TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Tvorba modelů pro přepis řeči v italštině

Bakalářská práce

Studijní program: B0613A140005 – Informační technologie
Studijní obor: B0613A140005IS – Inteligentní systémy

Autor práce: **David Šafařík**
Vedoucí práce: doc. Ing. Petr Červa, Ph.D.

Liberec 2022

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Zadání bakalářské práce

Tvorba modelů pro přepis řeči v italštině

Jméno a příjmení: **David Šafařík**
Osobní číslo: M19000043
Studijní program: B0613A140005 Informační technologie
Specializace: Inteligentní systémy
Zadávající katedra: Ústav informačních technologií a elektroniky
Akademický rok: **2021/2022**

Zásady pro vypracování:

1. Seznamte se s problematikou automatického rozpoznávání řeči z pohledu akustického a jazykového modelování a s metodami pro vyhodnocování úspěšnosti rozpoznávání řeči.
2. Proveďte rešerši dostupných zdrojů dat umožňujících vytvořit akustický a jazykový model pro italštinu (jedná se o akustická a textová data, slovníky a nástroje pro fonetickou transkripci).
3. Jednotlivé zdroje dat sjednoťte a popřípadě vhodně rozšiřte. Následně použijte co největší množství dostupných dat k vytvoření výše uvedených modelů pro italštinu.
4. Experimentálně vyhodnoťte kvalitu vytvořených modelů na vhodné testovací sadě.

Rozsah grafických prací:	dle potřeby dokumentace
Rozsah pracovní zprávy:	30-40
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	Čeština

Seznam odborné literatury:

[1] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR, May 2001.

[2] http://htk.eng.cam.ac.uk

Vedoucí práce:	Ing. Petr Červa, Ph.D. Ústav informačních technologií a elektroniky
Datum zadání práce:	12. října 2021
Předpokládaný termín odevzdání:	16. května 2022

L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
děkan

prof. Ing. Ondřej Novák, CSc.
vedoucí ústavu

V Liberci dne 19. října 2021

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

15. 5. 2022 David Šafařík

Poděkování

Rád bych poděkoval panu doc. Ing. Petru Červovi, Ph.D. za jeho pomoc a vedení při tvorbě této bakalářské práce. Zároveň bych chtěl poděkovat svému bratrovi Ing. Radku Šafaříkovi, Ph.D. za cenné rady a pomoc při vypracování této práce.

Tvorba modelů pro přepis řeči v italštině

Abstrakt

Tato bakalářská práce se zabývá tvorbou modelů pro přepis řeči v italštině. Rešerše této práce se zabývá popisem současného stavu systémů pro automatické rozpoznávání řeči (ASR). Automatické rozpoznávání řeči je nejprve popsáno obecně, následně se popis zaměřuje na modulární architekturu a end-to-end (E2E) architekturu systémů ASR. Následující kapitola je věnována popisu italštiny z obecného, gramatického, ale převážně z fonetického pohledu. Praktická část práce popisuje nejprve vývoj modulárního systému ASR. Postupně je tak popsáno vytvoření jednotlivých dílčích částí modulárního systému (výslovnostní slovník, jazykový model, akustický model). Dále je popsán vývoj E2E systému ASR, který zahrnuje popis shromáždění a úpravy volně dostupných řečových databází pro trénování systému a následně popis tvorby vlastní trénovací databáze k rozšíření trénovací množiny dat. Na konec jsou uvedeny výsledky vyhodnocení vytvořených modelů pro přepis řeči v italštině. Nejlepší vytvořený model dosahuje přesnosti přepisu 87,3 % na nahrávkách z nezávislé ručně anotované testovací databáze. Zároveň je výsledný systém ASR porovnán s komerčním systémem ASR, který je k dispozici v rámci cloudové platformy MS Azure.

Klíčová slova: automatické rozpoznávání řeči, skrytý Markovský model, end-to-end, hluboké neuronové sítě, hluboké učení, italština, jazykové modelování, akustické modelování

Creation of models for speech transcription in Italian

Abstract

This bachelor thesis is concerned with the creation of models for speech transcription in the Italian language. The research part of this thesis describes the current state of systems for automatic speech recognition (ASR). It begins with a general description of ASR, after which the modular and end-to-end (E2E) architectures of ASR systems are described. The next chapter describes the Italian language from the general, gramatical but mostly phonemic viewpoint. The practical part of this thesis begins with the description of the ASR modular system. The creation of all the parts needed to create the modular system is described (pronounciation lexicon, language model, acoustic model). Afterwards the development of the E2E ASR system is described which encompasses gathering and modification of free language databases followed by a description of the creation of a speech database for further training of the E2E system. Finally the results of the evaluation of created models are presented. The best created model has an accuracy of transcription of 87.3 % evaluated on an independent set of recordings annotated by hand. The final ASR system is also compared to the commercial ASR system within the cloud framework Microsoft Azure.

