Univerzita Hradec Králové
Fakulta informatiky a managementu
Katedra informačních technologií

Statistická analýza dat získaných technologií language identification

(Zvýšení úspěšnosti systému pro určení jazyka pomocí statistického přístupu)

Bakalářská práce

Autor: Pavel Komeštík

Studijní obor: Aplikovaná informatika - kombinovaná

Vedoucí práce: RNDr. Josef Dolejš, Ph.D.

Hradec Králové Únor 2015

Prohlášení:

Prohlašuji, že jsem bakalářskou práci zpracoval samostatně a s použitím uvedené literatury.

vlastnoruční podpis

V Hradci Králové dne 21.4.2015 Pavel Komeštík

Poděkování:

Děkuji vedoucímu bakalářské práce RNDr. Josefu Dolejšovi, Ph.D. za metodické vedení práce a velmi aktivní přístup při řešení vzniklých problémů.

Anotace

Název: Statistická analýza dat získaných technologií language identification

Bakalářská práce se zabývá postupy vedoucími ke zvýšení úspěšnosti identifikace jazyka mluvčího v neznámé skupině nahrávek. V teoretické části jsou popsány základní principy technologie pro identifikaci jazyka a dalších podpůrných technologií. Dále pak tato práce popisuje strukturu jednotlivých softwarových částí systémů pro identifikaci jazyka a jejich účel. V praktické části je popsán způsob automatického výběru jazyků do úzké testovací sady a porovnání úspěšnosti identifikace jazyka napříč různými nastaveními systému. V této hlavní části práce autor ukazuje možné cesty vedoucí ke zvýšení úspěšnosti identifikace jazyka. V závěrečných kapitolách jsou prezentovány výsledky výzkumu a popsány přínosy a doporučení při jejich praktickém využití.

Annotation

Title: Statistical analysis of data obtained by language identification technology

The Bachelor thesis deals with procedures leading to an increase in the success rate of speaker language identification in an unknown set of recordings. The theoretical part describes the basic principles of technology for language identification and other assistive technologies. Furthermore, this work describes the structure of an individual software components of language identification systems and their purpose. The practical part shows method for automatic selection of languages to close language set and a comparison of the success rate of language identification system across different settings. In this main part of the work the author shows a possible way to increase the success rate of language identification system. The final chapters are presenting results of research and describe the benefits and suggestions for their practical use.

Obsah

Seznam grafů

Seznam tabulek

1 Úvod

Problematika automatického určení jazyka mluvčího se v dnešní době stává nedílnou součástí technologie v širokém slova smyslu. Řečové technologie zažívají prudký rozvoj a jistě není daleko doba, kdy se stanou součástí našeho každodenního života.

Již dnes se můžeme setkat s výsledky mnohaletého vývoje v této oblasti například ve svých mobilních telefonech, automobilech či v bankách. Široce jsou pak tyto technologie využity v mnoha vědních odvětvích, jakými jsou například robotika, medicína či vývoj umělé inteligence.

Aby bylo možné ovládat zařízení hlasem, nebo dokonce komunikovat s nějakou formou umělé inteligence, je vždy nutné stanovit formální stránku takové interakce. V případě přirozené řeči se vlastně jedná o určení jazyka, v němž je komunikace vedena. Až na základě této znalosti je možné interpretovat hlasový záznam pomocí znaků, slov a vět a pochopit tak význam sdělení. Dosáhnout identifikace jazyka mluvené řeči, je možno několika způsoby. V dnešní době se ještě setkáme s nejjednodušší technikou tohoto určení například v automobilech, či v mobilních telefonech a jejich aplikacích. Zde je komunikační jazyk nastaven buď výrobcem zařízení, nebo pomocí uživatelského rozhraní. Jistě si však dokážeme představit systémy, v nichž toto řešení není dostatečně elegantní či dokonce vůbec dostupné. Tvůrci takového systému pak nezbývá, než se spolehnout na technologii, která se tak stává jedním ze základních stavebních prvků autonomních systémů, na automatickou identifikaci jazyka.

Současná technologie pro rozpoznání jazyka mluvčího prošla již dlouhým vývojem. Jejím základem je teoretická znalost jazykových charakteristik, z nichž vyplývají rozdíly následně použitelné pro rozlišení jednotlivých jazyků.

Základními metodami pro určení jazyka mluvčího jsou podle Lee (2008) přístupy založené na rozpoznání jednotlivých fonémů (nejmenší součást zvukové

stránky řeči) v mluvené řeči, dále pak systémy pro modelováním jazyka za pomoci spektrální analýzy signálu a nakonec systémy pracující na základě posloupností jednotlivých tokenů (konkrétní realizace slova) v řeči. Dnes nejčastěji doplňují tyto základní technologie systémy založené na reprezentaci jazyka ve vektorovém prostoru, které jsou použity pro konečné rozhodnutí o výsledku. Stále více využívané jsou dnes také různé kombinace výše zmíněných přístupů.

Přes veškerý vývoj, se však stále jedná o mladou technologii, která má svá omezení. Pohybujeme-li se tedy v problémové doméně, v níž je nutné řešit rozpoznání jazyka mluvčího automaticky, pak si toho musíme být vědomi. Následující text se bude zabývat právě těmito omezeními a zkoumat možnosti, jak se s nimi nejlépe vypořádat při praktickém použití dostupných technologií. K praktickým testům byl použit software založený na reprezentaci jazyka ve vektorovém prostoru a následném diskriminativním modelu skórování. Z tohoto důvodu jsou některé části následujícího textu zaměřeny právě na tuto techniku rozpoznání jazyka a týkají se zmíněné realizace softwaru.

2 Cíl práce

Hlavním cílem této bakalářské práce je zvýšení úspěšnosti určení jazyka s využitím automatického odhadu jazykového spektra skupiny nahrávek.

3 Teoretická východiska

Kapitola byla zpracována s využitím Psutka, Müller, Matoušek, Radová (2006).

3.1 Historie řečových technologií

Řečové technologie byly středem zájmu některých vědců i nadšenců již od druhé poloviny 18. století. První kroky k realizaci myšlenky zapojení stojů do běžného lidského hovoru pochází přibližně z tohoto období, kdy byly popsány experimenty s mechanickým syntetizérem lidského hlasu. Hlavní pokrok proběhl v souvislosti s nástupem číslicových počítačů a výpočetní techniky všeobecně. Prostředky umožňující praktické využití teoretických postupů poskytly až výkonné signálové procesory, které se objevily v první polovině 90. let. Od této doby bylo možné pracovat s řečovými technologiemi v reálném čase. O podobné nástroje projevovaly velký zájem pochopitelně i různé vládní organizace, jako například agentura výzkumných projektů Ministerstva obrany USA (angl. Defense Advanced Research Projects Agency). Pod jejím financováním vznikl v letech 1971 až 1976 také projekt DARPA-SUR (Druhá část názvu projektu vychází z anglického spojení Speech Understanding Research), který měl za cíl vývoj systému, který by rozuměl souvislé lidské řeči. Z dnešního hlediska můžeme takovéto plány hodnotit jako velmi smělé, protože při vývoji podobných systému narazily vědci na překážky, které prozatím nedovolují vyvinout technologii, která by takový cíl bezezbytku naplňovala. Nicméně dnešní stupeň poznání tohoto oboru nám již dovoluje prakticky využívat alespoň některé části, z kterých by mohla být nakonec technologie pro úplné porozumění souvislé řeči vytvořena.

3.2 Technologie pro rozpoznání jazyka mluvčího (LID)

3.2.1 **Teorie technik rozpoznání jazyka**

LID je všeobecně užívanou zkratkou technologie pro rozpoznání jazyka mluvčího. Jedná se o akronym anglického výrazu „language identification". Celá tato práce se zabývá právě touto technologií a jejími možnostmi při praktickém použití. V této části se seznámíme s teoretickými možnostmi rozlišení světových jazyků a z toho vyplývající strukturou systémů pro identifikaci jazyka mluvčího.

Techniky automatického rozpoznání jazyka mluvčího jsou založeny na několika různých přístupech. Jak uvádí Matějka (2008), jedná se především o následující:

- **Fonetika**

„Ačkoli je lidská řeč teoreticky schopna vyprodukovat obrovské množství různých zvuků, existuje v každém jazyce pouze omezené množství opakujících se, poměrně rozdílných jednotek řeči (hlásek/fonémů). Velké množství jazyků sdílí jakýsi "běžný" soubor fonémů. Frekvence použití fonémů se však mohou lišit. Stejný foném vyskytující se ve dvou jazycích tak může být v jednom z jazyků využíván častěji, než ve druhém. Počet fonémů v jazyce se pohybuje mezi 15 - 50. Většina jazyků má přibližně 30 fonémů.“

- **Fonotaktika**

„Jednotlivé světové jazyky se však neliší jen inventáři fonémů, ale také jejich vzájemnými kombinacemi a sekvencemi. Některé kombinace fonémů běžné v jednom jazyce mohou být v jiném nepřípustné. Fonotaktika se tedy zabývá vyhodnocováním těchto posloupností fonémů.“

- **Prozódie**

„Prozódie je termín pro melodii a rytmiku jazyka. Jazyky mají charakteristické zvukové vzory, které mohou být analyzovány co do délky fonémů, rychlosti řeči, intonace a důrazu.“

- **Morfologie - slovník**

„Konceptuálně je největším rozdílem mezi jednotlivými jazyky použití různých souborů slov tzn., jejich slovníky se liší. Nerodilý mluvčí, mluvící anglicky, používá fonémy a prozódii své rodné řeči, ale užití anglického slovníku rozhoduje o tom, že jazyk jeho řeči je vyhodnocen jako angličtina.“

- **Syntaxe**

„Dalším rozlišovacím faktorem může být způsob, kterým jsou spojována jednotlivá slova do vět. Pokud dva jazyky sdílejí stejné slovo, například "bin" v němčině a v angličtině, pak se mohou lišit sady slov, které mohou za tímto slovem následovat.“

Tyto teoretické znalosti o možnostech rozlišení jednotlivých jazyků nám ukazují cesty, jak přistupovat k tvorbě systémů pro rozpoznání jazyka a napomáhají nám přibližně stanovit jejich strukturu.

