|  |  |
| --- | --- |
|  | |
| **Srovnání současných modelů Speech-to-Text** | |
|  | |
| Bc. Juraj Králik | |
|  | |
|  |  |
| Diplomová práce  2025 |  |
|  |  |





**PROHLÁŠENÍ AUTORA DIPLOMOVÉ PRÁCE**

**Beru na vědomí, že**

* odevzdáním diplomové práce souhlasím se zveřejněním své práce podle zákona č. 111/1998 Sb., v platném znění bez ohledu na výsledek obhajoby;
* diplomová práce bude uložena v elektronické podobě v univerzitním informačním systému a bude dostupná k nahlédnutí;
* jedno vyhotovení diplomové práce v listinné podobě bude ponecháno Univerzitě Tomáše Bati ve Zlíně k uložení;
* na moji diplomovou práci se plně vztahuje zákon č. 121/2000 Sb. o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon) ve znění pozdějších právních předpisů, zejm. § 35 odst. 3;
* podle § 60 odst. 1 autorského zákona má Univerzita Tomáše Bati ve Zlíně právo na uzavření licenční smlouvy o užití školního díla v rozsahu § 12 odst. 4 autorského zákona;
* podle § 60 odst. 2 a 3 mohu užít své dílo – diplomovou práci – nebo poskytnout licenci k jejímu využití jen s předchozím písemným souhlasem Univerzity Tomáše Bati ve Zlíně, která je oprávněna v takovém případě ode mne požadovat přiměřený příspěvek na úhradu nákladů, které byly Univerzitou Tomáše Bati ve Zlíně na vytvoření díla vynaloženy (až do jejich skutečné výše);
* pokud bylo k vypracování diplomové práce využito softwaru poskytnutého Univerzitou Tomáše Bati ve Zlíně nebo jinými subjekty pouze ke studijním a výzkumným účelům (tj. k nekomerčnímu využití), nelze výsledky diplomové práce využít ke komerčním účelům;
* pokud je výstupem diplomové práce jakýkoliv softwarový produkt, považují se za součást práce rovněž i zdrojové kódy, popř. soubory, ze kterých se projekt skládá; neodevzdání této součásti může být důvodem k neobhájení práce.

**Prohlašuji, že**

* jsem na diplomové práci pracoval(a) samostatně a použitou literaturu jsem řádně citoval(a); v případě publikace výsledků budu uveden(a) jako spoluautor;
* odevzdaná verze diplomové práce a verze elektronická nahraná do IS/STAG jsou obsahově totožné.

Ve Zlíně, dne .............................. ...............................................................

podpis autora

ABSTRAKT

Diplomová práca sa zaoberá problematikou a porovnaním súčasných modelov technológie Speech-to-Text (STT). Cieľom práce je analyzovať dostupné modely z hľadiska ich presnosti, rýchlosti a schopnosti adaptácie na rôzne jazyky a prízvuky. Práca obsahuje teo-retický prehľad histórie a vývoja technológie STT, klasifikáciu modelov na základe použí-vaných technológií a podrobnú analýzu vybraných základných modelov.

Praktická časť je zameraná na zostavenie metodiky na porovnanie modelov a vyko-nanie testovania na vybraných datasetoch. Výsledky testov sú analyzované s dôrazom na porovnanie silných a slabých stránok jednotlivých modelov.

V závere práce sú zhrnuté hlavné poznatky, vrátane návrhov na možné reálne apli-kácie a odporúčaní pre ďalší výskum v oblasti Speech-to-Text technológií. Práca prispieva k lepšiemu pochopeniu možností súčasných modelov STT a ich využitia v rôznych oblastiach, ako sú automatizácia, prístupnosť alebo lingvistika.

Klíčová slova: Speech-to-Text, Speech recognition

ABSTRACT

The thesis addresses the issues and comparison of current Speech-to-Text (STT) technology models. The aim of the work is to analyze the available models in terms of their accuracy, speed, and adaptability to different languages and accents. The thesis includes a theoretical overview of the history and development of STT technology, a classification of models based on the technologies used, and a detailed analysis of selected foundational models.

The practical part focuses on developing a methodology for model comparison and conducting tests on selected datasets. The test results are analyzed with an emphasis on comparing the strengths and weaknesses of individual models.

In the conclusion, the thesis summarizes key findings, including suggestions for potential real-world applications and recommendations for further research in the field of Speech-to-Text technology. The work contributes to a better understanding of the capabilities of current STT models and their applications in various areas such as automation, accessibility, and linguistics.