Keywords: Automatic Speech Recognition, hidden Markov model, end-to-end, deep neural networks, deep learning, Italian, language model, acoustic model

Obsah

Seznam zkratek

TUL	Technická univerzita v Liberci
FM	Fakulta mechatroniky, informatiky a mezioborových studií Technické univerzity v Liberci
ASR	Automatic speech recognition
HMM	Hidden Markov Model
GMM	Gaussian mixed model
NN	Neural network
DNN	Deep neural network
AM	Akustický model
LM	Jazykový model
FSMN	Feedforward Sequential Memory Networks
E2E	End to End
WER	Word Error Rate
G2P	Grapheme to phonem
IPA	International phonetic alphabet
X-SAMPA	The Extended Speech Assessment Methods Phonetic Alphabet
FBC	Filter bank coefficients
ST	Speech translation
MLS	Multilingual Librispeech
YT	YouTube
CV	Common Voice
VF	VoxForge
TED	Multilingual TEDx
CER	Character error rate
ACC	Accuracy

Seznam obrázků

Seznam tabulek

Úvod

Automatické rozpoznávání řeči (angl. automatic speech recognition, ASR) je proces, při kterém je signál mluvené řeči převeden na text či jinou vhodnou reprezentaci. Běžný systém ASR má na vstupu akustický signál, který je převeden do lépe zpracovatelné formy, zanalyzován pomocí nějakého algoritmu či modelu, a na výstupu poskytne textový přepis. Tato práce se zabývá vývojem právě takového systému pro italský jazyk.

Modulární systémy ASR tradičně kombinují skrytý Markovův model (angl. hidden Markov model, HMM) a gaussovský mixturový model (GMM). V roce 2012 však začaly vznikat systémy kombinující HMM a hluboké neuronové sítě (angl. deep neural networks, DNN) [1]. Všechny tyto hybridní systémy vyžadují oddělené trénování částí pro akustické modelování, pro výslovnostní slovník a pro modelování jazyku.

V poslední době vzniká snaha posunout se od modulárních systémů směrem k tzv. end-to-end (E2E) systémům, které nevyžadují trénování několika oddělených modulů, ale právě jednoho jednotného. Velkou výhodou E2E systémů oproti modulárním je nižší náročnost jejich vývoje. Vývojáři modulárních systémů se totiž musí dopodrobna seznámit s rozpoznávaným jazykem, což u E2E systémů není nutné. Zároveň může dlouho trvat odladit jednotlivé části modulárního systému.

Pro vytvoření obou typů ASR systémů je nutné mít k dispozici co největší množství záznamů řeči s odpovídajícími textovými přepisy. V dnešní době existují různé projekty, které mají za úkol právě takováto data shromažďovat, a některé takové projekty následně shromážděná data zpřístupňují veřejnosti v podobě ucelené řečové databáze. Příkladem může být projekt Common Voice[1]. Většina takovýchto databází je ale zpoplatněna za velmi vysoké částky[2]. Někdy je tak vhodné všechny volně dostupné řečové databáze použít pro natrénování základní verze systému ASR, který následně slouží pro těžení dat z různých zdrojů, ze kterých je následně vytvořena vlastní řečová databáze.

V současné době existuje nemalé množství cloudových služeb, které nabízí převod řeči na text. Cloudová platforma Azure[3] od firmy Microsoft nabízí mimo jiné služby

[1] https://commonvoice.mozilla.org/
[2] https://catalogue.elra.info/
[3] https://azure.microsoft.com/cs-cz/

i přepis řeči ve více než 100 jazycích. Platforma Google Cloud[4] od firmy Google nabízí podobné služby.

Tyto platformy sice nabízí velmi přesný převod řeči na text, nicméně jejich využití je placené, a pro běžného uživatele může být jejich využití příliš drahé. Levnější alternativou může být například služba convertspeech[5]. Pro diktační účely pak služba speechnotes[6]. Konkrétně pro italštinu pak dále existují tyto systémy:

Ve článku *The 2014 KIT IWSLT speech-to-text systems for English, German and Italian* [2] autoři využívají modulární architektury pro tvorbu systému rozpoznávání řeči. Pro trénování systému využili řečovou databázi Quaero, záznamy z přednášek TED talks a různé neřečové nahrávky. Pro přepis italštiny dosáhl vytvořený systém word error rate (WER) 37,1 %.

Autoři *Phoenix: Deep Speech Based Automatic Speech Recognition System for Italian Language* [3] vytvořili E2E systém pro přepis italštiny, pro jehož trénování využili různé volně dostupné řečové databáze (VoxForge, CLIPS, SI-CALLOPE, Common Voice a audioknihy). Pro vytvoření systému využili projektu DeepSpeech[7] [4]. Vytvořený systém, který nazvali Phoenix, dosáhl WER 13,8 %.

Cílem této práce je vytvoření ASR modelů pro přepis italštiny za účelem rozšíření portfolia platformy pro nepřetržitý monitoring televizního a rádiového vysílání. Tato platforma vznikla mimo jiné i na pracovišti Laboratoře počítačového zpracování řeči na technické univerzitě v Liberci[8] a již podporuje největší slovanské a evropské jazyky.

V této práci jsou využívány jak modulární systémy založené na hybridní HMM/DNN architektuře, tak i E2E modely. Pro trénování těchto systémů je cílem nashromáždit velké množství textových a akustických dat (s odpovídajícími textovými přepisy) z různých zdrojů. Tato data je nutné vhodně upravit a pročistit.

Pro vytvoření modulárního systému je nutné vytvořit textový korpus, výslovnostní slovník a odpovídající pravidla pro fonetickou transkripci, jazykový model, a zároveň akustický model. Pro vytvoření E2E systému je nutné pouze nashromáždit a vhodně upravit nahrávky řeči s odpovídajícími textovými přepisy.

