 928	70	1,94	75,62
Lex195k	194 932	40	1,34	76,64
Lex257k	257 086	20	0,97	77,27
Lex312k	312 289	10	0,64	78,13

5.3. Fonetická transkripce

Fonetická transkripce vyjadřuje posloupnost zvuků mluvené řeči řetězci textových znaků. V oboru komunikace lidí s počítači se fonetická transkripce uplatňuje ve dvou úlohách. První úlohou je syntéza řeči podle textového vstupu. Možnou aplikací syntézy řeči je například software pro počítače pro slepé lidi a informační služby po telefonu. Druhou úlohou je automatické rozpoznání řeči. V obou případech musí být znaky textu psané formy jazyka (grafémy) namapovány na znaky zastupující zvuky vyslovované při čtení tohoto textu (fonémy). Fonetická transkripce slouží v obou úlohách jako spojka mezi textovou formou a akustickými modely slov.

Při řečové syntéze i rozpoznávání je fonetická transkripce užita ve dvou fázích. Nejdříve se musí trénovací databáze promluv přepsat na text a tento text musí být přepsán foneticky.

Výsledek je využit v procesu segmentace. Při segmentaci se zvukový signál s řečí dělí na segmenty obsahující jednotlivé fonémy. Tyto segmenty jsou potom využity pro trénování akustických modelů fonémů. Počáteční fáze segmentace se musí provádět ručně. Když jsou natrénovány spolehlivější akustické modely, segmentace může být více či méně automatická. Ve druhé fázi je fonetická transkripce aplikována na text, který má vyslovovat počítač, a v případě automatického rozpoznávání řeči se fonetická transkripce využívá k přepisu slov ve slovníku rozpoznávače.

Fonetickou transkripci SpeechLab realizuje pomocí vlastních počítačových programů. První takový program byl v této laboratoři vyvinut v roce 1999. Používal množinu přepisovacích pravidel navrženou pro český jazyk v knížce [9] (str. 101 – 106). Pomocí této množiny pravidel však nebylo přepsáno správě dostatečně velké množství slov. Základní pravidla musíme stále rozšiřovat o výjimky. Ohebnost českého jazyka množství těchto výjimek velmi zvětšuje.

Zkusili jsme také realizovat fonetickou transkripci pomocí neuronové sítě. Naše neuronová síť četla 5 znaků přepisovaného textu a odhadovala foném, na který se má přepsat prostřední znak z těchto 5 znaků. Síť se to nejdříve učila na anotovaném seznamu slov a jejich fonetických transkripcí. Naučená síť přepisovala nová slova. Výsledky dosažené s touto neuronovou sítí byly horší než se systémem založeným na pravidlech. Zatímco systém založený na pravidlech je možné zdokonalovat vylepšováním množiny pravidel, neuronová síť má některá principiální omezení, která například brání, aby se naučila výjimky.

V roce 2004 jsme použili genetické algoritmy pro automatizovaný sběr přepisovacích pravidel pro náš původní systém vyvíjený od roku 1999. V genetickém algoritmu se náhodně modifikují přepisovací pravidla a testují se na anotovaném seznamu slov a jejich fonetických transkripcí. Výstupem jsou pravidla, která úspěšně přepisují slova v seznamu a žádná nezkazí. Takto se podařilo obohatit náš systém pro fonetickou transkripci o nová pravidla.

Náš současný systém pro fonetickou transkripci dává každému slovu jen jedinou variantu výslovnosti. Mnoho slov má však z různým důvodů více variant. Dodatečné varianty zatím ve slovníku rozpoznávače doplňujeme ručně a jejich vliv na přesnost rozpoznávání je značný.

5.4. Jazykový model

5.4.1. Účel a podoba jazykového modelu

Jazykový model reprezentuje znalost o jazyce, která obvykle říká, jak jsou určitá slova pravděpodobná v určitém kontextu.

Ve většině rozpoznávačů spojité řeči jsou využívány takzvané *n*-gramové jazykové modely. *N*-gramový jazykový model je tabulka podmíněných pravděpodobností slova za podmínky, že určitá posloupnost $n - 1$ slov mu v promluvě předcházela. Tyto pravděpodobnosti jsou počítány z velkého trénovacího korpusu. Díky tomu jsou *n*-gramové jazykové modely také nazývány statistické. Při výpočtu jazykového modelu jsou z trénovacího korpusu vybírána pouze slova patřící do slovníku rozpoznávače. Náš rozpoznávač používá bigramové jazykové modely, které jsou speciálním případem *n*-gramových jazykových modelů pro $n = 2$.