3.2.2 Struktura systémů pro rozpoznání jazyka

Existuje mnoho různých přístupů k tomu, jak prakticky realizovat software pro rozpoznávání jazyka mluvčího. V zásadě, se však proces rozpoznání jazyka, podle Matějky (2008), skládá z následujících kroků.

- **Extrakce příznaků**

„Řečový signál je převeden do série vektorů, které by měli obsahovat pouze ukazatele důležité pro rozpoznání jazyka. Důležitým prvkem extrakce je potlačení informací, které jsou pro učení jazyka irelevantní. V současnosti je nejpopulárnější metoda Mel Frequency Cepstral Coefficients (MFCC) a její modifikace Shifted Delta Cepstra (SDC).“

- **Klasifikace**

„Buď přímo do finálních tříd (cílové jazyky), nebo do smysluplných skupin, které jsou použity pro další statistické vyhodnocování. Zde se využívají různé druhy klasifikátorů, jako jsou Gaussian Mixture Models (GMM), Neural Networks (NN), Support Vector Machines (SVM), Radial Basis Functions (RBF), a další.“

- **Statistické modely**

„V některých zapojeních LID je tento subsystém používán k statistickému modelování jednotek vytvořených klasifikací. Konvenční cestou modelování je jazykový model (n-gramy, binární rozhodovací stromy) nebo SVM.“

- **Fúze**

„Možné spojení s jiným systémem implementované často prostřednictvím "Linear Logistic Regret ion" (LLR), NN, GMM.“

- **Volba úrovně přijetí výsledku a rozhodnutí**

„Je použita pro konečné rozhodnutí, zda bude výsledek určení přijat.“

Většina dnešních rozpoznávačů dodržuje až na drobnosti toto vnitřní schéma a liší se především implementací jednotlivých kroků. Z pohledu této práce je zajímavá především poslední část (volba úrovně přijetí výsledku a rozhodnutí). Této úrovně tzv. thresholdu se bude týkat významná část práce a je tedy namístě upozornit, že její volba je součástí téměř každého systému pro rozpoznání jazyka mluvčího.

3.2.3 Základní rozdělení typů detektorů jazyka

Jedním ze základních způsobů rozpoznání jazyka mluvčího, je systém založený na rozeznávání jednotlivých fonémů (nejmenší součást zvukové stránky řeči) v mluvené řeči. Schwarz (2008) uvádí, že fonémové rozpoznávače jsou součástí nejrůznějších systémů pro automatické zpracování řeči. Tyto rozpoznávače tak mohou být základem nejen systémů pro rozpoznání jazyka mluvčího, ale také pro systémy vyhledávající v nahrávkách klíčová slova, či detekující téma hovoru. Přesnost všech těchto systémů je pak závislá na přesnosti takovéhoto fonémového rozpoznávače. Podle Schwarze (2008) se fonémové rozpoznávače zjednodušeně skládají ze tří oddělených bloků. První blok je zodpovědný za extrakci příznaků, při které dochází k potlačení nedůležitých charakteristik řeči a ke kompresi charakteristik důležitých pro daný úkol. Druhý blok následně provede porovnání extrahovaných příznaků s uloženými vzorky. Třetí blok tzv. dekodér hledá nejlepší cestu (pořadí) mezi jednotlivými akustickými jednotkami. Výstupem systému je pak rozpoznaný text.

Existují zde také různá vylepšení, která se snaží omezovat různé typy chyb při identifikaci jazyka. Například autor článku Zhang (2014) uvádí, že použitím

metody „gap-weighted subsequence kernel" dosáhli při evaluacích vyšší úspěšnosti identifikace, než je tomu u standardního modelu fonémového rozpoznávače následovaného modely vektorového prostoru.

Další způsob rozpoznání jazyka mluvčího využívají systémy založené na spektrální analýze signálů. Zde se jedná o porovnání frekvenčních charakteristik jednotlivých řečových segmentů s uloženými vzorky. Reynolds (2008) uvádí, že metody založené na spektrální analýze signálu jsou buď generativní, založené na „*Gaussian mixture model*" (GMM), či diskriminativní, prostřednictvím „*support vector machines"* (SVM).

Přístup založený na spektrální analýze se však dnes používá spíše u systémů pro identifikaci mluvčího, protože umožňuje sledovat fyzické charakteristiky mluvčího, jako jsou například dyšný hlas, chraptění, hypernazalita a podobně. U systémů LID je tento způsob používán podpůrně.

Nicméně v článku Siniscalchi (2012) autor popisuje univerzální přístup k spektrální identifikaci jazyka mluvčího za pomoci setu základních jednotek, které jsou přítomné ve všech jazycích. Jedná se o charakterizaci jednotlivých jazyků pomocí způsobu a umístění artikulačních atributů.

Nedílnou součástí detektorů jazyka jsou systémy pro interpretaci a porovnání výsledků. V dnešní době je v tomto ohledu velmi populární interpretovat zjištěné příznaky ve formě vektorů a tyto posléze porovnávat s vektory vypočtenými pro jednotlivé světové jazyky. Systémy využívající takovýto přístup jsou pak označovány za i-vektorové LID.

3.3 Identifikace řečových segmentů (VAD)

Aby bylo možné použít výše uvedené znalosti k vzájemnému rozlišení jednotlivých jazyků, je nejprve nutné určit, které části nahrávky jsou vlastně lidskou řečí. Právě tímto problémem se zabývají systémy pro detekci řeči v digitálních zvukových záznamech takzvané „Voice Activity detectors" (dále jen VAD). V následujícím textu

se tedy budeme zabývat charakteristikami lidské řeči a způsoby využití těchto charakteristik k odlišení lidské řeči od ostatních částí digitální nahrávky.

3.3.1 Tvorba řeči

Adamec (2008) uvádí poznatek Sigmunda z téhož roku takto:
„ Zdrojem všech znělých zvuků jsou kmitající hlasivky, které jsou umístěny v horní části hrtanu. Vzduch dodávaný plícemi prochází prostorem mezi hlasivkami, tzv. hlasivkovou štěrbinou (latinsky glottis). Hlasivky se rozkmitávají (uzavírají a otevírají hlasivkovou štěrbinu), a tím přeměňují proud vzduchu na pravidelný budící signál. Při kmitání hlasivek vznikají vzduchové rázy v intervalech přibližně 10 ms. Frekvence kmitů závisí jednak na tlaku vzduchu a jednak na svalovém napětí hlasivek. Frekvence kmitů hlasivek F_0 charakterizuje základní tón lidského hlasu.“

Tento základní tón se pak liší u věkových skupin a pohlaví a vnímáme jej jako výšku hlasu osoby. Adamec (2008) dále uvádí, že hodnoty frekvence základního tónu hlasu se pohybují mezi 50 – 400 Hz.

Díky těmto a dalším souvisejícím poznatkům o tvorbě a parametrech lidské řeči je možno nalézat způsoby její detekce v digitálních záznamech pomocí různých typů detektorů.

3.3.2 Struktura systémů pro rozpoznání řeči

Podle Adamce (2008) používá většina těchto detektorů obdobný postup při zpracování digitálních nahrávek.

1. *„Vstupní signál se rozdělí na časové rámce tzv. segmenty.“*
2. *„Stanoví se potřebné charakteristiky signálu. Například energie či kepstrum.“*
3. *„Vypočtená charakteristika se v každém rámci porovná s prahovou hodnotou.“*
4. *„Je-li daná charakteristika signálu v rámci větší než prahová, je segment označen jako řečový, jinak je označen jako pauza v řeči.“*

Výsledkem zpracování je tedy soubor časových segmentů, které jsou označeny informací o tom, zda se jedná o řečový rámec, či o pauzu. Tyto informace pak slouží

jako základ dalším technologiím pro zpracování hlasu, jako jsou například identifikace jazyka mluvčího, či přepis řeči na text.

3.3.3 Základní rozdělení typů detektorů řeči

Adamec (2008) dělí základní typy detektorů řeči podle principu na energetický detektor, kepstrální detektor, detektor intenzity, detektory založené na statistickém modelování a skrytých Markovových modelech (HMM) a nakonec na detektor podle standardu ITU-T G.729.

Nicméně, stejně jako u technologií pro identifikaci jazyka mluvčího, je i zde možné narazit na různé kombinace základních principů a metod. Základní metody identifikace řečových segmentů v digitálním záznamu vykazují vysokou úspěšnost především za ideálních podmínek. Tzn. u nahrávek s vysokým odstupem signálu od šumu. Ovšem pokud jsou vstupní data (nahrávky hlasu) pořízena v běžném životě, pak je jejich kvalita velmi vzdálena nahrávkám studiovým. Z tohoto důvodu vzniklo několik vědeckých prací, které se zabývají hledáním vhodných kombinací metod a postupů, vedoucích ke zvýšení úspěšnosti detekce řeči v nahrávkách s nízkým odstupem signálu od šumu.

Například Wu, Ji a Zhang Xiao-Lei (2011) uvádí, že kombinací dvou sub-algoritmů (rule-based energy detection algorithm a GMM-based multiple-observation log likelihood ratio algorithm) je možné dosáhnout lepších výsledků segmentace, než dosahuje každá tato metoda odděleně, hlavně pak v prostředí s vysokým okolním ruchem.

Stejně tak v kapitole autorů Tan, Ying-Wei a Liu, Wen-Ju (2014) nacházíme spojení dvou metod (short-term and long-term spectral patterns) za účelem dosažení lepších výsledků při segmentaci řečových dat. Z jejich práce vyplývá, že uvedená kombinace je vhodná zejména pro segmentaci nahrávek s nízkým odstupem signál šum.