[4] https://cloud.google.com/
[5] https://convertspeech.com/it-IT/
[6] https://speechnotes.co/
[7] https://github.com/mozilla/DeepSpeech
[8] http://speechlab.tul.cz/

1 Systémy automatického rozpoznávání řeči

1.1 Obecný přístup systémů automatického rozpoznávání řeči

Obecně lze říci, že systémy ASR [5] přijímají na vstupu audio signál, který dále zpracují, a vyprodukují odpovídající sekvenci cílových proměnných (též label) na výstupu. Zpracování vstupního audio signálu je prováděno pomocí parametrizačního modulu, který rozdělí signál na rámce (též frame), z nichž pro každý následně počítá příznakové vektory. Výstup ze systému ASR může být reprezentován znaky či celými slovy, které se nacházejí ve slovníku L, jenž přináleží samotnému systému.

Formálně lze tuto problematiku zapsat jako:

$$\hat{W} = \underset{W \in \mathcal{L}*}{\operatorname{argmax}}\, p(W|X) \tag{1.1}$$

Kde $\hat{W} = \{w_1, \ldots, w_N\}$ je výstupní sekvence N slov, $X = \{x_1, \ldots, x_T\}$ je sekvence vstupních příznakových vektorů délky T, $p(W|X)$ je podmíněná pravděpodobnost W za předpokladu, že známe X, a $\mathcal{L}*$ je množina všech možných sekvencí, které mohou být vytvořeny ze slov ve slovníku L.

Existují dvě hlavní architektury systémů ASR, které umožňují odhad $\hat{W}$ podle vzorce 1.1. První takovou je modulární architektura, která využívá Bayesův vzorec. Druhou takovou architekturou je tzv. end-to-end architektura, která má přímočařejší přístup. Obě architektury jsou představeny v následujících podkapitolách.

1.2 Modulární systém automatického rozpoznávání řeči

Modulární architektura ASR využívá Bayesův vzorec, a v takovém případě lze vzorec 1.1 přepsat jako:

$$\begin{aligned}\hat{W} &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} \frac{p(X|W)p(W)}{p(X)} \\ &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} p(X|W)p(W)\end{aligned} \tag{1.2}$$

Kde $p(X)$ je pravděpodobnost pozorování X, $p(W)$ je apriorní pravděpodobnost W, a $p(X|W)$ je podmíněná pravděpodobnost X, za předpokladu, že známe W.

Systémy ASR s touto architekturou většinou využívají skrytý Markovův model (HMM). Taková architektura se pak skládá ze tří nezávislých částí: akustický, výslovnostní a jazykový model. Všechny tři části jsou jazykově závislé. Dále jsou podrobněji představeny všechny tři části:

Akustický model (AM) funguje na principu mapování příznakových vektorů na příslušnou posloupnost fonémů. Výslovnostní model slouží k vytvoření výslovnostního slovníku, který obsahuje slova rozpoznatelná rozpoznávačem s jejich příslušnou fonetickou podobou. Jazykový model (LM) poskytuje pravděpodobnost výskytu jednotlivých slov v dané posloupnosti.

Zmíněného modelu HMM je však využito pouze v AM, pro tento případ lze tak přepsat vzorec 1.2 jako:

$$\begin{aligned}\hat{W} &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} p(X|W)p(W) \\ &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} p(X, W) \\ &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} p(X, S, W)\end{aligned} \tag{1.3}$$

Kde $S = \{S_t \in \{1, \ldots, J\} | t = 1, \ldots, T\}$ je posloupnost stavů HMM a J je počet stavů modelu HMM. Vzorec 1.3 lze pak dále rozvést následnovně:

$$\begin{aligned}\hat{W} &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} \sum_S p(X, S, W) \\ &= \underset{W \in \mathcal{L}*}{\operatorname{argmax}} \sum_S p(X|S, W)p(S|W)p(W)\end{aligned} \tag{1.4}$$

Díky podmíněné nezávislosti lze odhadovat $p(X|S, W)$ jako $p(X|S, W) \approx p(X|S)$, v tom případě lze vzorec 1.4 přepsat jako:

$$\hat{W} = \underset{W \in \mathcal{L}*}{\operatorname{argmax}} \sum_S p(X|S)p(S|W)p(W) \tag{1.5}$$

Kde $p(X|S)$ představuje akustický model, $p(S|W)$ představuje výslovnostní model, jenž je zakomponován ve slovníku výslovností, a $p(W)$ představuje jazykový model.

Samotný model HMM předpokládá nezávislost pozorování, tj. pozorování v jakémkoliv čase závisí pouze na skrytém stavu v daném čase. Když vezmeme v potaz tuto skutečnost a použijeme řetízkové pravidlo pravděpodobnosti, tak můžeme $p(X|S)$ rozepsat jako:

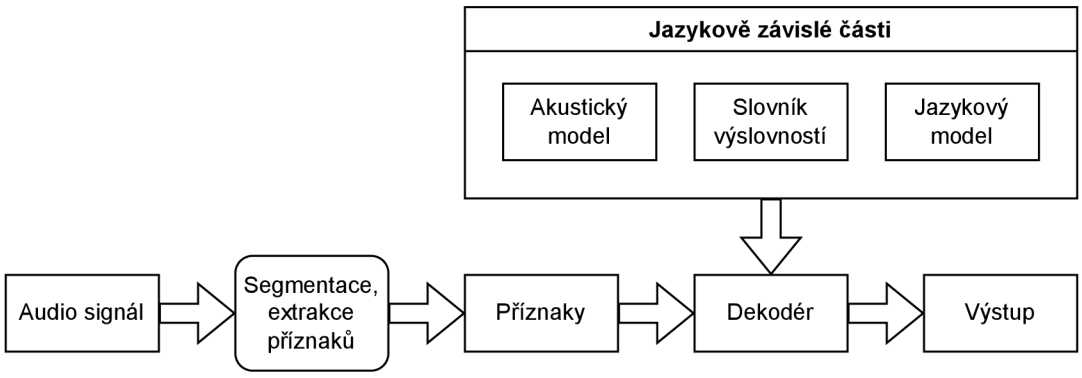
$$\begin{aligned} p(X|S) &= \prod_T p(x_t|x_1, \ldots, x_{t-1}, S) \\ &\approx \prod_T p(x_t|s_t) \end{aligned} \tag{1.6}$$