5.4.2. Výhody *n*-gramových jazykových modelů

1. *N*-gramové jazykové modely jsou dostatečně flexibilní, aby umožnily rozpoznávání libovolných promluv.
2. Algoritmus výpočtu *n*-gramové statistiky je nezávislý na jazyku.

3. Algoritmus výpočtu a používání *n*-gramového jazykového modelu je relativně jednoduchý, obzvláště v případě velmi malých slovníků.
4. *N*-gramové jazykové modely jsou levo-pravé. To znamená, že předpovídají budoucnost z minulosti. Z toho důvodu jsou dobře integrovatelné s akustickými modely založenými na skrytých markovských modelech, které jsou také levo-pravé.
5. *N*-gramové jazykové modely mohou být snadno přizpůsobeny určité tématické doméně. *N*-gramy patřící do té části korpusu, která se týká požadovaného tématu mohou dostat větší váhu než ostatní *n*-gramy a výsledné dvě skupiny *n*-gramů se mohou snadno sloučit. Výsledky této techniky jsou často překvapivě dobré.
6. *N*-gramové jazykové modely popisující pravděpodobnosti slovních tvarů mohou být snadno kombinovány s *n*-gramovými jazykovými modely popisujícími pravděpodobnosti gramatických kategorií slovních tvarů.

5.4.3. Nevýhody *n*-gramových jazykových modelů

1. *N*-gramové jazykové modely se skládají z velkého počtu parametrů rovného m^n, kde m je velikost slovníku a n je řád *n*-gramového jazykového modelu. Velikost slovníku m musí být obvykle veliká, což znamená, že n musí být obvykle menší než 4.
2. Prakticky použitelné *n*-gramové jazykové modely dokáží popsat jen lokální závislosti slov, ale reálné závislosti v jazyce často sahají dále než přes dvě nebo tři slova.
3. *N*-gramové jazykové modely by měly obsahovat pravděpodobnosti všech možných slovních dvojic nebo trojic. Ale žádný trénovací korpus není tak rozsáhlý, aby obsahoval každou možnou posloupnost dvou nebo tří slov. Například náš 2,6-GB trénovací textový korpus obsahuje 60 228 569 různých slovních párů složených ze slov v našem 312-tisícovém slovníku. Počet slovních párů, jejichž pravděpodobnost musí být v jazykovém modelu určena je ve skutečnosti $312\,000^2$. To znamená, že náš korpus obsahuje pouze 0,06 % všech možných slovních párů. Některé slovní páry, které chybí v korpusu by skutečně měly mít téměř nulovou pravděpodobnost. Některé jiné chybějící slovní páry jsou však z hlediska gramatiky daného jazyka přípustné. Umění jazykového modelování spočívá v rozlišení té jedné skupiny chybějících slovních párů od druhé a určení správných pravděpodobností pro přípustné slovní páry. Tuto úlohu řešíme takzvaným vyhlazováním jazykových modelů popsaným v kapitole č. 5.4.4.
4. *N*-gramové jazykové modely musí být vypočteny z velmi rozsáhlých trénovacích korpusů. Příprava takových korpusů je velmi pracná (viz kapitola č. 5.1), a výpočet frekvencí slovních posloupností musí být realizován nějakým efektivním způsobem.
5. Bigramové jazykové modely pro slovníky přesahující přibližně 10 tisíc slov musí být reprezentovány způsobem, který umožňuje jak jejich vměstnání do operační paměti současných osobních počítačů tak i efektivní využívání informací, které obsahují, během rozpoznávání řeči. Naše řešení tohoto problému je naznačeno na konci kapitoly č. 2.

K těmto nevýhodám *n*-gramových jazykových modelů platících obecně pro každý jazyk je nutno přičíst specifické vlastnosti českého jazyka.

Za prvé, díky ohebnosti češtiny je nutné mít ve slovníku rozpoznávače řádově sta tisíce položek, což znamená, že nemůžeme použít jazykový model vyššího řádu než bigramový.