4 Metodika šetření

4.1 Širší výzkumné cíle

Faktorů ovlivňujících výslednou úspěšnost identifikace jazyka nahrávky existuje mnoho. Tato práce se však nezabývá samotnými principy takovéto identifikace, ale aspekty technologie, které je možné ovlivnit z pozice běžného uživatele. Sem spadá hlavní cíl práce, výběr jazyků pro úzkou jazykovou sadu.

Dále se práce bude týkat praktického využití informací získaných z porovnání výsledků identifikace jazyka nahrávky tzv. expertním posudkem a systémem pro automatickou identifikaci. Počítáme-li totiž s možností že v budoucnosti bude třeba vyhodnocovat data ze stejného základního souboru, pak je možné využít získaných informací jako podkladu pro analytická rozhodnutí týkající se nastavení systému.

4.2 Výzkumné otázky

Je možné stanovit metodologii pro vytvoření úzké jazykové sady, pokud není známo skutečné jazykové spektrum dat, která bude systém vyhodnocovat?

4.3 Pracovní hypotézy

Pracovní hypotézou je domněnka, že i když při použití široké jazykové sady je zpracování méně přesné, tak je tato přesnost stále postačující pro odhad jazykového spektra v testované skupině nahrávek a tedy, že při použití níže uvedeného postupu je možné automaticky stanovit jazykové spektrum a určit tak úzkou jazykovou sadu. Takto určená jazyková sada bude při zpracování vykazovat lepší výsledky, než sada široká. Na základě této hypotézy by bylo možné zúžení sady jazyků podle výsledků předchozího zpracování širokou sadou. Zde vycházíme také z principu, že zanedbání jazyka, který je v testované skupině nahrávek minoritní může mít v praxi pozitivní vliv na celkovou úspěšnost zpracování velkého množství dat.

4.4 Sběr informací

Pro potvrzení testované hypotézy bylo nezbytné pořízení široké skupiny nahrávek a jejich expertní ohodnocení, při kterém byl určen jazyk každé konkrétní nahrávky. Jako základní soubor těchto dat posloužily nahrávky call centra z jednoho odběrového média (kanálu) společnosti. Z těchto dat bylo na základě generátoru náhodných čísel vybráno 1670 nahrávek. Tyto nahrávky byly předkládány jazykovým expertům společnosti, kteří stanovili jazyk každé z nich.

Z vybraných nahrávek byly následně sejmuty jejich i-vektorové reprezentace použité při testování.

4.5 Metoda ověřování hypotézy

Získané i-vektorové reprezentace nahrávek byly zpracovány za použití široké jazykové sady, v níž je obsaženo 42 světových jazyků. Z výsledných dat byla určena křivka přesnosti a křivka úplnosti určení jazyka a zjištěny F-míry pro jednotlivé úrovně thresholdu. Z výsledků testování širokou sadou jazyků byla následně určena úzká sada, obsahující 11 jazyků, a opět byly určeny výsledné křivky a F-míry. Nakonec byla data zpracována jazykovou sadou odpovídající skutečnému jazykovému spektru (12 jazyků) se stejným dohodnocením.

Určení úzké jazykové sady probíhá tak, že z jazykového spektra získaného zpracováním širokou sadou vybereme jazyky, které jsou zastoupeny více než 3 % nahrávek. Tato hranice byla stanovena iterativně a bylo empiricky zjištěno, že na testovaných datech vykazuje jazyková sada vybraná na základě této hranice lepší úspěšnost, než sady vybrané na základě hranice vyšší či nižší.

Nižší hranice znamená více vybraných jazyků a tím nižší přesnost zpracování, protože budou vybrány i minoritně zastoupené jazyky či jazyky, které se v testovaných nahrávkách nevyskytují vůbec. Na druhou stranu vyšší hranice znamená možné zamítnutí významně zastoupeného jazyka a tím také významné snížení úspěšnosti zpracování.

Threshold je úroveň zamítnutí výsledku zpracování. Pokud je tedy skóre nahrávky s vítězným jazykem vyšší, než je hodnota thresholdu, pak výsledek zpracování přijímáme (nahrávku považujeme za určenou). Pokud je naopak nižší, pak výsledek zpracování zamítáme (nahrávku považujeme za neurčenou).

Úplností určení jazyka rozumíme poměr mezi počtem zamítnutých a přijatých výsledků pro konkrétní threshold (vyšší threshold způsobuje větší počet zamítnutí a z toho vyplývající nižší úplnost).

Přesností určení jazyka rozumíme poměr mezi všemi přijatými a správně určenými výsledky (při porovnání s expertním posudkem).

F-míra je harmonickým průměrem mezi hodnotou přesnosti a úplnosti určení jazyka na jednotlivých prazích - thresholdech.

Vztah pro výpočet F-míry:

$$F = 2.\frac{P.R}{P+R} \qquad (1)$$

Jedná se o standardní vztah pro výpočet harmonického průměru s následujícím významem proměnných:

P - Je přesnost určení na daném thresholdu.
R - Je úplnost určení na daném thresholdu.

Konečným kritériem pro potvrzení hypotézy je dosažení vyšší hodnoty nejlepší F-míry s nalezenou (automaticky stanovenou) úzkou sadou oproti sadě široké. Tento výsledek totiž znamená, že došlo automatizovaným postupem ke zlepšení úspěšnosti identifikace oproti zpracování standardnímu.

4.6 Popis a analýza informací

4.6.1 Jazyková sada a její složení

Vstupním nastavením při používání softwaru pro automatickou identifikaci jazyka je volba tzv. jazykové sady, tedy stanovení skupiny jazyků, z nichž bude systémem vybrán jeden výsledný. K těmto jazykům je nejprve nutné získat velké množství trénovacích dat (jedná se o desítky až stovky hodin řeči v konkrétním jazyce), z nichž jsou sejmuty jejich i-vektorové reprezentace, které jsou následně použity k vytvoření modelu zmíněné testovací sady. Jelikož se jedná o diskriminativní model, tak zde neexistují modely jednotlivých jazyků, ale pouze výsledný model celé jazykové sady, který obsahuje všechny požadované jazyky.

Takto zvolená jazyková sada a její model je použit při identifikaci nahrávek tak, že i-vektorová reprezentace každé testované nahrávky je porovnána s reprezentacemi jazyků v modelu jazykové sady a získá procentuální hodnotu shody (dále jen skóre) s každou i-vektorovou reprezentací jazyka, který je v sadě obsažen. Toto skóre vyjadřuje víru systému ve shodu reprezentace testované nahrávky s reprezentací konkrétního jazyka sady. Vzhledem k diskriminativitě tohoto skórování je součet všech skóre nahrávky vůči jednotlivým jazykům jazykové sady roven vždy 100 %. Jazyk s nevyšším skóre se označuje jako vítězný jazyk.

Širokou jazykovou sadou se rozumí sada, která obsahuje všechny jazyky, pro které jsou dostupná trénovací data. Úzká jazyková sada představuje výběr jazyků z široké sady.

Použitím úzké jazykové sady je možné významně zvýšit úspěšnost identifikace jazyka. V principu totiž platí, že zúžením sady o jazyk, o kterém „jistě“ víme, že se v testovaných nahrávkách nevyskytne, dáváme systému zásadní informaci, která má za následek menší možnost omylu při identifikaci. Například pokud máme v široké testovací sadě exotický jazyk sanskrt, tak jeho apriorním vyloučením zvýšíme úspěšnost identifikace. Naopak jeho ponechání zbytečně zvyšuje možnost chyby.

Pozoruhodné je, že díky tomuto principu, může prakticky nastat situace, kdy odebrání jazyka, který je v testovaných datech sice zastoupen, avšak pouze velmi malým procentem výskytů, povede nakonec také ke zvýšení úspěšnosti identifikace v dané skupině nahrávek. Například máme-li tisíc jazykově neurčených nahrávek, které chceme identifikovat, a dojde-li k situaci, že se mezi těmito nahrávkami vyskytne jediná nahrávka v češtině, pak bude jistě z hlediska úspěšnosti identifikace lepší vyřadit češtinu z jazykové sady. Systém tak sice jistě určí tuto jedinou českou nahrávku chybně, avšak nemá možnost udělat chybu druhého typu u všech ostatních testovaných nahrávek.

Základním problémem tak zůstává, jakým způsobem co nejvhodněji odhadnout jazykové spektrum v neznámé skupině nahrávek a stanovit tak onu úzkou jazykovou sadu.

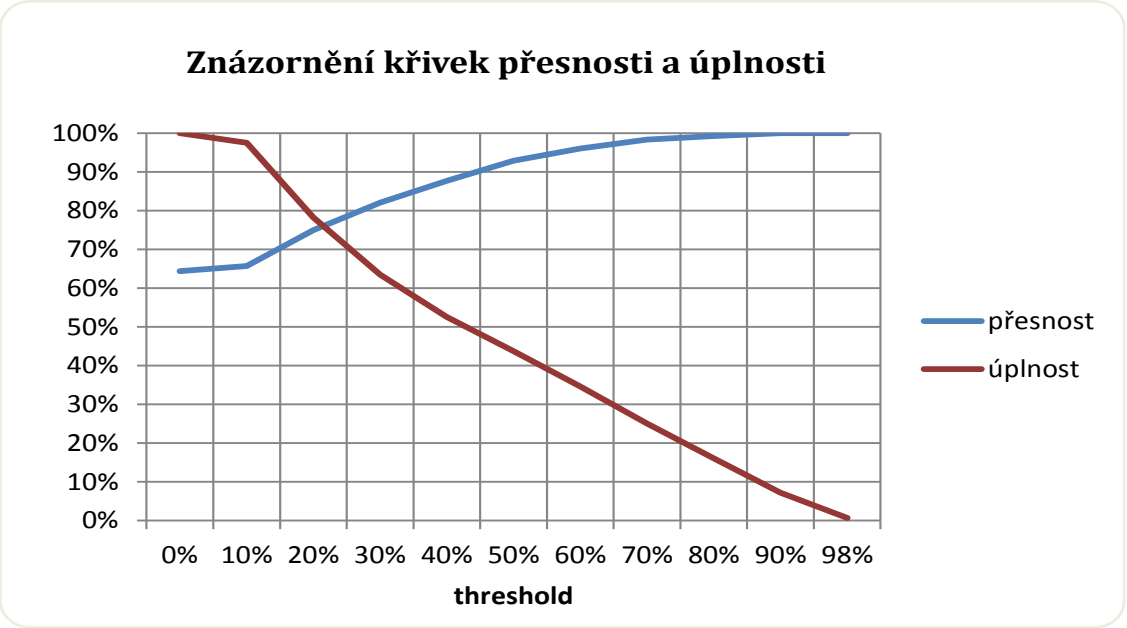
4.6.2 **Křivky přesnosti a úplnosti**

Známe-li skutečný jazyk nahrávek a také jazyk určený systémem pro automatickou identifikaci, můžeme tyto výsledky porovnat a určit tak úspěšnost automatické identifikace.