Pro výpočet pravděpodobnosti $p(x_t|s_t)$ existují dvě různé architektury akustického modelu. První z nich využívá Gaussian mixture models (GMM). Taková topologie se nazývá GMM-HMM. Architektury state-of-the-art využívají hluboké neuronové sítě (DNN) [1]. DNN-HMM topologie následně počítá $p(x_t|s_t)$ výpočtem z aposteriorní pravděpodobnosti $p(s_t|x_t)$ a apriorní pravděpodobnosti $p(s_t)$ podle vzorce:

$$p(x_t|s_t) \propto \frac{p(s_t|x_t)}{p(s_t)} \tag{1.7}$$

Existuje mnoho různých architektur neuronových sítí pro akustický model.

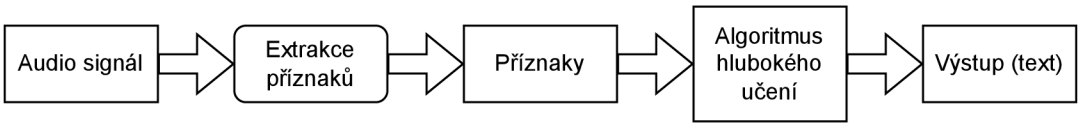
Mimo všech popsaných částí musí takovýto systém obsahovat ještě dekodér, který dekóduje optimální $\hat{W}$ přes všechna řešení $\mathcal{L}*$ pomocí všech tří dílčích modelů [6]. Schéma modulárního systému ASR je znázorněno na obrázku 1.1.



Obrázek 1.1: Schéma modulárního systému ASR

1.3 End-to-end systém automatického rozpoznávání řeči

End-to-end (E2E) systémy [7] přímo mapují sekvence vstupních akustických charakteristik na sekvence grafémů či celých slov. Výhodou E2E systémů oproti modulárním systémům je větší přímočarost, jak je znázorněno v 1.2. U modulárních systémů je totiž nutné trénovat akustický, výslovnostní a jazykový model zvlášť (obrázek 1.1). Zároveň jsou u modulárních systémů kladeny větší nároky na jejich vývojáře.



Obrázek 1.2: Schéma E2E systému ASR

Z toho vyplývá, že E2E architektura, v porovnání s modulární architekturou, razantně zjednodušuje proces spojený s automatickým rozpoznáváním řeči. E2E eliminuje potřebu manuálně označovat informace, jazyková a výslovnostní pravidla se naučí neuronová síť sama. Existují dva hlavní principy, které E2E systémy využivají: connectionist temporal classification (CTC) a attention model.

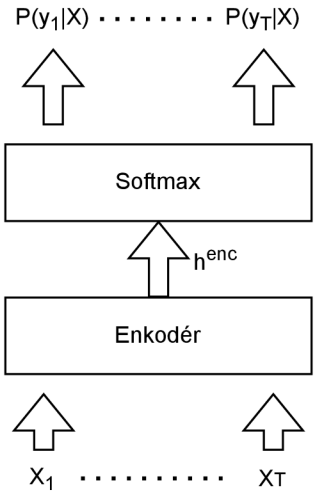
1.3.1 Metoda CTC

Metoda CTC [7] byla poprvé představena v [8] jako způsob trénování akustického modelu s nesegmentovanými sekvencemi dat. CTC tedy eliminuje potřebu předsegmentace trénovacích dat a následného post-processingu výstupních klasifikací. V kontextu ASR předsegmentace znamená zarovnání na úrovní rámců (frame). Post-processing představuje namapování rámců na slova. Schéma struktury CTC je vyznačeno na obrázku 1.3.

Trénování akustického modelu za použití CTC jako loss funkce využívá tzv. *blank* [9] symbol, jenž značí, že daný frame nemá predikovanou hodnotu.

Za předpokladu, že známe vstupní vektor x, tak pravděpodobnost, že výstupní posloupnost bude y, je dána vzorcem:

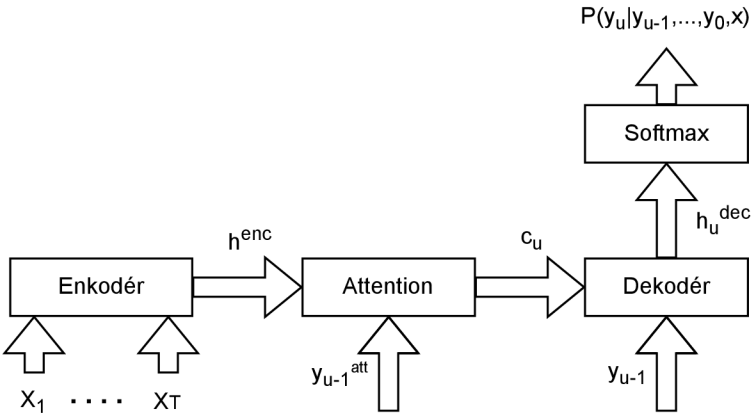
$$p(y|x) = \sum_{\hat{y} \in B(y,x)} \prod_{t=1}^{T} p(\hat{y}_t|x) \tag{1.8}$$



Obrázek 1.3: Struktura CTC [7]

1.3.2 Architektura enkodér-dekodér s Attention mechanismem

Hlavní myšlenka attention mechanizmu je překonat limitace tradiční enkodér-dekodér struktury [10], která je závislá na vektoru neměnné délky. Architektura attention modelu je vyobrazena na obrázku 1.4.