Za druhé, v češtině je mnoho pravidel o shodě mezi rodem, číslem a pádem, takže velké množství slovních párů se nikdy v gramatické české větě neobjeví. To má za následek velmi řídkou matici četností slovních párů nalezených v trénovacím korpusu. Tato matice musí být

nějak vyhlazena (viz kapitola č. 5.4.4), ale vyhlazení, které nebere v úvahu informaci o české gramatické shodě dává příliš velikou pravděpodobnost bigramům, které mají zůstat nulové.

Za třetí, vedlejším efektem pravidel gramatické shody je relativně volný pořádek slov v české větě. Mluvnická shoda nese informaci o vztazích podmětů, předmětů a přísudků, takže není nutné určitým způsobem slova ve větě řadit. To sice redukuje řídkost jazykového modelu, ale jeho prediktivní schopnost se tím spíše zhoršuje.

5.4.4. Vyhlazování jazykových modelů

Vyhlazování nahrazuje nulové pravděpodobnosti v jazykovém modelu nenulovými pravděpodobnostmi. I když je takto dána nenulová pravděpodobnost slovním vazbám, které nejsou gramaticky povoleny, máme experimentálně ověřeno, že přesnost rozpoznávání se vyhlazením jazykového modelu výrazně zvýší.

Pro vyhlazování používáme metodu zvanou Witten-Bell discounting, viz vzorce (2) a (3), kde $C(w_1, w_2)$ je četnost dvojice slov w_1, w_2 v trénovacím korpusu, $C(w_1)$ je četnost slova w_1, $T(w_1)$ je počet druhů slov, které se v trénovacím korpusu objevily za slovem w_1, a V je velikost slovníku neboli počet všech možných druhů slov.

$$P(w_2|w_1) = \frac{C(w_1, w_2)}{C(w_1) + T(w_1)} \text{ když } C(w_1, w_2) > 0 \tag{2}$$

$$P(w_2|w_1) = \frac{T(w_1)}{(V - T(w_1)) \cdot (C(w_1) + T(w_1))} \text{ když } C(w_1, w_2) = 0 \tag{3}$$

K této metodě jsme přidali dvě pravidla vyjádřená vzorci (4) a (5). Ověřili jsme pomocí výpočtu křížové perplexity jazykového modelu proti textu z testovací databáze, že takto upravené vyhlazování lépe vystihuje předem neznámá data než původní forma Witten-Bell discounting (2) a (3) nebo Witten-Bell discounting kombinované s vyhlazováním, které zvýší četnost všech možných dvojic slov o jednu (takzvané add-one smoothing).

$$P(w_2|w_1) = \frac{C(w_1, w_2) \cdot (C(w_1) + 2T(w_1) - V)}{C(w_1) \cdot (C(w_1) + T(w_1))} \text{ když } 2T(w_1) > V \text{ a } C(w_1, w_2) > 0 \tag{4}$$

$$P(w_2|w_1) = \frac{1}{C(w_1) + T(w_1)} \text{ když } 2T(w_1) > V \text{ a } C(w_1, w_2) = 0 \tag{5}$$

Některé jiné vyhlazovací metody by mohly vést i k lepším výsledkům. Nepoužíváme je proto, že jejich počítačová reprezentace by zpomalila proces rozpoznávání. Souvisí to se způsobem, jak reprezentujeme jazykový model, viz konec kapitoly č. 2.

5.5. Testovací databáze promluv

Výsledky publikované v této práci byly pořízeny na třech testovacích databázích promluv.

První databáze obsahuje 1 600 vět namluvených neprofesionálními mluvčími, kteří četli texty z novin do mikrofonu. Tato databáze má celkem 16 027 slov. Její popis spolu s výsledky byl publikován v článcích [4] a [10].

Druhá databáze je tvořena třemi kompletními zpravodajskými televizními pořady. Tato databáze má celkem 8 451 slov. Její popis spolu s výsledky byl publikován v článku [11].

Třetí databáze byla vytvořena pro tuto práci a bude využívána i v budoucnu. Je tvořena různými rozhlasovými zpravodajskými pořady, je v ní řečeno 13 081 slov 61 mluvčími během

1,5 hodiny. Výsledek základního rozpoznávacího experimentu na této databázi ukazuje Tabulka č. 2.