Keywords: Speech-to-Text, Speech recognition

Zde je místo pro případné poděkování, popř. motto, úryvky knih atp.

OBSAH

TO DO – vygenerovať a formátovať az na konci

Nebyla nalezena položka obsahu.

Úvod

TO DO:

(translate) The task of speech recognition is to map an acoustic signal containing a spoken natural language utterance into the corresponding sequence of words intended by the speaker.

Modely

Praktická časť - overview

|  |  |
| --- | --- |
|  | teoRetická část |

1. Teoretický prehľad technológie speech-to-text

Rozpoznávanie hovorenej reči, či prepis hovoreného slova do písma jeden z prvých cieľov počítačového spracovania jazyka. Spracovanie reči ako také predchádza počítač o približne 14 rokov. Prvý stroj, ktorý dokázal rozpoznávať reč bola hračka z roku 1922. Jednalo sa o hračku „Radio Rex“, celuloidový pes, ktorý sa pohyboval pomocou pružiny, ktorú uvoľnila akustická energia 500 Hz. Prvý formant samohlásky [eh] v „Rex“ je približne 500 Hz a vyvolávalo to dojem, že Rex na povel vyjde z búdy.

V moderných časoch sa však od automatického rozpoznávania reči očakáva omnoho viac a to prevod akejkoľvek vlnovej formy na príslušný reťazec slov. Táto disciplína prešla mnohými fázami do aktuálneho stavu životaschopnosti a širokého využitia v domácich spotrebičoch, osobných asistentoch, automatické generovanie titulkov z audio a video podkladov či využitie ako náhrada klávesnice.

Úloha rozoznávania reči sa líši vo viacerých rozmeroch. Jedným je slovná zásoba, pričom sa presnosť mení v závislosti od jej rozsahu – či sa jedná o rozoznávanie dvoch slov („yes“ a „no“), jedenásť slov (rozoznávanie číslic „zero“ až „nine“ plus „oh“), alebo rozoznávanie celej hovorenej konverzácie s veľkou slovnou zásobou v rozsahu až 60 000 slov. Druhým môže byť to, s kým hovoriaci hovorí. Ľudia hovoriaci so strojmi (diktovanie, alebo rozprávanie do dialógového systému) sú ľahšie rozoznateľní, ako ľudia hovoriaci s ľuďmi. Rozpoznanie reči ľudí, ktorí spolu vedú konverzáciu (napríklad prepis obchodného stretnutia) je najnáročnejšie. Zdá sa, že keď ľudia hovoria priamo zo strojmi a hovoria bez prítomnosti publika, zjednodušujú svoje vety a rozprávajú pomalšie a jasnejšie. Tretím rozmerom je šum. Je veľký rozdiel v tom, či sa hovoriaci nahráva v tichej miestnosti, alebo na ulici plnej iných hovoriacich ľudí, či iných zdrojov zvuku. Posledným rozmerom je prízvuk, či charakteristika hovoriaceho. Rozoznávanie reči môže mať problémy, pokiaľ má hovoriaci regionálny, či etnický dialekt, alebo je hovoriacim dieťa, ak je systém trénovaný iba na hovorcoch v štandardných dialektoch alebo dospelých hovorcoch [1].

.

* 1. História a vývoj STT technológií

1.1.3 Začiatky vývoja STT (60

Vývoj automatického rozvoja reči prebieha už takmer štyri storočia. Prvé pokusy o prevod reči boli už v roku 1952, keď v Bell Laboratories, Davis, Biddulph a Balashek zostavili systém na rozoznávanie jednotlivých číslic pre jedného hovoriaceho [6]. Systém sa spoliehal hlavne na meranie spektrálnych rezonancií samohláskovej časti každej číslice. V nezávislom projekte v RCA Laboratories v roku 1956 sa Olson a Belar pokúsili rozpoznať 10 rôznych slabík jedného hovoriaceho, stvárnených ako 10 jednoslabičných slov [7]. Systém sa opäť spoliehal na spektrálne merania najmä počas samohláskových častí. V roku 1959 sa na University College v Anglicku pokúsili Fry a Denes vybudovať fonémový rozpoznávač, ktorý rozpoznával štyri samohlásky a deväť spoluhlások [8]. Používali spektrálny analyzátor a vzorový porovnávač na prijímanie rozhodnutí o rozpoznávaní. Novým aspektom tohto výskumu bolo využitie štatistických informácií o prípustných sekvenciách foném v angličtine (primitívna forma jazykovej syntaxe) na zlepšenie celkovej presnosti rozpoznávania foném pre slová obsahujúce dve alebo viac foném. Ďalší významný projekt v tomto období bol rozpoznávač samohlások od Forgieho a Forgieho, vytvorený v MIT Lincoln Laboratories v roku 1959, v ktorom bolo rozpoznávaných 10 samohlások vložených vo formáte /b/-samohláska-/t/ nezávisle od hovoriaceho [9].