Obrázek 1.4: Struktura attention modelu [7]

Enkodér (analogický k AM) přetváří vstupní signál do reprezentace vyššího řádu podle vztahu:

$$h_i^{enc} = Encoder(x_i) \tag{1.9}$$

Attention (model pro zarovnání) modul identifikuje rámce, které jsou důležité pro generování aktuálního výstupu. Počítá míru podobnosti mezi dekodérem a každým

rámcem enkodéru:

$$\begin{aligned} e_{u,t} &= score(h_{u-1}^{att}, h_t^{enc}) \\ a_{u,t} &= \frac{\exp(e_u, t)}{\sum_{t'}^{T} \exp(e_u, t')} \\ c_u &= \sum_{t=1}^{T} a_{u,t} h_t^{enc} \end{aligned} \tag{1.10}$$

Dekodér provádí auto regresivní operace odhadováním výstupních tokenů jako funkci předchozích odhadů. Výstup skrytého stavu h_t^{dec}:

$$h_t^{dec} = Decoder(y_{t-1}, c_{t-1}) \tag{1.11}$$

Generování odhadu třídy:

$$y_t = softmax(U[h_t^{dec}; c_t]) \tag{1.12}$$

2 Italština z pohledu automatického rozpoznávání řeči

V této kapitole jsou popsány detaily italštiny důležité z pohledu tvorby systémů ASR. Pro tvorbu takového systému není vyloženě nutné být expertem na daný jazyk, ale stále je přínosné seznámit se s jeho základními rysy a charakteristikami, a to zejména při tvorbě klasického systému ASR.

V Evropské unii je italština druhým nejrozšířenějším jazykem s 67 miliony rodilými mluvčími, a 13,4 milionů občanů EU ovládá italštinu jako druhý jazyk. Společně se všemi italsky hovořícími lidmi ve zbytku světa činí odhad 85 milionů lidí hovořících italštinou [11]. Italština je úředním jazykem v Itálii, San Marinu a Švýcarských kantonech Ticino a Grigioni. Ve Vatikánu je druhým oficiálním jazykem, stejně jako v některých částech Slovinska a Chorvatska. Ve světě je italština používaná skupinami migrantů například v USA, Brazílii nebo Austrálii. Díky takovému množství italsky mluvících osob existuje nemalé množství dostupných dat použitelných pro vývoj ASR systémů.

V následujících podkapitolách je blíže popsána fonetika, ortografie, morfologie a gramatika italštiny. Vzhledem k rozmanitosti italských dialektů je jedna podkapitola věnována i této problematice.

2.1 Fonetická podoba italštiny

Italština využívá 23 souhláskových fonémů (viz tabulka 2.1) a 7 samohláskových fonémů (viz tabulka 2.2) [12][13]. Vzhledem k tomu, že některé z fonémů uvedených zde v tabulkách mohou být pro českého čtenáře neznáme, tak si některé z nich přiblížíme z pohledu češtiny:

- /ɲ/ se podobá českému *ň*;
- /t͡s/ se vyslovuje jako *ts*, v podstatě *c*;
- /d͡z/ se vyslovuje jako *dz*;

- /t͡ʃ/ se vyslovuje jako české *č*;
- /d͡ʒ/ se vyslovuje jako české *dž*;
- /ʃ/ se vyslovuje jako české *š*;
- /ʎ/ se sice nevyskytuje v češtině, ale ve slovenštině se objevuje jako *ľ*, tedy měkké l;
- v češtině nerozlišujeme samohlásky /ɛ/ a /ɔ/ od samohlásek /e/ a /o/, ale v italštině jsou rozlišeny od samohlásek /e/ a /o/ větším důrazem.

Tabulka 2.1: Fonetická abeceda souhlásek italštiny

Souhlásky	**Labiála**	**Dentála/ Alveolára**	**Postalveolára/ Palatála**	**Velára**
Nazála	m	n	ɲ	
Ploziva	p b	t d		k g
Afrikáta		t͡s d͡z	t͡ʃ d͡ʒ	
Frikativa	f v	s z	ʃ	
Aproximanta			j	w
Laterála		l	ʎ	
Vibranta		r		

Tabulka 2.2: Fonetická abeceda samohlásek italštiny

Samohlásky	**Přední**	**Střední**	**Zadní**
Zavřená	i		u
Polozavřená	e		o
Polootevřená	ɛ		ɔ
Otevřená		a	

Souhlásky mohou tvořit jediný foném nebo nebo být geminované, pokud leží mezi dvěma samohláskami, nebo mezi samohláskou a aproximantem. Geminované souhlásky zkrátí předchozí samohlásku a první prvek gemináty přechází v plozivní souhlásku. Výjimkou jsou fonémy /j/, /w/ a /z/, které nemohou být geminované. *S* a *z* jsou na začátku slov foneticky nerozlišitelné, když přímo předchází další souhlásce.