Tabulka č. 2 Analytický pohled na přesnost (Accuracy) v procentech základního rozpoznávacího experimentu na testovací databázi rozhlasových zpráv (Všechna slova byla převedena na malá písmena. Jazykový model byl z celých slov.)

Rozhlasová stanice	**Pohlaví**	**Styl promluvy**	**Zvukové podmínky**			**Celkem**	**Celkem**
			Čisté	**Nízká věrnost (např. telefon)**	**Hluk na pozadí (řeč, hudba)**		
Radiožurnál	muž	profesionál	87,851	81,053	77,635	84,730	82,827
Radiožurnál	muž	host	76,739	54,869	78,689	69,109	
Radiožurnál	žena	profesionál	90,709	83,099	83,668	89,768	
Radiožurnál	žena	host	83,333	69,748	8,571	58,721	
BBC	muž	profesionál	83,006	.	75,749	81,850	80,589
BBC	muž	host	73,036	67,130	78,298	70,043	
BBC	žena	profesionál	87,681	.	80,465	87,077	
BBC	žena	host	.	.	.	.	
Celkem	muž		81,703	64,193	77,186	76,684	
Celkem	žena		89,295	74,737	78,130	87,700	
Celkem	profesionál		87,618	81,928	79,167	86,349	
Celkem	host		75,586	63,672	70,997	69,094	
Celkem			85,826	65,211	77,529	81,676	81,676

5.6. Ladění parametrů rozpoznávače

Prvním úkolem po naprogramování rozpoznávače bylo nalezení optimální kombinace všech jeho parametrů. Tento úkol je velmi těžký, protože každý parametr ovlivňuje vliv ostatních parametrů. Bylo nutné pátrat v prostoru o několika rozměrech, tudíž provést několik stovek experimentů. Tolik experimentů však není možné provádět na velkých datech, což zase může vést k tomu, že výsledná sada parametrů bude optimální jen pro data, na kterých byla optimalizována. Rozpoznávač používaný ve všech našich experimentech popsaných v této práci je popsán v kapitole č. 2. Následující parametry ovlivňují kvalitu rozpoznávání:

1. *Akustický model* (skryté markovské modely ve formě buďto 16 nebo 32-mixturových monofónů),
2. *Jazykový model*,
3. *LM Factor* (váha jazykového modelu),
4. *Word Insertion Penalty*,
5. *Prune Threshold*,
6. *Number of Word-End Hypotheses*,
7. *Slovník rozpoznávače*.

Pro ladění parametrů jsme využívali množinu 1 600 vět představenou v kapitole č. 5.5. Po několika úvodních experimentech jsme se rozhodli hledat správnou kombinaci výše vyjmenovaných parametrů v prostoru *Jazykového modelu*, parametru *LM Factor* a *Word*

Insertion Penalty. V případě *Akustického modelu* stačí jen malé množství experimentů pro dokázání, že 32-mixturové monofóny dosahují vyšší přesnosti než 16-mixturové monofóny s pouze malým nárůstem spotřebovaného času, viz výsledky v článcích [3] a [4]. Role parametrů *Prune Threshold* a *Number of Word-End Hypotheses* je prořezat strom možných posloupností slov v rozpoznávané větě. Parametr *Prune Threshold* to dělá na úrovni stavů skrytých markovských modelů a parametr *Number of Word-End Hypotheses* to dělá na úrovni slov. O obou parametrech se dá říci, že čím jsou vyšší, tím vyšší je přesnost rozpoznávání, ale každý z nich má jistou úroveň nasycení. Když parametr dosáhne tuto úroveň, přesnost rozpoznávání již stoupá zanedbatelně, ale spotřeba času stoupá stále. *Slovník rozpoznávače* je velmi závislý na testovacích datech. V našich ladících experimentech na množině 1 600 vět jsme používali takzvaný uzavřený slovník. Pojem „uzavřený“ znamená, že v něm byla všechna slova z rozpoznávané databáze (v tomto případě pouze slova z ní). Později jsme začali experimentovat s otevřenými slovníky. V těchto experimentech jsme studovali vliv velikosti slovníku, kolokací, dekomponovaných slov a vícenásobných fonetických transkripcí.

