|  |  |
| --- | --- |
|  | |
| **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 časť |

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

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 [10]. 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 [11]. 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 [12].

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 [13]. 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 [14]. 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 [15]. 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 rozšírili využívanie pattern-recognition v rozpoznávaní reči [16]; japonský výskum preukázal úspešné aplikovanie metód dynamického programovania [17]; 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 [18].

Ďalším míľnikom 70. rokov bol začiatok dlhodobého, veľmi úspešného skupinového úsilia v oblasti rozpoznávania reči s veľkou slovnou zásobou v IBM, v rámci ktorého skúmali tri rôzne úlohy v priebehu takmer dvoch desaťročí, predovšetkým jazyk New Relic [19] pre jednoduché databázové dotazy, využitie laser patent text language [20] pre transkripciu laser patentov a úlohu korešpondencie v kancelárskom prostredí, nazývanú Tangora [21], na diktovanie zápisníc a poznámok.

V AT&T Bell Labs výskumníci začali sériu experimentov zameraných na vytvorenie systémov rozpoznávania reči, ktoré by boli skutočne nezávislé na hovorcoch [22]. Na dosiahnutie tohto cieľa sa použila široká škála sofistikovaných zhlukových algoritmov na určenie počtu odlišných vzorov potrebných na reprezentáciu všetkých variácií rôznych slov v rámci širokej populácie používateľov. Tento výskum sa viac ako desať rokov zdokonaľoval, takže techniky na vytváranie vzorov nezávislých od rečníkov sú teraz dobre definované a široko používané.

Tak ako bolo v 70. rokoch 20. storočia kľúčovým predmetom výskumu izolované rozpoznávanie slov, v 80. rokoch sa výskum zameral na problém spojeného rozpoznávania slov. Tu bolo cieľom vytvoriť robustný systém schopný rozpoznať plynule hovorený reťazec slov (napr. číslice) na základe porovnávania zreťazeného vzoru jednotlivých slov. Bola formulovaná a implementovaná široká škála prepojených algoritmov na rozpoznávanie slov, vrátane dvojúrovňového dynamického programovania Sakoe v Nippon Electric Corporation (NEC) [23], jednopriechodovej metódy Bridle a Brown v Joint Speech Research Unit (JSRU) v Anglicku [24], prístupu k budovaniu úrovní Myersa a Rabinera na úrovni rámca Bell Labs a Rabin [25, 26]. Každý z týchto „optimálnych“ párovacích postupov mal svoje vlastné implementačné výhody, ktoré sa využívali pre širokú škálu úloh.

Výskum reči v 80. rokoch bol charakterizovaný posunom v technológii od prístupov založených na šablónach k metódam štatistického modelovania – najmä prístupu hidden Markov model [27, 28]. Hoci metodológia hidden Markov modeling (HMM) bola dobre známa a pochopená v niekoľkých laboratóriách, až po rozsiahlej publikácii metód a teórie HMM v polovici 80-tych rokov sa táto technika stala široko používanou prakticky v každom výskumnom laboratóriu na rozpoznávanie reči na svete.

Ďalšou „novou“ technológiou, ktorá bola znovu zavedená koncom osemdesiatych rokov, bola myšlienka aplikácie neurónových sietí na problémy s rozpoznávaním reči. Neurónové siete boli prvýkrát predstavené v 50. rokoch 20. storočia, ale spočiatku sa neukázali ako užitočné, pretože mali veľa praktických problémov. V 80. rokoch však došlo k hlbšiemu pochopeniu silných stránok a obmedzení technológie, ako aj vzťahov tejto technológie ku klasickým metódam klasifikácie signálov. Bolo tiež navrhnutých niekoľko nových spôsobov implementácie systémov [29, 30].