Pro vyjádření úspěšnosti je třeba, kromě vítězného jazyka nahrávky, vzít v úvahu také hodnotu vítězného skóre. Řekněme, že uživatel použil pro identifikaci jazyka model jazykové sady obsahující jazyky čeština a ruština. Každá testovaná nahrávka tak získala dvě skóre, která vyjadřují víru systému, že nahrávka je v daném jazyce. Vyšší z těchto skóre je takzvaným „vítězným“ skóre a jazyk, kterému toto skóre patří, se nazývá vítězným jazykem.

Vyvstává tak otázka, jak vysoké vítězné skóre budeme považovat za dostatečné k tomu, abychom uznali výsledek identifikace za platný. Toto je podstata thresholdu, který je právě onou hranicí, pod kterou výsledek testování zamítáme a nad kterou ho přijímáme.

Pokud takovouto hranici použijeme při prezentaci výsledků. Pak můžeme získat křivky popisující přesnost a úplnost identifikace jazyka v závislosti na hodnotě tohoto thresholdu. Příklad výsledných křivek můžeme vidět na následujícím grafu (graf 1). Tyto dvě křivky jsou pak velmi důležité pro rozhodování o tom, jakou úroveň thresholdu použijeme při identifikaci dalších nahrávek ze stejného základního souboru.



Graf 1 Křivky přesnosti a úplnosti identifikace

Zdroj: vlastní zpracování

Na grafu (Graf. 1) vidíme, jaký vliv má volba úrovně thresholdu na výsledky identifikace. Pokud uživatel volí vysokou hodnotu, pak je identifikace přesnější ovšem za cenu většího počtu neidentifikovaných nahrávek. Naopak je-li zvolen nízký threshold, pak je identifikováno větší množství nahrávek ovšem za cenu nižší přesnosti identifikace.

Toto rozhodnutí je důležité například v situaci, kdy jsou identifikované nahrávky dále zpracovávány lidmi. Například pokud je ve firmě dostatek lidských zdrojů pro takové zpracování, pak nemusí vadit jejich zatížení chybně určenými nahrávkami a analytik tak volí nižší threshold, naopak pokud je lidských zdrojů

nedostatek, pak je třeba, aby se k těmto lidem dostávaly pouze správně identifikované nahrávky, a analytik tak volí vyšší úroveň threshold.

4.6.3 Intervaly spolehlivosti

Jelikož však hodnoty přesnosti a úplnosti vychází z podílů a jsou stanoveny na základě výběru ze základního souboru, tak zde existuje určitá míra nejistoty, kterou je třeba vzít v úvahu.

Intervaly spolehlivosti pro zjištěné hodnoty přesnosti na jednotlivých thresholdech byly stanoveny pro spolehlivost 95 %. Pro jejich stanovení byly použity standardní vztahy pro parametr alternativního rozdělení π :

$$\text{dolní mez} = p - u_{0,975} \cdot \sqrt{\frac{p.(1-p)}{n}} = p - 1{,}95996 \cdot \sqrt{\frac{p.(1-p)}{n}} \qquad (2)$$

$$\text{horní mez} = p + u_{0,975} \cdot \sqrt{\frac{p.(1-p)}{n}} = p + 1{,}95996 \cdot \sqrt{\frac{p.(1-p)}{n}} \qquad (3)$$

, kde n je počet případů a p je pozorovaný podíl v dané úrovni threshold.

4.6.4 Hustota rozložení výsledných skóre

Dalším důležitým prvkem, který ovlivní rozhodnutí o volbě správného thresholdu pro zamítnutí či přijetí výsledku, je hustota rozložení výsledných skóre. Pokud určíme, kolik procent hodnot skóre se nachází v jednotlivých intervalech thresholdu, pak můžeme usuzovat o vlastnostech systému zpracování.

Diskriminativní skórování systému má za následek, že čím více jazyků použijeme do jazykové sady, tím nižší bude průměrné vítězné skóre nahrávky. Je to způsobeno tím, že nahrávka získává nějakou hodnotu skóre s každým jazykem sady a vzhledem k tomu, že součet všech skóre nahrávky je vždy 100 %, tak hodnota vítězného skóre je snížena o hodnoty skóre všech ostatních jazyků.

Dále se dá z hustoty rozložení skóre odhadovat vhodnost výběru jazykové sady. Obsahuje-li totiž jazyková sada jazyky, které se v testovaných datech nevyskytují nebo neobsahuje-li naopak jazyky, které se v testovaných datech vyskytují, pak to má za následek nižší průměrnou hodnotu vítězného skóre.

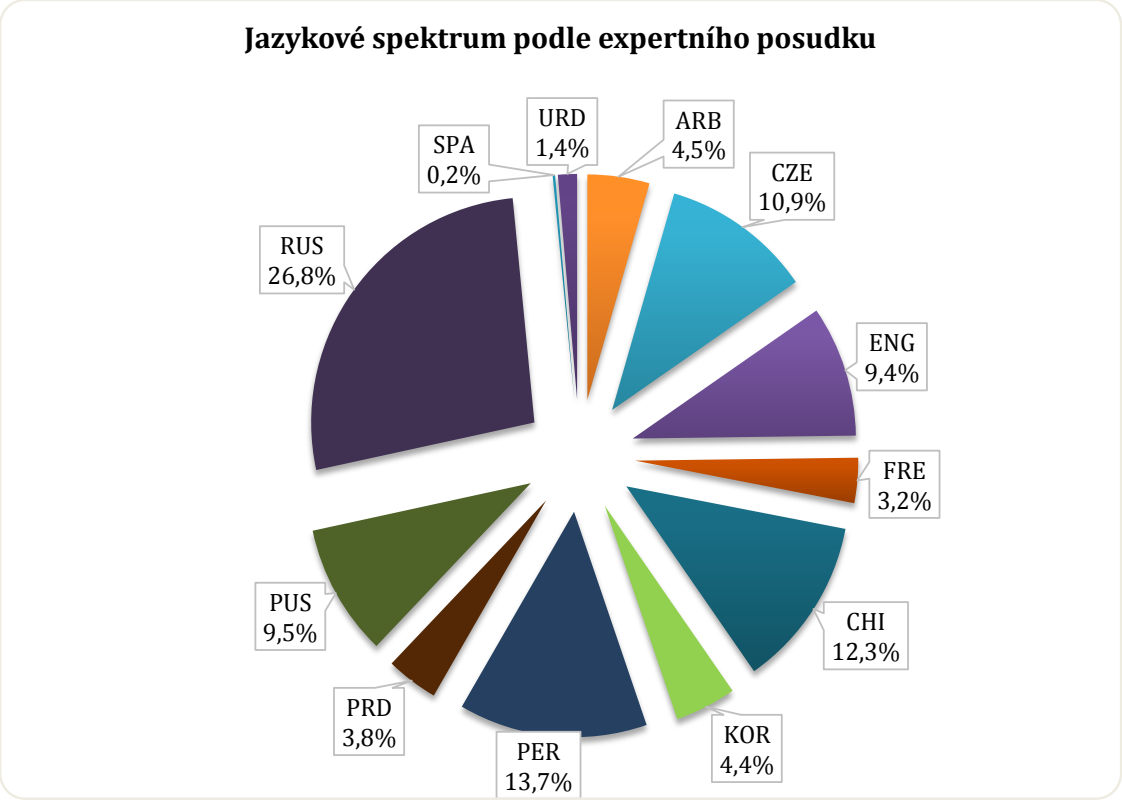
Z pohledu analytika uvažujícího nad vhodnou úrovní thresholdu, je pak informace o hustotě rozložení výsledných skóre velmi důležitá. Pokud se například 90 % vítězných skóre nachází v intervalu mezi 30–50 %, pak je jistě třeba, vzít tuto informaci v úvahu při volbě thresholdu pro další zpracování dat ze stejného základního souboru.

Kapitola byla zpracována s využitím Cyhelský, Kahounová, Hindls (1999) a Anděl (2011).