Přechod z uzavřeného slovníku 7 033 slov databáze 1 600 vět, se kterou jsme experimentovali v roce 2002 k otevřenému slovníku 312 000 slov, který jsme začali používat v roce 2005, podstatně změnil ideální kombinaci parametrů nalezenou v roce 2002. Náš velký slovník může být testován pouze na velké testovací množině a takový experiment trvá několik hodin. Takže jsme provedli jen několik ladících experimentů, abychom nalezli nové optimální hodnoty parametrů. Tyto hodnoty ukazuje Tabulka č. 3.

Tabulka č. 3 Porovnání optimální množiny parametrů nalezené v roce 2002 pro uzavřený slovník 7 033 slov [10] s hodnotami parametrů používanými v roce 2005 pro otevřený slovník 312 tisíc slov

Parametr	**7 033 slov**	**312 000 slov**
Jazykový model	Witten-Bell	Witten-Bell
LM Factor	6	7
Word Insertion Penalty	– 5	0
Prune Threshold	130	120
Number of Word-End Hypotheses	10	40

Rozdíly v optimálních hodnotách parametrů, které ukazuje Tabulka č. 3, lze vysvětlit podstatným zvětšením velikosti slovníku. V takovém případě klesá průměrná pravděpodobnost bigramů v jazykovém modelu a tak musí být zvětšena váha jazykového modelu (*LM Factor*). Během rozpoznávání musí být také bráno do úvahy větší množství slov, tudíž se musel podstatně zvětšit parametr *Number of Word-End Hypotheses*, možná také parametr *Word Insertion Penalty*.

5.7. Vyhodnocování rozpoznávání

Standardní míry kvality rozpoznávání jsou založeny na rozdílnostech mezi referenčními transkripcemi a výstupem rozpoznávače. Je možné používat následující vzorce ze zdrojů [3], a [2] (str. 271):

$$\textit{Correctness}\ (\textit{správnost})\ [\%] = 100 \cdot \frac{N - D - S}{N} \tag{6}$$

$$\textit{Accuracy}\ (\textit{přesnost})\ [\%] = 100 \cdot \frac{N - D - S - I}{N} \tag{7}$$

$$\textit{Word Error Rate}\ (\textit{chybovost}) = \textit{WER}\ [\%] = 100 \cdot \frac{D + S + I}{N} = 100 - \textit{Accuracy} \qquad (8)$$

N je celkový počet slov ve správné referenční transkripci.
D (*deletions*) je počet slov, která rozpoznávač vynechal.
S (*substitutions*) je počet slov, která rozpoznávač zaměnil za jiná.
I (*insertions*) je počet slov, která rozpoznávač zapsal navíc.

Hodnoty D, S a I jsou výsledkem algoritmu pro výpočet **minimální editační vzdálenosti** popsané například v knize [2] (str. 153 – 156). Výsledek tohoto algoritmu, suma D, S a I je někdy nazývána **Levenshteinova vzdálenost**. V jedné variantě této míry má každá z těchto tří editačních operací cenu rovnou 1. V jiné variantě má operace vynechání a vložení slova cenu 1 a operace záměny slova má cenu 2, protože je ekvivalentní jednomu vymazání a jednomu vložení slova. Při výpočtu přesnosti (*Accuracy*) rozpoznávání používáme v našich výsledcích tu první variantu.

Jinou důležitou mírou kvality rozpoznávání je spotřeba času. To se vyjadřuje ukazatelem zvaným **real-time factor** *xRT*:

$$xRT = \frac{\textit{Čas spotřebovaný na rozpoznávání}}{\textit{Trvání rozpoznávané promluvy}} \qquad (9)$$

Rychlost rozpoznávání měřená vzorcem (9) je důležitá zvláště pro úlohy, které musí být prováděny v reálném čase, například za účelem opatření živých televizních zpravodajských pořadů titulky. Nevýhodou ukazatele *real-time factor* je jeho závislost na použitém hardwaru. Měl by být vždy doplněn o informaci, na jakém počítači bylo rozpoznávání realizováno.

6. Závěr

6.1. Co bylo v disertační práci vykonáno

Tato práce popsala následující úkoly, které jsou součástí složité úlohy rozpoznávání spojité řeči:

1. Příprava velkého textového korpusu,
2. Sestavení slovníku pro rozpoznávač obsahujícího 312 tisíc slov,
3. Fonetická transkripce slov ve slovníku pro rozpoznávač,
4. Výpočet bigramových jazykových modelů pro rozpoznávač spojité řeči,
5. Příprava 1,5-hodinové testovací databáze promluv,
6. Ladění parametrů rozpoznávače,
7. Vyhodnocování rozpoznávání.