V italštině není žádné rozlišení mezi krátkými a dlouhými samohláskami. Nicméně samohlásky v přízvučných otevřených slabikách jsou prodlouženy na konci

intonační fráze, pokud nejsou na konci slova. Pokud spolu sousedí dvě identické samohlásky na konci slova a na začátku slova následujícího, tak se nerozlišují jejich slabiky, a při rychlé řeči mohou splynout v jeden samohláskový foném. Polozavřené a polootevřené samohlásky jsou rozlišitelné v přízvučných slabikách, jinak je rozdíl neznatelný.

2.2 Písemná a výslovnostní podoba italštiny

Italská abeceda je upravenou formou latinky. Typicky se skládá z 21 písmen, a tedy neobsahuje písmena j, k, w, x, y, která se ale objevují u přejatých slov.

Italský pravopis je téměř fonemický. To znamená, že zápis slov často odpovídá tomu, jak se vyslovují. Existují ale výjimky, které bylo potřeba pro potřeby této práce nastudovat. K tomu lze využít různé písemné zdroje [14][12], případně je nutné na některé výslovnosti přijít vlastními silami poslouchání jazyka. Uveďme si tedy běžné výjimky [13]:

- písmeno *c* představuje foném /k/ na konci slov a před samohláskami a, o, u, ale před samohláskami e nebo i představuje foném /tʃ/;
- písmeno *g* představuje foném /g/ na konci slov a před samohláskami a, o, u, ale před samohláskami e nebo i představuje foném /dʒ/;
- spřežka *gl* představuje foném /ʎ/;
- spřežka *gn* představuje foném /ɲ/;
- písmeno *h* se nečte. Když ale leží za písmenem c nebo g, tak s nimi utvoří spřežku a mění jejich výslovnost na /k/ a /g/ před samohláskami e nebo i;
- spřežky *ci* a *gi* představují fonémy /tʃ/ a /dʒ/ bez fonému /i/, pokud leží před další samohláskou, ale pokud by měl být přízvuk na *i* ve spřežce, tak se /i/ vyslovuje. Jinak *ci* a *gi* představují /tʃ/ a /dʒ/ následované /i/;
- písmeno n většinou představuje foném /n/, ale představuje foném /ɲ/ před písmeny *c* a *g*, pokud se vyslovují jako /k/ a /g/. Pokud se *c* a *g* vyslovují jako /tʃ/ a /dʒ/, tak *n* představuje foném /n/;
- písmeno *q* se vyslovuje jako /k/ a vždy se objevuje před písmenem *u* a jinou samohláskou. Písmeno *u* pak přechází ve /v/;
- písmeno *s* představuje foném /z/, pokud leží mezi dvěma samohláskami;
- písmeno *z* představuje foném /ts/ i /dz/. Neexistují však přesná pravidla, je potřeba naučit se výslovnost pro každé slovo.

Geminát, nebo-li zdvojené souhlásky se v řeči vyznačují prodloužením a větším důrazem. Délka většiny souhlásek se dá takto rozlišovat, výjimku tvoří souhlásky /ʃ/, /dz/, /ts/, /ʎ/ a /ɲ/, které jsou vždy geminovány mezi samohláskami, a /z/, které je naopak vždy krátké. Ukažme si geminát na několika příkladech:

- ve slově *notte* (noc) vznikne před vyslovením /t/ krátká pauza a následně je /t/ vysloveno s větším důrazem. Naopak například ve slově *notizie* (zprávy) se na /t/ žádný důraz neklade;
- ve slově *galleria* (tunel) vznikne před vyslovením /l/ krátká pauza a následně je /l/ vysloveno s větším důrazem. Naopak zase například ve slově *inalazione* (inhalace) se na /l/ žádný důraz neklade.

2.3 Morfologie a gramatika italštiny

Italština patří do rodiny románských jazyků, které se vyvinuly z latiny. Morfologicky lze italštinu řadit jak mezi flektivní, tak analytické jazyky. Z toho vyplývá, že slovní zásoba je u italštiny větší než například u angličtiny, která se řadí mezi analytické jazyky, ale menší než u češtiny, nebo jiných flektivních jazyků.

Flektivní jazyky se vyznačují využitím ohýbání (flexe) pro vyjádření gramatické funkce. Ohýbáním se rozumí skloňování a časování. Flektivní rysy se v italštině objevují díky vývinu z latiny. Lze je pozorovat u sloves, které vyjadřují osobu, číslo a čas ohýbáním koncovek. Italština má celkem osm slovesných časů, čtyři jednoduché a čtyři složené, jejichž tvorba je ale otázkou analytického principu. Slovosled je v italštině dost volný. Například u otázky je často stejný jako u oznamovací věty, a mění se pouze intonace. Zároveň není nutné vyjadřovat podmět, když je zřejmý z kontextu.

Analytické jazyky nepoužívají ohýbání slov, gramatickou funkci vyjadřují pomocí slovosledu a předložek. K analytickému principu se italštinu posouvá zmizením pádů, které se zachovaly pouze u zájmen, jinak byly nahrazeny předložkovými vazbami. Tento princip se také uplatňuje při tvoření složených časů sloves, kdy každý ze čtyř složených časů je vyjádřen vlastním pomocným slovesem.