5 Výsledky šetření

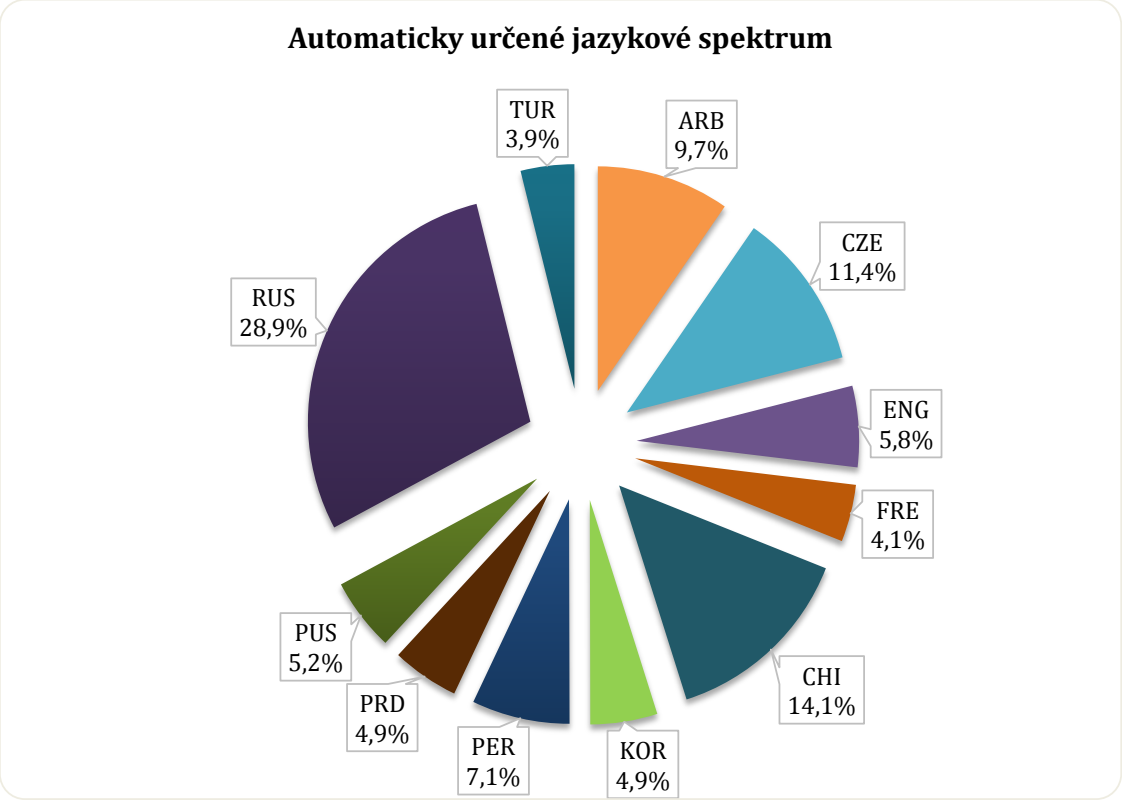
5.1 Jazyková sada a její složení

Na následujících grafech můžeme vidět, že při použití hranice 3% zastoupení jazyka, se podařilo, z výsledků získaných širokou jazykovou sadou, automaticky určit všechny významné jazyky ve skupině (v porovnání se skutečným jazykovým spektrem určeným na základě expertního posudku). Nebyly detekovány pouze jazyky španělština (SPA) a urdu (URD), které jsou, podle expertního posudku, ve skupině nahrávek zastoupeny podílem mezi 0–1 %. Chybně byl detekován jazyk TUR (turečtina).



Graf 2 Skutečné jazykové spektrum testovaných nahrávek (expertní posudek)

Posuzovanou metodou bylo v testované skupině nahrávek určeno následující jazyková spektrum.

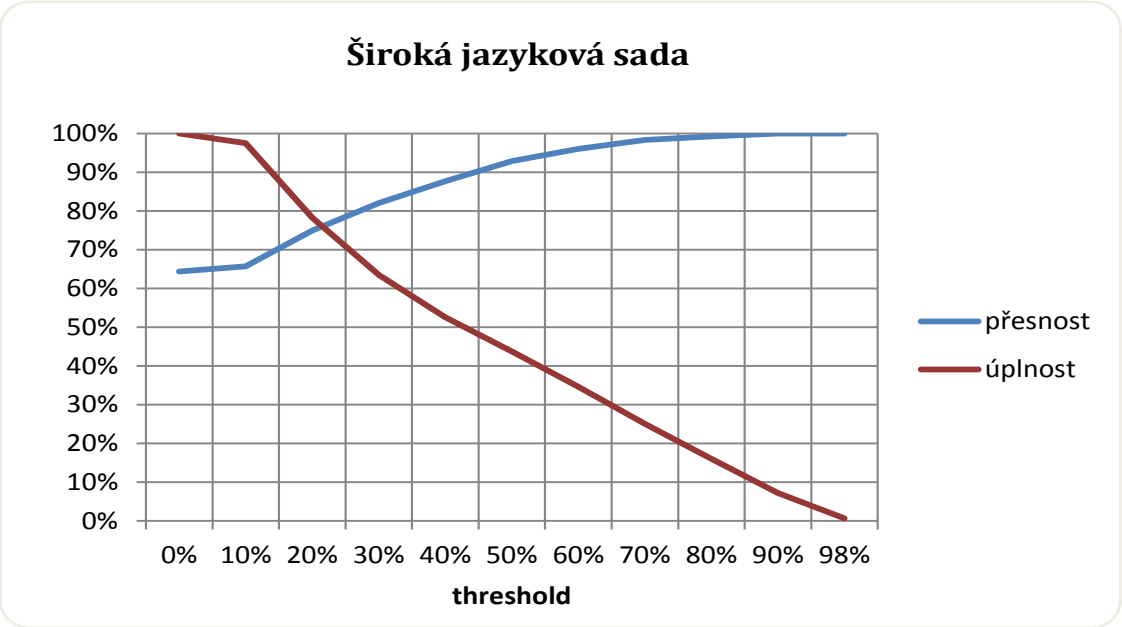


Graf 3 Jazykové spektrum určené automaticky na základě výsledků široké jazykové sady

5.2 Výsledné F-míry jednotlivých sad

Na základě výsledků testování byly stanoveny křivky přesnosti a úplnosti pro každou jazykovou sadu, které můžeme sledovat na následujících grafech.

První graf znázorňuje výsledky při identifikaci nahrávek za použití široké jazykové sady obsahující 42 světových jazyků.

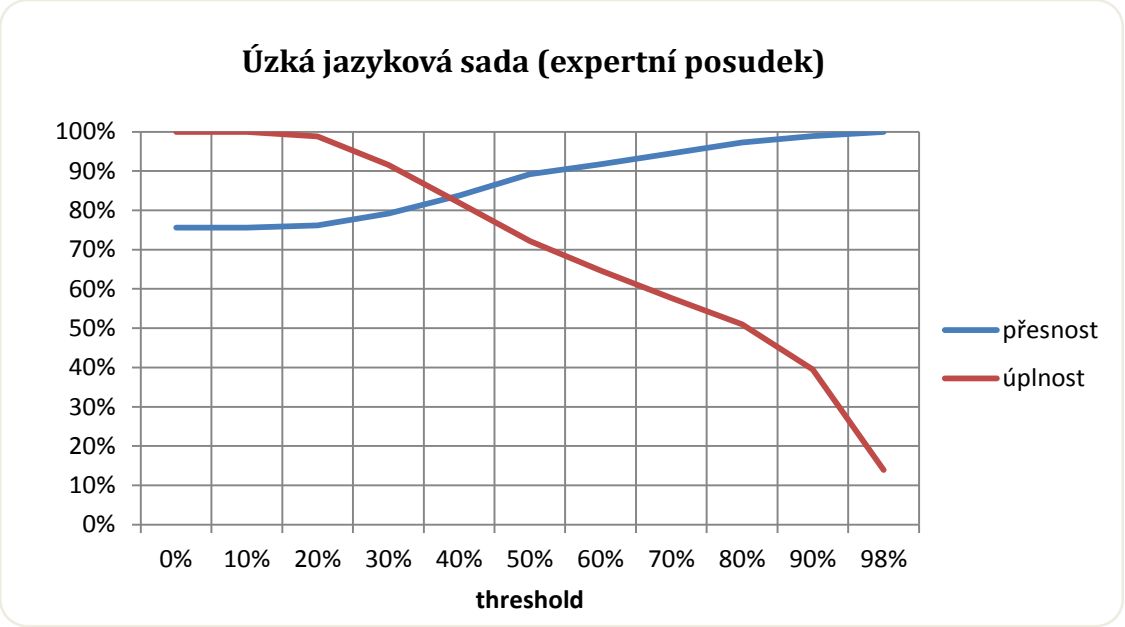


Graf 4 Křivky přesnosti a úplnosti pro širokou jazykovou sadu

Výsledná nejvyšší F-Míra:

F = 78,5 %.

Další graf znázorňuje výsledky při identifikaci nahrávek za použití jazykové sady odpovídající expertnímu posudku, která obsahuje 12 světových jazyků.

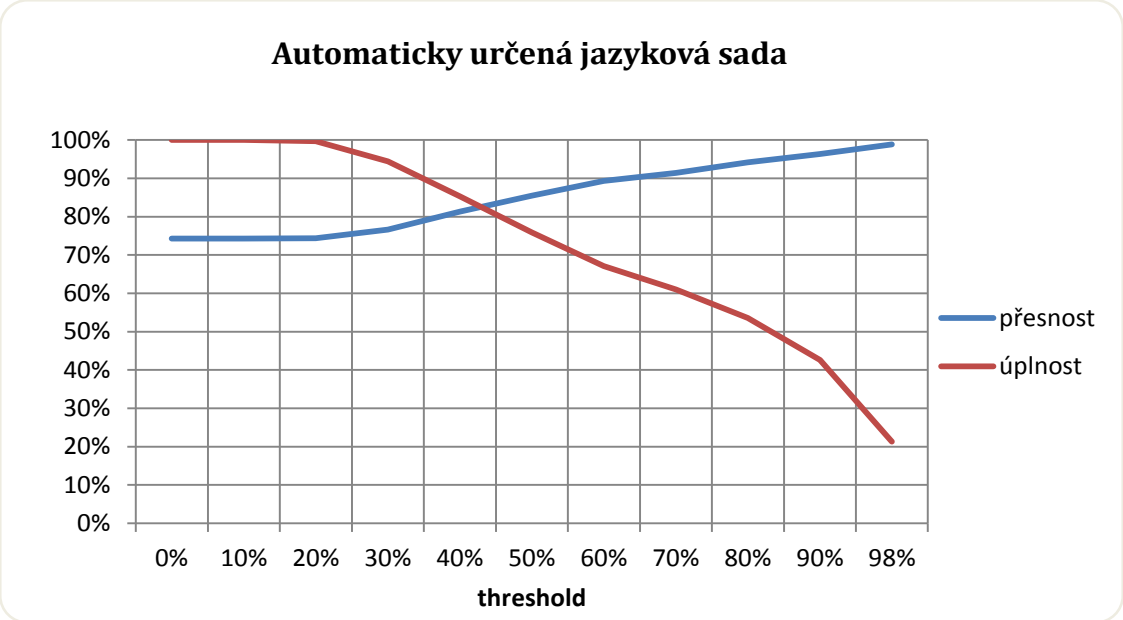


Graf 5 Křivky přesnosti a úplnosti pro úzkou jazykovou sadu

Výsledná nejvyšší F-Míra:

F = 86,1 %.