Napokon, osemdesiate roky boli desaťročím, v ktorom komunita Defense Advanced Research Projects Agency (DARPA), ktorá sponzorovala rozsiahly výskumný program zameraný na dosiahnutie vysokej presnosti slov na 1 000 slov, priebežné rozpoznávanie reči a úlohu správy databázy. Hlavné výskumné príspevky vyplynuli z úsilia v CMU (najmä známy systém SPHINX) [31], BBN so systémom BYBLOS [32], Lincoln Labs [33], SRI [34], MIT [35] a AT&T Bell Labs [36]. Program DARPA pokračoval do 90. rokov 20. storočia, pričom dôraz sa presunul do prirodzeného jazyka na rozpoznávanie a úloha sa presunula na vyhľadávanie informácií o leteckej doprave. Súčasne sa v telefónnych sieťach čoraz viac využíva technológia rozpoznávania reči na automatizáciu a skvalitnenie služieb operátorov.

* 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 |

Príprava Testovania

Na realizáciu testovania modelov automatického rozpoznávania reči (Speech-to-Text, STT) bol zvolený programovací jazyk Python 3.12.3 kvôli širokej podpore knižníc STT a nástrojov určených na prácu so zvukovými materiálmi.

Kritéria výberu dostupných modelov

Základnými kritériami bola technická dostupnosť, spracovanie a relevancia výsledkov.

Open-source

Všetky vybrané modely sú dostupné ako open-source, čo zabezpečuje:

* transparentnosť fungovania modelu a jeho architektúry
* možnosť analýzy a úprav podľa potreby
* nezávislosť od komerčných licencií
* dlhodobú udržateľnosť v akademickom a vývojárskom prostredí

**Integrácia v Pythone**

Ďalším kľúčovým kritériom bola možnosť priamej integrácie v jazyku Python, ktorý bol zvolený ako hlavný jazyk pre spracovanie testov. Výhody tohto prístupu zahŕňajú:

* širokú podporu na strane knižníc a frameworkov
* jednoduchú implementáciu a prispôsobenie testovacích skriptov
* možnosť efektívneho využitia existujúcich nástrojov na spracovanie audia a hodnotenie výsledkov
* paralelné testovanie

Výber modelov s natívnou podporou Pythonu výrazne urýchlil vývoj testovacieho prostredia a minimalizoval potrebu zložitej integrácie.

Predtrénované modely

Pre účely testovania boli vybrané predtrénované modely, teda modely, ktoré boli už vopred naučené na rozsiahlych dátových sadách. Tento výber bol motivovaný:

* snahou zabezpečiť reálnu aplikovateľnosť modelov bez potreby ďalšieho rozsiahleho trénovania
* možnosťou okamžitého porovnania kvality modelov v rovnakých podmienkach
* elimináciou vplyvu rôznych stratégií vlastného tréningu na výsledky

Použitie predtrénovaných modelov zároveň reflektuje praktické potreby väčšiny používateľov STT technológií, ktorí typicky využívajú modely v stave pripravenom na použitie.

Speech Recognition

Knižnica speech\_recognition bola zvolená vďaka jej jednoduchosti a podpore viacerých backendov, vrátane cloudových a lokálnych modelov.

Konkrétne boli využité:

* Google Web Speech API – bezplatné cloudové API pre rozpoznávanie reči, ktoré poskytuje vysokú presnosť a rýchlu odozvu
* Whisper API cez adaptáciu v speech\_recognition – umožňuje využívať open-source model Whisper od OpenAI v rámci známeho a jednoduchého rozhrania

Dôvody výberu:

* rýchla integrácia a jednoduché ovládanie pomocou jazyka Python
* možnosť porovnať lokálne a cloudové riešenia v jednotnom rozhraní
* získanie referenčných výsledkov od renomovaného cloudového poskytovateľa (Google) bez potreby komplexného nastavovania

Whisper

Whisper od OpenAI je open-source model trénovaný na rozsiahlych multilanguage dátových sadách.