2.4 Italské dialekty

Nesjednocenost italštiny může představovat překážku při automatickém rozpoznávání řeči. Existuje totiž nepřeberné množství italských dialektů, lze až skoro říci, že každé italské město má vlastní dialekt. Mezi uznávané dialekty jsou řazeny

například toskánský, piemonstký, abruzzský, apulský, umbrijský, aj. Některé dialekty jsou dokonce uznávány za samostané jazyky. V této souvistlosti lze hovořit o sicilštině, sardštině a benátštině. Je ale nutné brát v potaz rozdíl mezi italským dialektem, tedy nářečím italštiny, a italským jazykem, jenž je takový jazyk, který má stejný základ v latině jako italština, ale vyvíjel se paralelně s italštinou.

Vzhledem k tomu, že různé výslovnosti jednotlivých slov mohou představovat problém při automatickém rozpoznávání řeči, tak si dále uvedeme některé rozdíly pro vybrané dialekty.

Toskánský dialekt má několik vlastních dialektů, které člení podle regionů v toskánsku. Jako dialekt se toskánština liší od italštiny hlavně výslovností, ale narozdíl od italštiny má sjednocená vlastní zájmena, u kterých se tedy rozlišuje pouze osoba. Co se týče výslovnostních rozdílů, tak *c* se vyslovuje jako /h/, pokud není zdvojené a je obklopeno samohláskami. Například slovo *formica* (mravenec) se vyslovuje jako *formiha*. V některých případech se ale *c* čte jako /g/. Například ve slově *cabina* (chata), které se čte v toskánštině jako *gabina*. [15]

Piemonstské nářečí je rozšířeno v italském kraji Piemont. Od standardní italštiny se liší například absencí souhlásky *S* (výslovnost jako *š*), která je nahrazována obyčejným *s*. Na rozdíl od italštiny ale piemontština obsahuje tzv. slovesná zájmena, která musí být použita vždy, když věta obsahuje předmět. Dalším z mnoha rozdílů je výslovnost písmene *u*, které se vyslovuje jako německé přehlasované *ü*. [16]

3 Provedené vývojové práce

Tato kapitola popisuje vývojové práce provedené pro tvorbu systémů rozpoznávání řeči pro italštinu. První podkapitola se věnuje popisu vývoje modulárního systému automatického rozpoznávání řeči pro italštinu. Druhá podkapitola se pak věnuje tvorbě end-to-end systému automatického rozpoznávání řeči pro italštinu.

3.1 Modulární systém

V této podkapitole jsou popsány kroky provedené k vytvoření modulárního systému ASR pro italštinu. Tyto kroky zahrnují tvorbu textového korpusu, výslovnostního slovníku, a následně také jazykového a akustického modelu:

- textový korpus slouží jako základ pro vytvoření výslovnostního slovníku a jazykového modelu;
- výslovnostní slovník slouží k modelování fonetické podoby jazyka;
- jazykový model slouží k modelování pravděpodobností výskytu jednotlivých slov z textového korpusu;
- akustický model je trénován ke klasifikaci fonémů v audio nahrávce.

Na závěr podkapitoly je popsáno vyhodnocení vytvořeného modelu, který je porovnán s paralelně vytvořeným modelem E2E (popis v 3.2).

3.1.1 Tvorba korpusu

Pro vytvoření textového korpusu pro jakýkoliv jazyk je potřeba nashromáždit velké množství textů, které se následně sjednotí a je z nich vytvořen slovník výslovností a jazykový model. Takový korpus může být vytvořen nashromážděním volně dostupných dat z internetu, případně lze využít některý z již existujících a veřejně dostupných korpusů.

Aby korpus co nejlépe pokryl rozsah jazyka určeného k rozpoznávání, tak je potřeba nashromáždit dostačující množství textových dat. Pro jazyk s tak rozsáhlou slovní zásobou, jako je italština, je ideální velikost korpusu v řádu jednotek GB. Po nashromáždění textových dat je nutné je dále zpracovat, což zahrnuje normalizaci textů a vyfiltrování nežádoucích znaků.

Zdroje textových dat

První krok tvorby korpusu je hledání zdrojů textových dat. Ideální zdroje jsou takové, které obsahují velké množství textových dat s co nejmenším počtem gramatických chyb. Jako významný zdroj se tedy nabízí webové stránky zpravodajských portálů, které zpravidla prochází jazykovou korekturou a obsahují tak italštinu s malým množstvím gramatických chyb. Kromě toho většinou obsahují široký okruh témat, díky čemuž pokrývají velkou část slovní zásoby. Navíc ještě obsahují mnoho článků, což znamená velké množství dostupných textových dat.

Bez vhledu do prostředí italských zpravodajských portálů se ukázal být dobrým zdrojem pro hledání webových stránek těchto portálů web ABYZ News Links[1], který shromažďuje odkazy na zpravodajské weby jednotlivých států světa.

V rámci této práce se podařilo nashromáždit množství textových dat o velikosti 4 GB. Po dalším zpracování, které je popsáno dále v textu, má výsledný korpus velikost 2,65 GB. V tabulce 3.1 jsou vypsány nejvýznamnější zdroje textových dat použitých pro tvorbu korpusu.