Poslední graf znázorňuje výsledky při identifikaci nahrávek za použití automaticky stanovené úzké jazykové sady obsahující 11 světových jazyků.



Graf 6 Křivky přesnosti a úplnosti pro automaticky určenou sadu

Výsledná nejvyšší F-Míra:

F = 85,3 %.

5.3 Intervaly spolehlivosti

5.3.1 Intervaly spolehlivosti pro hodnoty přesnosti identifikace

Následují tabulky s vypočtenými intervaly spolehlivosti pro hodnoty přesnosti na jednotlivých úrovních thresholdu.

Uvedené zkratky v záhlaví tabulek:

THR – Je hodnota uvažovaného thresholdu.

k – Je počet nahrávek, jejichž vítězné skóre je nad úrovní thresholdu a jejichž jazyk byl určen správně (shoda s expertním posudkem).

n – Je počet všech nahrávek, jejichž vítězné skóre je nad úrovní thresholdu.

p – Vyjadřuje pravděpodobnost, že nahrávka s vítězným skóre nad threshold bude určena správně.

95 % I. S. – Je rozmezí, v němž se spolehlivostí 95 % leží skutečný podíl v základním souboru.

Tab. 1 Intervaly spolehlivosti přesnosti pro Širokou jazykovou sadu

THR	k	n	p	95 % I. S.
1%	1075	1670	0,6437	±0,0230
10%	1071	1629	0,6575	±0,0230
20%	980	1307	0,7498	±0,0235
30%	870	1060	0,8208	±0,0231
40%	770	878	0,8770	±0,0217
50%	678	730	0,9288	±0,0187
60%	555	578	0,9602	±0,0159
70%	412	419	0,9833	±0,0123
80%	266	268	0,9925	±0,0103
90%	120	120	1	---
98%	11	11	1	---

Tab. 2 Intervaly spolehlivosti přesnosti pro úzkou jazykovou sadu (expertní posudek)

THR	k	n	p	95 % I. S.
1%	1262	1670	0,7557	±0,0206
10%	1262	1670	0,7557	±0,0206
20%	1257	1651	0,7614	±0,0206
30%	1212	1530	0,7922	±0,0203
40%	1147	1369	0,8378	±0,0195
50%	1076	1206	0,8922	±0,0175
60%	991	1080	0,9176	±0,0164
70%	910	963	0,9450	±0,0144
80%	829	852	0,9730	±0,0090
90%	653	660	0,9894	±0,0078
98%	232	232	1	---

Tab. 3 Intervaly spolehlivosti přesnosti pro automaticky určenou jazykovou sadu

THR	k	n	p	95 % I. S.
1%	1240	1670	0,7425	±0,0210
10%	1240	1670	0,7425	±0,0210
20%	1238	1665	0,7435	±0,0210
30%	1208	1577	0,7660	±0,0209
40%	1158	1425	0,8126	±0,0203
50%	1083	1267	0,8548	±0,0194
60%	1001	1121	0,8930	±0,0181
70%	931	1019	0,9136	±0,0172
80%	842	894	0,9418	±0,0153
90%	685	711	0,9634	±0,0138
98%	352	356	0,9888	±0,0109

5.3.2 Intervaly spolehlivosti pro hodnoty úplnosti identifikace

Následují tabulky s vypočtenými intervaly spolehlivosti pro hodnoty úplnosti na jednotlivých úrovních threshold.

Pro následující tabulky se mění význam proměnných takto:

THR – Je hodnota uvažovaného thresholdu.

k – Je počet nahrávek, jejichž vítězné skóre je nad úrovní threshold.

N – Je zde počet všech uvažovaných nahrávek.

p – Vyjadřuje pravděpodobnost, že nahrávka s vítězným skóre nad úrovní thresholdu bude určena správně.

95 % I. S. – Je rozmezí, v němž se spolehlivostí 95 % leží skutečný podíl v základním souboru.

Tab. 4 Intervaly spolehlivosti úplnosti pro širokou jazykovou sadu

THR	k	n	p	95 % I. S.
1%	1670	1670	1	---
10%	1629	1670	0,9754	±0,0074
20%	1307	1670	0,7826	±0,0198
30%	1060	1670	0,6347	±0,0231
40%	878	1670	0,5257	±0,0239
50%	730	1670	0,4371	±0,0238
60%	578	1670	0,3461	±0,0228
70%	419	1670	0,2509	±0,0208
80%	268	1670	0,1605	±0,0176
90%	120	1670	0,0719	±0,0124
98%	11	1670	0,0066	±0,0039

Tab. 5 Intervaly spolehlivosti úplnosti pro úzkou jazykovou sadu (expertní posudek)

THR	k	n	p	95 % I. S.
1%	1670	1670	1	---
10%	1670	1670	1	---
20%	1651	1670	0,9886	±0,0051
30%	1530	1670	0,9162	±0,0133
40%	1369	1670	0,8198	±0,0184
50%	1206	1670	0,7222	±0,0215
60%	1080	1670	0,6467	±0,0229
70%	963	1670	0,5766	±0,0237
80%	852	1670	0,5102	±0,0240
90%	660	1670	0,3952	±0,0234
98%	232	1670	0,1389	±0,0166

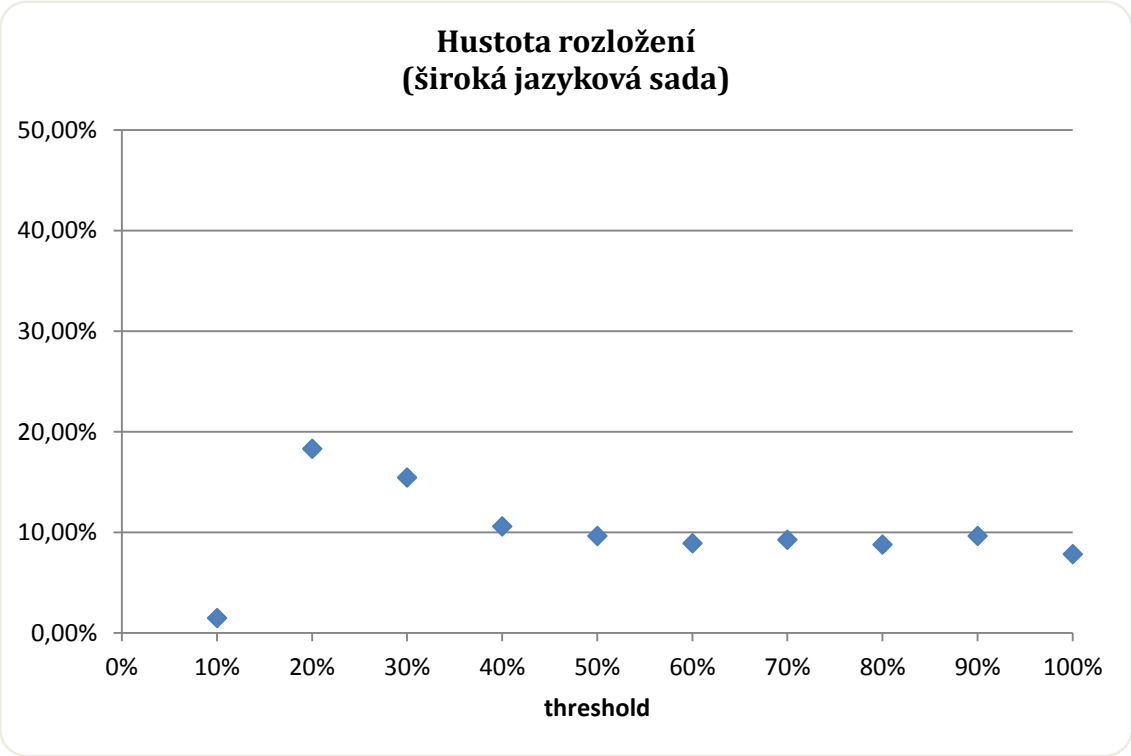
Tab. 6 Intervaly spolehlivosti úplnosti pro automaticky určenou jazykovou sadu

THR	k	n	p	95 % I. S.
1%	1670	1670	1	---
10%	1670	1670	1	----
20%	1665	1670	0,9970	±0,0026
30%	1577	1670	0,9443	±0,0110
40%	1425	1670	0,8533	±0,0170
50%	1267	1670	0,7587	±0,0205
60%	1121	1670	0,6713	±0,0225
70%	1019	1670	0,6102	±0,0234
80%	894	1670	0,5353	±0,0239
90%	711	1670	0,4257	±0,0237
98%	356	1670	0,2132	±0,0196