Použitý bol priamo prostredníctvom knižnice openai-whisper alebo cez adaptáciu v knižnici speech\_recognition

Dôvody výberu:

* vysoká odolnosť voči šumu a rôznym nárečiam
* dobrá dostupnosť predtrénovaných verzií (small, medium, large)
* presnosť aj na nahrávkach s nižšou kvalitou
* možnosť lokálneho testovania bez potreby internetového pripojenia
* možnosť porovnania modelu použitom v knižniciach openai-whisper a speech\_recognition

Wav2Vec

Model Wav2Vec2.0 Large od Facebook AI (Meta) je jedným z najvýznamnejších modelov pre rozpoznávanie reči založených na self-supervised learning. Taktiež bol pre porovnanie zvolený základný model Wav2Vec2.0 Base.

Bol použitý cez knižnicu Hugging Face transformers

Dôvody výberu:

* predstaviteľ špičky v oblasti self-supervised učenia reči
* vysoká presnosť pri relatívne nízkom množstve trénovacích dát
* jednoduchá integrácia cez Python API

Príprava modelov

Pre každý z testovaných modelov boli nainštalované potrebné knižnice a závislosti pomocou správcu balíkov pip.

Použité balíky:

* SpeechRecognition (vrátane podpory pre Google Web Speech API a Whisper)
* Whisper (openai-whisper)
* Transformers (pre Wav2Vec2 model od Facebook AI)
* Torch a Torchaudio s podporou CUDA, Librosa (základné knižnice pre prácu s modelmi a zabezpečenie GPU akcelerácie)

Všetky knižnice a ich dependencies boli štandardne uložené do súboru requirements.txt pre ich jednoduchšiu opakovanú inštaláciu na iných zariadeniach.

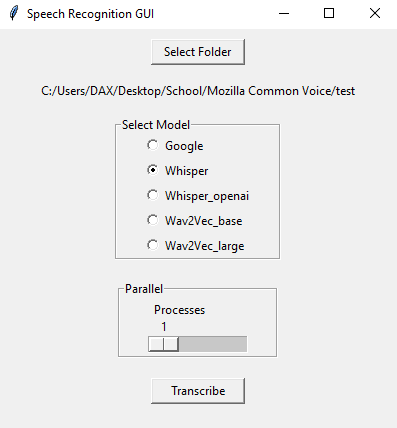
Pre účely správy kódu, verziovania a jednoduchého zdieľania projektu bol vytvorený vlastný Git repozitár a zmeny boli pravidelne nahrávané na GitHub do verejne dostupného repozitára.

Moduly pre prácu s modelmi

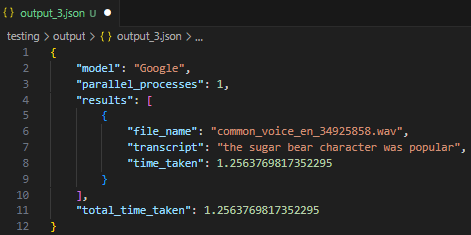
Pre každú použitú knižnicu bol vytvorený jeden modul, ktorý obsahoval triedu s  inicializovaným modelom, ktorý si používateľ pre testovanie zvolil. Model ostával pripravený v pamäti, aby sa zamedzilo opätovnej inicializácií pre každý testovaný súbor a zbytočnému spomaleniu testovania.

Hlavný script

Hlavný skript obsahuje užívateľské rozhranie (GUI) založené na štandardnej knižnici tkinter, ktoré umožňuje voľbu zložky s testovacím materiálom, testovaného modelu, počtu paralelných procesov a možnosť spustenia testovania.