Předmětem těchto úloh byl automatický přepis zpravodajských pořadů v českém jazyce.

6.2. Jaký je přínos této disertační práce pro vědecký obor automatického rozpoznávání řeči

Tato práce popisuje úlohy specifikované v kapitole č. 6.1 podrobněji, než je obvyklé v článcích zabývajících se automatickým rozpoznáváním řeči.

Byl sestaven slovník 800 tisíc nejfrekventovanějších českých slovních tvarů za účelem robustního odhadu pokrytí nezávislého českého textu slovníky o různých velikostech.

Byly popsány tři alternativní metody fonetické transkripce a rozpoznávací experimenty potvrdily důležitost správných fonetických transkripcí ve slovníku rozpoznávače.

Metoda vyhlazování bigramového jazykového modelu zvaná „Witten-Bell discounting“ byla vylepšena a výsledná křížová perplexita takto vyhlazeného jazykového modelu oproti testovacímu korpusu byla skutečně nižší než křížová perplexita počítaná z jazykových modelů vyhlazených alternativními metodami.

Bylo objeveno, že rozložení slov na části ve slovníku a jazykovém modelu rozpoznávače v průměru nevylepšuje přesnost rozpoznávání. Avšak opačný postup – spojování slov, které se často objevují vedle sebe v textu – přesnost rozpoznávání významně vylepšilo.

6.3. Jaký je přínos této disertační práce pro praxi

Výsledky popsané v této práci – textový korpus, slovník, fonetická transkripce, jazykový model a testovací databáze promluv – slouží jako východisko k budoucímu zdokonalování rozpoznávače spojité řeči v laboratoři SpeechLab.

Práce ukázala, že v ní popsané metody řešení mohou pravděpodobně vést ke komerčně využitelným aplikacím pro přepis zpravodajských pořadů v českém jazyce.

6.4. Co by mělo být vykonáno v budoucnu

Zvyšování přesnosti rozpoznávání je výsledkem práce v mnoha oborech. Zde uvádíme pouze úlohy týkající se lingvistické části problému.

1. Mělo by být nalezeno více pravidel pro čištění textového korpusu, zejména pravidla pro přepisování čísel.
2. Tabulka č. 1 v kapitole č. 5.2.3 ukazuje, že přidání jednoho nebo dvou set tisíc nových slov do našeho slovníku pro rozpoznávač, který má zatím 312 tisíc slov, by stále ještě mohlo vylepšit přesnost rozpoznávání.
3. Náš systém pro fonetickou transkripci by měl obsahovat více pravidel a mít je lépe organizována. Měl by být také schopen generovat alternativní fonetické transkripce.
4. Jazykové modely pro náš rozpoznávač byly vždy omezeny kapacitou aktuálně dostupného hardwaru osobních počítačů. Vývoj hardwaru je však tak rychlý, že bychom se měli snažit navrhovat nové jazykové modely implementovatelné na budoucím hardwaru již dnes. Nejslibnějším přístupem podle našeho názoru je vyvíjet jazykové modely, které by se automaticky přizpůsobovaly tématu právě rozpoznávané řeči.
5. Výstup rozpoznávače by měl být co nejbližší psané podobě jazyka. To zahrnuje správné psaní velkých a malých písmen a interpunkce. Také některé číslovky by měly být psány číslicemi a některé slovy. Pro vyřešení tohoto problému musí být sestaveny speciální jazykové modely.

Literatura

[1] Jan Nouza (editor): Počítačové zpracování řeči, cíle, problémy a aplikace. (Sborník článků). Technická univerzita v Liberci. Fakulta mechatroniky a mezioborových inženýrských studií. Katedra elektroniky a zpracování signálů – Laboratoř počítačového zpracování řeči. Liberec 2001, ISBN 80-7083-551-6.

[2] Daniel Jurafsky, James H. Martin: Speech and Language Processing. Prentice Hall, Inc., New Jersey, 2000, ISBN 0-13-095069-6.