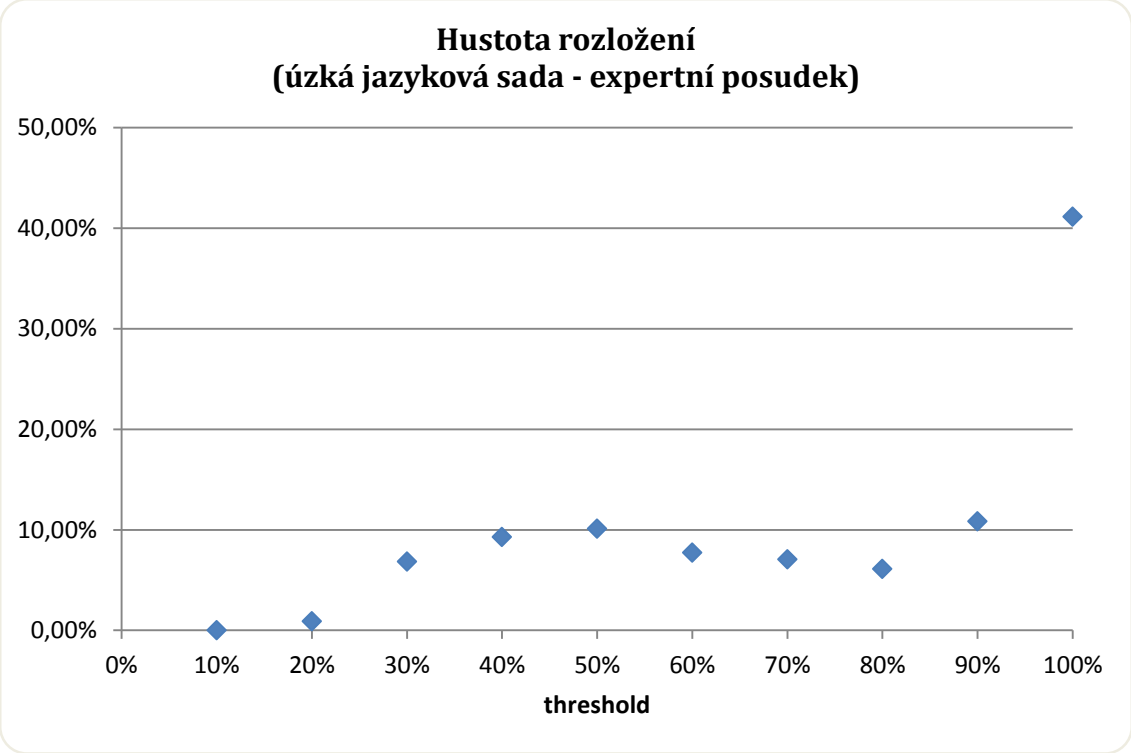
5.4 Hustota rozložení vítězných skóre

Hustota rozložení skóre je určena tak, že jsou určeny intervaly na ose hodnot thresholdu a pro tyto intervaly jsou sečteny výskyty hodnot do nich spadajících.

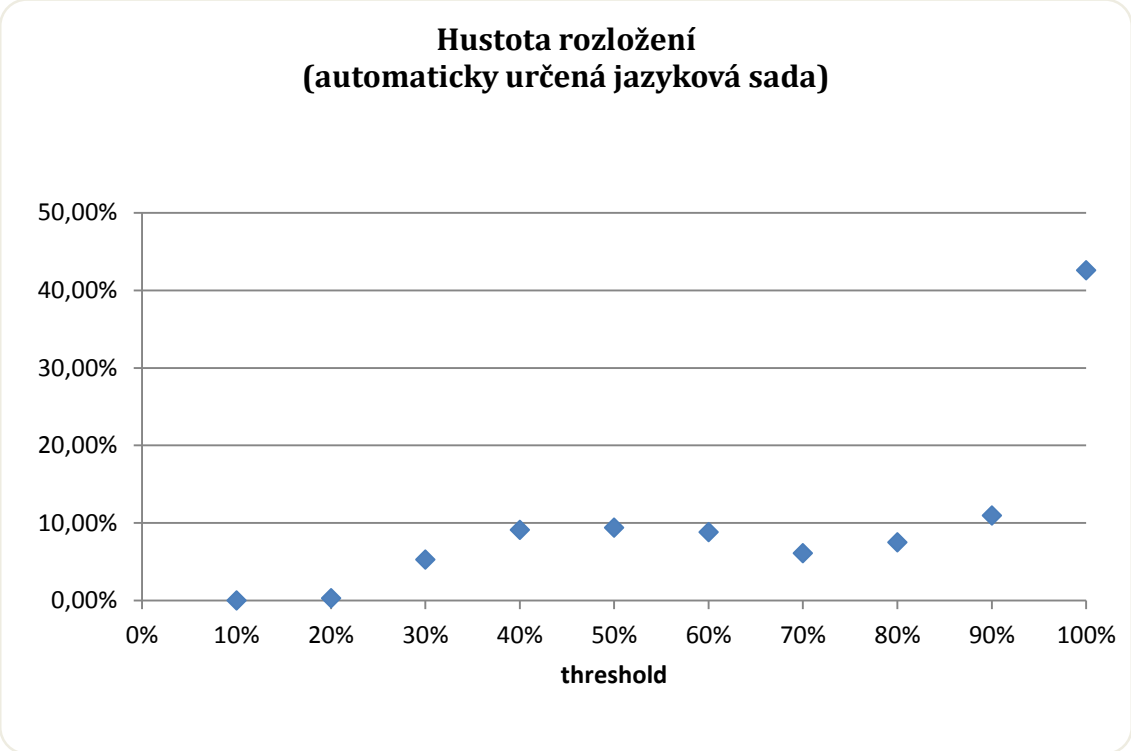
Následující grafy zobrazují procentuální zastoupení skóre nahrávek pro jednotlivé intervaly thresholdu, pro každou testovanou jazykovou sadu.



Graf 7 Hustota rozložení výsledných skóre pro širokou jazykovou sadu



Graf 8 Hustota rozložení výsledných skóre pro expertní jazykovou sadu



Graf 9 Hustota rozložení výsledných skóre pro automaticky určenou jazykovou sadu

6 Shrnutí výsledků

6.1 Vyhodnocení úspěšnosti automatické identifikace

Pro srovnání výsledků identifikace jazyka testovaných nahrávek je využita hodnota nejlepší F-míry jednotlivých sad. Při reálném použití existuje jistě více kritérií, podle nichž je třeba porovnávat jednotlivé výsledky. Často bude záležet na způsobu dalšího zpracování, podle něhož je třeba preferovat jednu ze složek výsledku (přesnost/úplnost), nicméně pro potřeby objektivního porovnání dosažených výsledků se hodnota nejvyšší dosažené F-míry nabízí jako vhodné řešení.

Výsledky F-míry dosažené použitím výše nastíněného postupu jsou následující.

Pro širokou sadu:

F = 78,5 %

Pro sadu určenou expertním posudkem:

F = 86,1 %

Pro automaticky určenou sadu:

F = 85,3 %

Na základě zjištěných výsledků můžeme tedy potvrdit hypotézu, že zpracování širokou jazykovou sadou je dostatečně přesné pro přibližné určení jazykového spektra v neznámé skupině nahrávek. Díky tomu, je možné určit úzkou sadu jazyků, která bude mít lepší výsledky při zpracování než sada široká. Dokonce docházíme k závěru, že výsledky takto vytvořené jazykové sady se blíží výsledkům sady odpovídající expertně určenému jazykovému spektru.

6.2 Intervaly spolehlivosti

Hodnoty intervalů spolehlivosti pro jednotlivé sady se pohybují vždy pod hranicí 2,5 %. Zjištěné výsledky jsou tedy ze statistického hlediska dostatečně vypovídající.

Toto je dáno především vysokým počtem nahrávek (1670), na kterých byl výzkum prováděn.

6.3 Hustota rozložení vítězných skóre

Hustoty rozložení vítězných skóre pro jednotlivé jazykové sady ukazují, že jejich určení usnadní uživateli rozhodnutí o nastavení systému pro budoucí zpracování dat ze stejného základního souboru. Jak je patrné z jednotlivých grafů, rozložení vítězných skóre také vypovídá o vhodnosti zvolené jazykové sady. Dále můžeme pozorovat, že šířka a vhodnost jazykové sady velmi významně ovlivňuje průměrnou hodnotu vítězného skóre nahrávek.

Hustota rozložení vítězných skóre nám také může posloužit jako zdroj informací o samotném systému LID. Použijeme-li totiž k testování nahrávek jazykovou sadu, která odpovídá skutečnému jazykovému rozložení v dané skupině nahrávek, pak můžeme posoudit vlastnosti. Stejně tak porovnání hustot rozložení vítězných skóre pro vhodnou a nevhodnou jazykovou sadu může napovědět, jak bude systém skórovat při reálném použití. Sám o sobě takovýto test nemusí vypovídat přímo o úspěšnosti použitého systému, nicméně naznačí jeho chování. Navíc je možné takovéto testy provádět bez čerpání lidských zdrojů potřebných pro určení jazyka skupiny nahrávek pro testovací účely.

7 Závěry a doporučení

Cíl práce byl splněn a zkoumaná hypotéza byla potvrzena. Výsledky identifikace jazyka širokou jazykovou sadou jsou dostatečně přesné k odhadu jazykového spektra v neznámé skupině nahrávek. Pomocí uvedené metody bylo dosaženo lepších výsledků při identifikaci jazyka na široké skupině nahrávek.

V případě, že nahrávky jsou ve špatné kvalitě, či jazykově neodpovídají použitým jazykovým modelům, pak bude tento postup nejspíše méně přesný a bude třeba jistá korekce například v úrovni zamítnutí výběru jazyka do automaticky zjištěného spektra (zde voleno 3 %). Proto by bylo vhodné zamyslet se nad způsobem detekce takového stavu například s využitím hustot rozložení vítězných skóre.

Určení hustot rozložení vítězných skóre usnadní rozhodování o volbě nastavení systému pro budoucí zpracování dat ze stejného základního souboru. A je tedy přínosem pro praktické využití.