V 60. rokoch vstúpili do oblasti rozpoznávania reči japonské laboratóriá a začali budovať špecializovaný hardvér ako súčasť svojich systémov. Jeden z prvých japonských systémov, vyvinutý Suzuki a Nakatom z Radio Research Lab, bol hardvérový rozpoznávač samohlások. Zložitý analyzátor filtrového banku bol použitý spolu s logickým obvodom, ktorý prepojil výstupy každého kanála spektrálneho analyzátora na rozhodovací obvod pre výber samohlásky, a na určenie vyslovenej samohlásky sa použilo rozhodnutie väčšiny (majority decision logic scheme). Ďalší hardvérový rozpoznávač v Japonsku bol vyvinutý Sakaiom a Doshitom z univerzity v Kjóte v roku 1962, ktorí vytvorili hardvérový rozpoznávač foném [10]. Bol použitý hardvérový segmentátor reči spolu s analýzou pomocou prahového prechodu na rôznych častiach hovoreného vstupu, aby sa vytvoril výstup rozpoznávania. Tretím japonským pokusom bol rozpoznávač číslic, vyvinutý Nagatom a jeho kolegami v NEC Laboratories v roku 1963.

V 60. rokoch začali tri kľúčové výskumné projekty, ktoré mali zásadný význam pre výskum a vývoj rozpoznávania reči počas nasledujúcich 20 rokov. Prvý z týchto projektov boli snahy Martina a jeho kolegov v RCA Laboratories, začínajúce koncom 60. rokov, ktorých cieľom bolo vyvinúť realistické riešenia problémov spojených s nejednotnosťou časových mierok v hovorených udalostiach. Martin vyvinul súbor základných metód časovej normalizácie, založených na schopnosti spoľahlivo detegovať začiatky a konce rečových segmentov, čo výrazne znížilo variabilitu rozpoznávacieho skóre [11]. Martin nakoniec metódu zdokonalil a založil jednu z prvých firiem, Threshold Technology, ktorá produkovala a predávala produkty pre rozpoznávanie reči.

Približne v tom istom čase, v Sovietskom zväze, Vintsyuk navrhol použitie metód dynamického programovania na zarovnávanie párov rečových viet [12]. Hoci podstata konceptov dynamického preťahovania (warping), ako aj základné verzie algoritmov pre spojité rozpoznávanie slov, boli zakotvené v práci Vintsyuka, boli na Západe do značnej miery neznáme a do popredia sa dostali až začiatkom 80. rokov — teda dlho po tom, čo sa už rozšírili formálnejšie metódy navrhnuté a implementované inými.

Posledným významným výsledkom 60. rokov bol výskum Reddyho v oblasti spojitého rozpoznávania reči pomocou dynamického sledovania foném [13]. Reddyho výskum nakoniec vyústil do dlhej a úspešnej línie výskumu rozpoznávania reči na Carnegie Mellon University a ktorá dodnes zostáva svetovým lídrom v oblasti systémov spojitého rozpoznávania reči.

V 70. rokoch výskum rozpoznávania reči dosiahol množstvo významných míľnikov. Poprvé sa oblasť rozpoznávania izolovaných slov alebo diskrétnych výrokov stala životaschopnou a použiteľnou technológiou na základe štúdií Velichka a Zagoruyka v Rusku, Sakoa a Chibu v Japonsku, a Itakuru v Spojených štátoch. Ruské štúdie pomohli rozšíriť použitie konceptov dynamického programovania na rozpoznávanie reči; japonský výskum ukázal, že metódy dynamického programovania môžu byť úspešne aplikované; a Itakurov výskum ukázal, ako princípy linear predictive coding (LPC), ktoré už boli úspešne využívané v low-bit-rate speech coding, mohli byť rozšírené do systémov na rozoznávanie reči za použitia korektných meraní dĺžok podľa LPC spektrálnych parametrov.

[1] chapter 16.2

[3] chapter **1.4**

* 1. Prehľad základných princípov a metód (akustické modely, jazykové modely, neurónové siete)

[2] chaptr 12.3, 12.4 (?)