Na začiatku skriptu sa importujú triedy pre prácu s jednotlivými modelmi, ktoré je možné inicializovať podľa užívateľovho výberu. Používateľ si pomocou jednoduchého GUI vyberie priečinok so zvukovými súbormi cez filedialog.askdirectory() a textový label následne zobrazuje cestu ku zvolenému priečinku. Voľba počtu paralelných procesov umožňuje urýchliť testovanie na viacerých jadrách CPU (čo je však efektívne využiteľné iba pri niektorých modeloch). Skript po spustení testov paralelne inicializuje zvolený počet modelov a rovnako paralelne začne spracúvať zvukové súbory v priečinku. Po spracovaní jednotlivého súboru sa zaznamenáva, názov súboru, prevedený text a čas spracovania. Tieto informácie sú priebežne ukladané do súboru JSON ktorý je automaticky vygenerovaný pre každý test, pod názvom „output“ a číselným indexom na konci. Okrem týchto výsledných prevodoch, ktoré sú ukladané do štruktúry {"results": [ ... ]}, sú do súboru zaznamenané aj dáta ako meno použitého modelu, počet paralelných procesov a na konci testovania aj úplné trvanie testov, nakoľko sa pri paralelnom spracovaní dát nezhoduje so súčtom trvania jednotlivých testov.



Príprava testovacieho materiálu

Pre potreby testovania modelov automatického rozpoznávania reči bol použitý Mozilla Common Voice dataset, ktorý predstavuje jednu z najväčších verejne dostupných databáz hovoreného slova.

Charakteristika datasetu:

* veľkosť: približne 3670 hodín zvukových nahrávok vo formáte .mp3 v angličtine, 23 hodín v slovenčine a 268 hodín v češtine
* dĺžka vzoriek: priemerná dĺžka jednotlivých nahrávok je približne 5 sekúnd
* jazyky: dataset pokrýva 134 jazykov vrátane angličtiny, slovenčiny, češtiny a nemčiny
* nahrávky sú priebežne overované dataset obsahuje aj súbor vo formáte .tsv s validovaným textom jednotlivých nahrávok, čo uľahčuje proces kontroly správnosti prevodu

Mozilla Common Voice poskytuje dostatočné množstvo testovacieho materiálu na prevod reči na text. Aby bola zabezpečená kvalita referenčných údajov a spoľahlivosť výsledkov, bol vybraný tzv. delta segment datasetu. Delta segment v Mozilla Common Voice označuje podmnožinu audionahrávok, ktorá obsahuje iba novšie pridané nahrávky. Bol vybraný delta segment Common Voice Delta Segment 12.0, ktorý obsahoval 6.717 validovaných zvukových súborov z celkových 40.707, čo predstavuje vyše 10 hodín validovaných zvukových nahrávok.

Na spracovanie stiahnutého datasetu bol vytvorený skript v jazyku Python, ktorý do vstupu dostane súbor stiahnutých nahrávok a následne konvertuje nahrávky z formátu mp3 do formátu wav, pre jednoduchšie spracovanie modelmi. Modely ako Google prevod reči zo súborov vo formáte mp3 nepodporujú a preto bol zvolený formát ktorý bol kompatibilný so všetkými modelmi. Skript zároveň vyhodnocuje, či sa názov súboru nachádza v súbore validated.tsv a teda jeho text je validovaný a je možné správnosť prevodu následne vyhodnocovať; v opačnom prípade, konvertovanie vynechá. Konvertované nahrávky sú následne uložené do samostatného súboru a pripravené na testovanie.

**TODO** 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

[8] D. B. Fry, “Theoretical Aspects of Mechanical Speech Recognition”; and P. Denes, “The Design and Operation of the Mechanical Speech Recognizer at University College London,” J. British Inst. Radio Engr., 19, 4: 211–229, 1959.

[9] W. R. Forgie and C. D. Forgie, “Results Obtained From a Vowel Recognition Computer Program,” J. Acoust. Soc. Am., 31 (11): 1480–1489, 1959.

[10] J. Suzuki and K. Nakata, “Recognition of Japanese Vowels—Preliminary to the Recognition of Speech,” J. Radio Res. Lab., 37 (8): 193–212, 1961.

[11] T. Sakai and S. Doshita, “The Phonetic Typewriter, Information Processing 1962,” Proc. IFIP Congress, Munich, 1962.

[12] K. Nagata, Y. Kato, and S. Chiba, “Spoken Digit Recognizer for Japanese Language,” NEC Res. Develop., No. 6, 1963.