[3] Dana Nejedlová, Jan Nouza: Language Model Support for Continuous Speech Recognition in Czech Language. In: Signal Processing, Pattern Recognition, and Application, Anaheim (USA), Calgary (Kanada), Curych (Švýcarsko) 2002, ISBN 0-88986-338-5, str. 541 – 546, ISSN 1482-7921.

[4] Jan Nouza: Strategies for Developing a Real-Time Continuous Speech Recognition System for Czech Language. In: Text, Speech and Dialogue (eds. Petr Sojka, Ivan Kopeček, Karel Pala) Springer-Verlag, Heidelberg, 2002, str. 189 – 196, ISBN 3-540-44129-8, ISSN 0302-9743.

[5] Dana Nejedlová, Jindra Drábková, Jan Kolorenč, Jan Nouza: Lexical, Phonetic, and Grammatical Aspects of Very-Large-Vocabulary Continuous Speech Recognition of Czech Language. In: Electronic Speech Signal Processing, Proceedings of the 16th Conference on Electronic Speech Signal Processing joint with the 15th Czech-German Workshop on Speech Processing, Dresden, Německo, září 2005, TUDpress, str. 224 – 231, ISBN 3-938863-17-X, ISSN 0940-6832.

[6] Pavel Ircing: Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language (Czech). [Disertační práce] Západočeská univerzita v Plzni. Fakulta aplikovaných věd. Plzeň 2003.

[7] Jan Nouza, Jindřich Žďánský, Petr David, Petr Červa, Jan Kolorenč, Dana Nejedlová: Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon. In: Interspeech, Lisabon-Portugalsko, 2005, ISCA, Bonn, Německo, str. 1681 – 1684, ISSN 1018-4074.

[8] Dana Nejedlová: Building and Evaluation of a Large Vocabulary for a Czech Voice Dictation System. In: ECMS (The 6th International Workshop on Electronics, Control, Measurement and Signals), Liberec, 2003, str. 74 – 78, ISBN 80-7083-708-X.

[9] Josef Psutka: Komunikace s počítačem mluvenou řečí, Academia, Prague 1995, ISBN 80-200-0203-0.

[10] Dana Nejedlová: Comparative Study on Bigram Language Models for Spoken Czech Recognition. In: Text, Speech and Dialogue (eds. Petr Sojka, Ivan Kopeček, Karel Pala) Springer-Verlag, Heidelberg, 2002, str. 197 – 204, ISBN 3-540-44129-8, ISSN 0302-9743.

[11] Jan Nouza, Dana Nejedlová, Jindřich Žďánský, Jan Kolorenč: Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs. In: ICSLP (eds. Soon Hyob Kim and Dae Hee Youn), Sunjin Printing Co., 2004, str. 409 – 412, ISSN 1225-441x.

Vlastní publikované práce

Dana Nejedlová, Jindra Drábková, Jan Kolorenč, Jan Nouza: "Lexical, Phonetic, and Grammatical Aspects of Very-Large-Vocabulary Continuous Speech Recognition of Czech Language". Prezentováno na 16. konferenci „Electronic Speech Signal Processing" spojené s „15th Czech-German Workshop on Speech Processing" Ústavu radiotechniky a elektroniky Akademie věd České republiky v Lichtenštejnském paláci v Praze 27. září 2005. In: Electronic Speech Signal Processing, Proceedings of the 16th Conference on Electronic Speech Signal Processing joint with the 15th Czech-German Workshop on Speech Processing, Drážďany, Německo, Září 2005, TUDpress, str. 224 – 231, ISBN 3-938863-17-X, ISSN 0940-6832.

Jan Nouza, Jindřich Žďánský, Petr David, Petr Červa, Jan Kolorenč, Dana Nejedlová: "Fully Automated System for Czech Spoken Broadcast Transcription with Very Large (300K+) Lexicon". In: proc. 9th European Conference on Speech Communication and Technology Interspeech 2005 (CD-ROM), Lisabon, Portugalsko, 2005, ISCA, Bonn, Německo, str. 1681 – 1684, ISSN 1018-4074.