[3] chapters 2 - 9

* 1. Klasifikácia STT modelov podľa technológií a použití

[3] chapters 6.5 – 6.11, 6.13(lepšie do záveru - (6)?), **7, 8**

[4] chapter **1.3!, celá kniha – modely do podpodkapitol**

|  |  |
| --- | --- |
|  | praktická část |

Výber a charakteristika modelov

text

Kritéria výberu dostupných modelov

Open-source

Python integration,

Prelearned

Modely:

* Python Speech recognition lib
* Whisper OpenAI
* Wav2Vec Facebook
* Kaldi (TODO: in WSL)

Príprava modelov

* Inštalácia modelov (+dependencies)
* Python modul pre každý model (funkcia s parametrom file\_name, ktorý má previesť na text, return text string)
* Main python script spája moduly a spúšťa testing za každý modul na všetky testovacie materiály, vytvára JSON file outputy

Príprava testovacieho materiálu

Mozilla Common Voice (open-source, ponuka 89GB audionahrávok ~5sec .mp3 v en, ďalšie jazyky ak to umožňujú modely –sk, cz, de)

- oddeliť iba validated audio

- generovať kontrolný JSON file z validated sentences pre každé audio

- vytvorenie duplicít nahrávok s upravenou rýchlosťou/pridaný šum

Metodika testovania a porovnávania

Druhý python script na porovnávanie vygenerovaného textu s korektným textom, rozkladanie textu na slová a bodovanie každého jednotlivého testu

Kritéria hodnotenia modelov

Každá prevedená nahrávka ohodnotená hodnotou float [0, 1]. Output je dict[názov: str, úspešnosť: float]. Každý prevedený text sa rozdelí na slová a hodnotenie sa upravuje podľa 1/počet slov. Vyhodnotenie celkovej chybovosti modelu.

Druhý output – doba prevedenia každého testu dict[názov: str, time\_elapsed: float].

Praktické testovanie modelov

Príprava datasetov

Postup prípravy, prípadne dodatočné úpravy

Postup testovania jednotlivých modelov

TODO

Zber a spracovanie výsledkov

TODO Script na spracovanie výsledkov

Analýza a prezentácia výsledkov

Aktuálny stav po teste

Voľba vizualizácie výsledkov

Porovnanie modelov na základe zvolených metrík

Vyjadrenie chybovosti modelov

Vyhodnotenie rýchlosti prevodu

Porovnanie chybovosti bežného textu a textu zaťaženého accentom/rýchlosťou/šumom pozadia

Vizualizácia výsledkov

TODO

Rozbor silných a slabých stránok modelov

Vyhodnotenie chybovosti a rýchlosti spracovania, vrátane zmien vznikajúcich zmenou rýchlosti/pridanie šumu

Možnosti využitia STT technológií v praxi

Vyhodnotiť využiteľnosť použitých modelov

Použitie v softwaroch

Príklady využitia v praxi

Budúci vývoj a trendy v oblasti STT

[3] chapter 6,

Pridať online zdroje?

ZávEr

Využiteľnosť použitých modelov, skúsenoti s prípravou a použiteľnosť pre bežného užívateľa.

Zoznam použitej literatúry

[1] JURAFSKY Daniel, MARTIN James. *Speech and Language Processing*. Pearson, Harlow. 2014. ISBN 9781292025438.

[2] GOODFELLOW, Ian. *Deep Learning*, The MIT Press. Cambridge, MA. 2017. ISBN 9798841205937.

[3] RABINER Lawrence, JUANG Biing-Hwang. *Fundamentals of speech recognition*. Pearson Education, Delhi, 2005. ISBN 9788129701381.

[4] YU Dong, DENG Li. *Automatic speech recognition : a deep learning approach*. Springer, London. 2015. ISBN 9781447157786.

[5] KAMATH Uday, LIU John, WHITAKER James. *Deep learning for NLP and speech recognition.* Springer, Cham, Switzerland. 2019. ISBN 9783030145965.

[6] K. H. Davis, R. Biddulph, and S. Balashek, “*Automatic Recognition of Spoken Digits*,” J. Acoust. Soc. Am., 24 (6): 637–642, 1952

[7] H. F. Olson and H. Belar, “Phonetic Typewriter,” J. Acoust. Soc. Am., 28 (6): 1072–1081, 1956

TODO Vyhľadať viac online

Seznam obrázků

[Obrázek 1 Popisek obrázku 17](#_Toc173623594)

Seznam tabulek

[Tabulka 1 Popisek tabulky 17](#_Toc173262262)

Seznam použitých symbolů a zkratek

STT Speech-to-text

Seznam příloh

Příloha P I: Název první přílohy

Příloha P I: Název první přílohy

Obsah první přílohy.