[13] T. B. Martin, A. L. Nelson, and H. J. Zadell, “Speech Recognition by Feature Abstraction Techniques,” Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964.

[14] T. K. Vintsyuk, “Speech Discrimination by Dynamic Programming,” Kibernetika, 4 (2): 81–88, Jan.–Feb. 1968.

[15] D. R. Reddy, “An Approach to Computer Speech Recognition by Direct Analysis of the Speech Wave,” Tech. Report No. C549, Computer Science Dept., Stanford Univ., September 1966.

[16] V. M. Velichko and N. G. Zagoruyko, “Automatic Recognition of 200 Words,” Int. J. Man-Machine Studies, 2: 223, June 1970.

[17] H. Sakoe and S. Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-26 (1): 43–49, February 1978.

[18] F. Itakura, “Minimum Prediction Residual Applied to Speech Recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-23 (1): 67–72, February 1975.

[19] C. C. Tappert, N. R. Dixon, A. S. Rabinowitz, and W. D. Chapman, “Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery,” Rome Air Dev. Cen, Rome, NY, Tech. Report TR-71-146, 1971.

[15] F. Jelinek, L. R. Bahl, and R. L. Mercer, “Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech,” IEEE Trans. Information Theory, IT-21: 250–256, 1975.

[16] F. Jelinek, “The Development of an Experimental Discrete Dictation Recognizer,” Proc. IEEE, 73 (11): 1616–1624, 1985.

[17] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, “Speaker Independent Recognition of Isolated Words Using Clustering Techniques,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27: 336–349, August 1979.

[18] H. Sakoe, “Two Level DP Matching—A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-27: 588–595, December 1979.

[19] J. S. Bridle and M. D. Brown, “Connected Word Recognition Using Whole Word Templates,” Proc. Inst. Acoust. Autumn Conf., 25–28, November 1979.

[20] C. S. Myers and L. R. Rabiner, “A Level Building Dynamic Time Warping Algorithm for Connected Word Recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., ASSP-29: 284–297, April 1981.

[21] C. H. Lee and L. R. Rabiner, “A Frame Synchronous Network Search Algorithm for Connected Word Recognition,” IEEE Trans. Acoustics, Speech, Signal Proc., 37 (11): 1649–1658, November 1989.

[22] J. Ferguson, Ed., Hidden Markov Models for Speech, IDA, Princeton, NJ, 1980.

[23] L. R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Proc. IEEE, 77 (2): 257–286, February 1989.

[24] R. P. Lippmann, “An Introduction to Computing with Neural Nets,” IEEE ASSP Mag., 4 (2): 4–22, April 1987.

[25] A. Weibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, “Phoneme Recognition Using Time-Delay Neural Networks,” IEEE Trans. Acoustics, Speech, Signal Proc., 37: 393–404, 1989.

[26] K. F. Lee, H. W. Hon, and D. R. Reddy, “An Overview of the SPHINX Speech Recognition System,” IEEE Trans. Acoustics, Speech, Signal Proc., 38: 600–610, 1990.

[27] Y. L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, S. Roucos, and R. M. Schwartz, “BYBLOS: The BBN Continuous Speech Recognition System,” Proc. ICASSP 87, 89–92, April 1987.

[28] D. B. Paul, “The Lincoln Robust Continuous Speech Recognizer,” Proc. ICASSP 89, Glasgow, Scotland, 449–452, May 1989.

[29] M. Weintraub et al., “Linguistic Constraints in Hidden Markov Model Based Speech Recognition,” Proc. ICASSP 89, Glasgow, Scotland, 699–702, May 1989.

[30] V. Zue, J. Glass, M. Phillips, and S. Seneff, “The MIT Summit Speech Recognition System: A Progress Report,” Proc. DARPA Speech and Natural Language Workshop, 179–189, February 1989.

[31] C. H. Lee, L. R. Rabiner, R. Pieraccini, and J. G. Wilpon, “Acoustic Modeling for Large Vocabulary Speech Recognition,” Computer Speech and Language, 4: 127–165, 1990.

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.