Jan Nouza, Dana Nejedlová, Jindřich Žďánský, Jan Kolorenč: "Very Large Vocabulary Speech Recognition System for Automatic Transcription of Czech Broadcast Programs". In: proc. 8th International Conference on Spoken Language Processing ICSLP 2004 (editoři: Soon Hyob Kim and Dae Hee Youn) (4 svazky a CD-ROM), Jeju Island, Korea, říjen 2004, Sunjin Printing Co., str. 409 – 412, ISSN 1225-441x.

Dana Nejedlová: "Lexicon and Language Model Building for Czech Very-Large-Vocabulary Speech Recognition". Prezentováno na „The 14th Czech-German Workshop on Speech Processing" Ústavu radiotechniky a elektroniky Akademie věd České republiky v Karlově univerzitě v Praze 14. září 2004. In: Speech Processing, 14th Czech-German Workshop, Praha 2004, str. 82 – 92, ISBN 80-86269-11-6.

Dana Nejedlová: "Construction of a Dictation System for Czech Physicians". Prezentováno na „The 13th Czech-German Workshop on Speech Processing" Ústavu radiotechniky a elektroniky Akademie věd České republiky v Karlově univerzitě v Praze 16. září 2003. In: Speech Processing, 13th Czech-German Workshop, Praha 2004, str. 115 – 117, ISBN 80-86269-10-8.

Dana Nejedlová, Jan Nouza: "Building of a Vocabulary for the Automatic Voice-Dictation System". Prezentováno na 6. mezinárodní konferenci TSD 2003 v Českých Budějovicích 9. září 2003. In: Text, Speech and Dialogue (editoři: Václav Matoušek, Pavel Mautner) Springer-Verlag, Heidelberg, 2003, str. 301 – 308, ISBN 3-540-20024-X, ISSN 0302-9743.

Dana Nejedlová: "Building and Evaluation of a Large Vocabulary for a Czech Voice Dictation System". Prezentováno na „The 6th International Workshop on Electronics, Control, Measurement and Signals – ECMS 2003" 3. června 2003. In: ECMS 2003, Liberec, June 2003, str. 74 – 78, ISBN 80-7083-708-X.

Dana Nejedlová: "Building a 20K Vocabulary and Language Model for Czech Language". In: Speech Processing, 12th Czech-German Workshop, Praha 2002, str. 67 – 70, ISBN 80-86269-09-4.

Dana Nejedlová: “Comparative Study on Bigram Language Models for Spoken Czech Recognition”. Prezentováno na 5. mezinárodní konferenci TSD 2002 v Brně 9. září 2002. In: Text, Speech and Dialogue (editoři: Petr Sojka, Ivan Kopeček, Karel Pala) Springer-Verlag, Heidelberg, 2002, str. 197 – 204, ISBN 3-540-44129-8, ISSN 0302-9743.

Dana Nejedlová, Jan Nouza: “Language Model Support for Continuous Speech Recognition in Czech Language”. Prezentováno na „The IASTED International Conference SPPRA 2002” v Řecku na ostrově Kréta 27. června 2002. In: Signal Processing, Pattern Recognition, and Application, Anaheim (USA), Calgary (Kanada), Curych (Švýcarsko) 2002, str. 541 – 546, ISBN 0-88986-338-5, ISSN 1482-7921.

Jan Nouza, Dana Nejedlová: “Experiments with Read Speech Recognition in Czech”. Prezentováno na „The 11th Czech-German Workshop on Speech Processing“ Ústavu radiotechniky a elektroniky Akademie věd České republiky v Karlově univerzitě v Praze 18. září 2001. In: Speech Processing, 11th Czech-German Workshop, Praha 2001, str. 46 – 49, ISBN 80-86269-07-8.

Dana Nejedlová, Marek Volejník: „Transkripce psaného českého textu do fonetické podoby“. In: Počítačové zpracování řeči – cíle, problémy, metody a aplikace (symposium), Technická univerzita v Liberci, Liberec 2001, str. 10 – 22, ISBN 80-7083-551-6.

Dana Nejedlová, Jan Nouza: “Phonetic Transcription of Czech Language Using a NETtalk-type Neural Network”. Prezentováno na „The 10th Czech-German Workshop on Speech Processing“ Ústavu radiotechniky a elektroniky Akademie věd České republiky v Karlově univerzitě v Praze 20. září 2000. In: Speech Processing, 10th Czech-German Workshop, Prague 2000, str. 37 – 40, ISBN 80-86269-05-1.