Při zpracování výsledků výpočtů hustot rozložení vítězných skóre se ukazuje, že tato rozložení vypovídají jak o vlastnostech použitého systému pro automatickou identifikaci, tak i o vhodnosti použité jazykové sady. V tomto ohledu by bylo vhodné další vyhodnocení těchto dat.

8 Seznam použité literatury

[1] ADAMEC, Michal. *Moderní rozpoznávače řečové aktivity* (online). Diplomová práce, Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2008. 75 s. Vedoucí práce Mgr. Pavel Rajmic, Ph.D. (cit. 2015-02-01). Přístup z internetu: http://www.vutbr.cz/www_base/zav_prace_soubor_verejne.php?file_id=5788

[2] ANDĚL, Jiří. *Základy matematické statistiky.* 3. vydání. Praha: Matfyzpress, 2011, 358 s. ISBN 978-80-7378-162-0

[3] BOLDIŠ, Petr. *Bibliografické citace dokumentů podle ČSN ISO 690 a ČSN ISO 690-2: Část 1 – Citace: metodika a obecná pravidla*. Verze 3.3. 199-2004, poslední aktualizace 11.11. 2004. Přístup z internetu: http://www.boldis.cz/citace/citace1.pdf

[4] CYHELSKÝ, Lubomír, Jana KAHOUNOVÁ a Richard HINDLS. *Elementární statistická analýza*. 2. vydání. Praha: Management Press, 1999. 319 s. ISBN 80-7261-003-1

[5] ECO, U. *Jak napsat diplomovou práci*. Olomouc: Votobia, 1997. 271 s. ISBN 80-7198-173-7

[6] LEE, Chin-Hui. *Principles of Spoken Language Recognition*. In: Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008, s. 785-796. ISBN 978-3-540-49127-9

[7] MATĚJKA, Pavel. *Phonotactic and Acoustic Language Recognition* (online). Doctoral thesis, Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 2008. 91 l. Vedoucí práce prof. Milan Sigmund. Přístup z internetu: http://www.fit.vutbr.cz/research/groups/speech/publi/2009/matejka_thesis.pdf

[8] PSUTKA, Josef, MÜLLER, Luděk, MATOUŠEK, Jindřich, RADOVÁ Vlasta. *Mluvíme s počítačem česky.* Vyd. 1. Praha: Academia, 2006. 752 s. ISBN 80-200-1309-1

[9] REYNOLDS, Douglas A., et al. *Automatic language recognition via spectral and token based approaches*. In: Springer Handbook of Speech Processing. Springer Berlin Heidelberg, 2008. s. 811-824. ISBN 978-3-540-49127-9

[10] SCHWARZ, Petr. *Phoneme recognition based on long temporal context* (online). Doctoral thesis, Brno: Brno University of Technology, Faculty of Information Technology, 2008. 95 l. Vedoucí práce Doc. Dr. Ing. Jan Černocký. Přístup z internetu: http://www.fit.vutbr.cz/~schwarzp/publi/thesis.pdf

[11] SIGMUND, Milan. Analýza řečových signálů. 1. vydání. VUT Brno 2000. ISBN 80-214-1783-8

[12] SINISCALCHI, Sabato Marco, et al. *Universal attribute characterization of spoken languages for automatic spoken language recognition*. In Computer Speech & Language. Faculty of Engineering and Architecture, Kore University of Enna, Sicily, Italy, vol. 27, 2013. ISSN 0885-2308

[13] TAN, Ying-Wei, LIU, Wen-Ju. *Robust Voice Activity Detection Using the Combination of Short-Term and Long-Term Spectral Patterns.* In Pattern Recognition. Springer Berlin Heidelberg, vol. 484, 2014. ISBN 978-3-662-45643-9

[14] WU, Ji, ZHANG, Xiao-Lei. *An efficient voice activity detection algorithm by combining statistical model and energy detection*. In Eurasip Journal on Advances in Signal Processing. Springer International Publishing AG, vol. 2011. ISSN 1687-6180

[15] ZHANG, Wei-Qiang, et al. *Spoken language recognition based on gap-weighted subsequence kernels*. In Speech Communication. Tsinghua National Laboratory for information Science and Technology, Tsinghua University, Beijing, China, vol. 60, 2014. ISSN 0167-6393

Přílohy

1) Doplňující tabulky ke grafům.

Tabulka 1 data pro graf č. 4. Přesnosti a Úplnosti určení pro jednotlivé hodnoty thresholdu

Threshold	Celkem nahrávek	Nad threshold	Nad threshold a správně	Přesnost určení	Úplnost určení
0,00%	1670	1670	1075	64,37%	100,00%
10,00%	1670	1629	1071	65,75%	97,54%
20,00%	1670	1307	980	74,98%	78,26%
30,00%	1670	1060	870	82,08%	63,47%
40,00%	1670	878	770	87,70%	52,57%
50,00%	1670	730	678	92,88%	43,71%
60,00%	1670	578	555	96,02%	34,61%
70,00%	1670	419	412	98,33%	25,09%
80,00%	1670	268	266	99,25%	16,05%
90,00%	1670	120	120	100,00%	7,19%
98,00%	1670	11	11	100,00%	0,66%

Zdroj: vlastní zpracování

Tabulka 2 data pro graf č. 5. Přesnosti a Úplnosti určení pro jednotlivé hodnoty thresholdu

Threshold	Celkem nahrávek	Nad threshold	Nad threshold a správně	Přesnost určení	Úplnost určení
0,00%	1670	1670	1262	75,57%	100,00%
10,00%	1670	1670	1262	75,57%	100,00%
20,00%	1670	1651	1257	76,14%	98,86%
30,00%	1670	1530	1212	79,22%	91,62%
40,00%	1670	1369	1147	83,78%	81,98%
50,00%	1670	1206	1076	89,22%	72,22%
60,00%	1670	1080	991	91,76%	64,67%
70,00%	1670	963	910	94,50%	57,66%
80,00%	1670	852	829	97,30%	51,02%
90,00%	1670	660	653	98,94%	39,52%
98,00%	1670	232	232	100,00%	13,89%

Zdroj: vlastní zpracování

Tabulka 3 data pro graf č. 6. Úplnosti určení pro jednotlivé hodnoty thresholdu

Threshold	Celkem nahrávek	Nad threshold	Nad threshold a správně	Přesnost určení	Úplnost určení
0,00%	1670	1670	1240	74,25%	100,00%
10,00%	1670	1670	1240	74,25%	100,00%
20,00%	1670	1665	1238	74,35%	99,70%
30,00%	1670	1577	1208	76,60%	94,43%
40,00%	1670	1425	1158	81,26%	85,33%
50,00%	1670	1267	1083	85,48%	75,87%
60,00%	1670	1121	1001	89,30%	67,13%
70,00%	1670	1019	931	91,36%	61,02%
80,00%	1670	894	842	94,18%	53,53%
90,00%	1670	711	685	96,34%	42,57%
98,00%	1670	356	352	98,88%	21,32%

Zdroj: vlastní zpracování

Tabulka 4 data pro graf č. 7. Hustota rozložení výsledných skóre pro širokou jazykovou sadu

Threshold	Pod threshold	%	kumulativně
10%	25	1.50%	25
20%	306	18.32%	331
30%	258	15.45%	589
40%	177	10.60%	766
50%	161	9.64%	927
60%	149	8.92%	1076
70%	155	9.28%	1231
80%	147	8.80%	1378
90%	161	9.64%	1539
100%	131	7.84%	1670

Zdroj: vlastní zpracování

Tabulka 5 data pro graf č. 8. Hustota rozložení výsledných skóre pro expertní jazykovou sadu

Threshold	Pod threshold	%	kumulativně
10%	0	0.00%	0
20%	15	0.90%	15
30%	114	6.83%	129
40%	155	9.28%	284
50%	169	10.12%	453
60%	129	7.72%	582
70%	118	7.07%	700
80%	102	6.11%	802
90%	181	10.84%	983
100%	687	41.14%	1670

Zdroj: vlastní zpracování

Tabulka 6 data pro graf č. 9 Hustota rozložení výsledných skóre pro automaticky určenou jazykovou sadu

Threshold	Pod threshold	%	kumulativně
10%	0	0.00%	0
20%	5	0.30%	5
30%	88	5.27%	93
40%	152	9.10%	245
50%	157	9.40%	402
60%	147	8.80%	549
70%	102	6.11%	651
80%	125	7.49%	776
90%	183	10.96%	959
100%	711	42.57%	1670

Zdroj: vlastní zpracování

FIM UHK
UNIVERZITA HRADEC KRÁLOVÉ
Fakulta informatiky a managementu
Rokitanského 62, 500 03 Hradec Králové, tel.: 493 331 111, fax: 493 332 235

Zadání k závěrečné práci

Jméno a příjmení studenta: **Pavel Komeštík**

Obor studia: Aplikovaná informatika
Jméno a příjmení vedoucího práce: **Josef Dolejš**

Název práce:
Statistická analýza dat získaných technologií Language Identification

Název práce v AJ:
Statistical analysis of data acquired by language identification technology

Podtitul práce:
Zvýšení úspěšnosti systému pro určení jazyka pomocí statistického přístupu

Podtitul práce v AJ:
Increasing of succes rate of language identification system using statistical approach

Cíl práce: Cílem práce je zvýšení úspěšnosti určení jazyka nahrávky s využitím automatického odhadu jazykového spektra v dané skupině nahrávek.

Osnova práce:

1. ANOTACE
2. ÚVOD
3. TEORETICKÁ VÝCHODISKA
4. PRŮBĚH A VÝSLEDKY ŠETŘENÍ
5. NÁVRHY A DOPORUČENÍ
6. ZÁVĚR
7. SEZNAM ZDROJŮ
8. PŘÍLOHY

Projednáno dne:

Podpis studenta

Podpis vedoucího práce