Tabulka 3.1: Zdroje dat pro korpus

Zdroje	**Staženo dat**	**Použito dat**
Napoli Today, Il Friuli, Il Mattino, Il Corriere del Sud, Il Denaro, Italpress, Albenga Corsara, Prima Brescia, Corriere dell'Irpinia, Il Tempo, Gazzetta di Reggio, Gazzetta di Modena, Liberta, Prima Illevante, TG Com 24, Corriere della Sera, Voce di Mantova, RaiNews, Il Resto del Carlino, AGI, Il Giorno, radio24, Liguria Notizie, Roma Today, Avvenire, ...	4 GB	2,6 GB

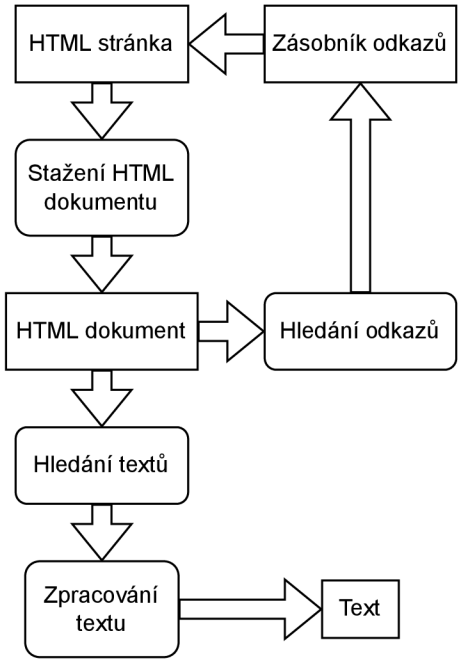
Stahování dat

Pro stahování textů z webových stránek byl vytvořen program, který prohledává danou webovou stránku a ukládá si odkazy na další stránky, které následně také

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

prohledává. Během toho obsah jednotlivých stránek zpracovává a ukládá. Zásobník odkazů drží kromě samotných odkazů informaci o již zpracovaných odkazích. Jako první odkaz je do zásobníku vložen odkaz na hlavní stránku vybraného webu, a postupně jsou zpracovávány všechny přidávané odkazy v zásobníku.

Na právě zpracovávané stránce se nejprve vyhledají všechny odkazy, které směřují na stejný web. Pokud ještě nejsou v zásobníku odkazů, tak jsou do něj přidány. Následně jsou v dokumentu nalezeny všechny textové elementy, které jsou vytaženy z dokumentu, uloženy a následně zpracovány (popis zpracování textů dále v sekci). Celý tento proces je znázorněn na obrázku 3.1.



Obrázek 3.1: Schéma pro stahování textů z webových stránek

Zpracování textů

Pro zlepšení výsledné přesnosti modelu je ještě nutné provést dodatečné zpracování textového korpusu. Tento proces zahrnuje odstranění nežádoucích elementů a zpracování číslic, dat, častých zkratek, jednotek a dalších symbolů.

Nejprve byly z textu odstraněny veškeré HTML a XML tagy, odkazy na webové stránky, e-mailové adresy a jiné nejazykové elementy. Dále bylo sjednoceno kódování jednotlivých textů a byly sjednoceny různé znaky, jako jsou například uvozovky, apostrofy nebo různé písmenné znaky.

Následně byla čísla a data zapsaná v číselné podobě převedena na tokeny, které slouží k označní pro další zpracování. K těmto tokenům byla vytvořena převodní

tabulka, která obsahuje přípustné tvary a výslovnosti pro jednotlivá čísla a data. Tokeny jsou používány z toho důvodu, že některá slova mohou mít několik tvarů. Například slovo „uno“ (jedna) lze nalézt i ve tvaru „una“ pro ženský rod. Zástupný token je ve tvaru *A#*, kde A je textový řetězec symbolizující odpovídající číslo či datum. Příklad tokenu pro číslo 1 je následující: *1#*.

3.1.2 Tvorba výslovnostního slovníku

Z vytvořeného textového korpusu je následně možné vytvořit výslovnostní slovník. Ten obsahuje všechna slova, která mohou být rozpoznána rozpoznávačem. Každé z těchto slov má ve slovníku přiřazenou výslovnost, která je vytvořena podle pravidel, které jsou popsány dále v textu.

Při zpracování textového korpusu v něm byla nalezena slova, která nepatří do italštiny. Jedná se o dvoupísmenná a třípísmenná slova (např. "aaa"), příliš dlouhá slova obsahující podivné znaky či slova neobsahující samohlásky. Takováto slova bylo nutné vyfiltrovat a nebyla tak vybrána do výslovnostního slovníku.

Samotný slovník výslovností je vytvořen frekvenční analýzou slov v korpusu. Slovy se rozumí souvislé řetězce znaků oddělené mezerou. Pro každé unikátní slovo v korpusu je spočítána jeho četnost, podle které jsou pak tato slova seřazena ve slovníku. Ke každému takovému slovu jsou zároveň přiřazeny všechny jeho přípustné výslovnosti. Pro určení dostatečného množství slov vybraných pro výslovnostní slovník se používá míra out-of-vocabulary (OOV). OOV je spočítáno jako poměr počtu unikátních slov v korpusu, která nebyla vybrána do slovníku, k celkovému počtu slov. Podařilo se dosáhnou míry OOV 4,3 %, kterou lze považovat za dostatečnou. Po všech úpravách a vyfiltrování nežádoucích slov bylo do slovníku vybráno přibližně 150 tisíc slov z korpusu. Souhrnné statistiky o výslovnostním slovníku jsou k nalezení v tabulce 3.2.

Tabulka 3.2: Statistika dat ve výslovnostním slovníku

	Slovník výslovností
Počet slov	154513
Počet výslovností	167383
Míra OOV	4,32 %

Fonetická abeceda

Před popisem pravidel pro převod z ortografické podoby na fonetickou je potřeba definovat vlastní fonetickou abecedu. Fonetické studie zpravidla používají pro zápis fonémů mezinárodní fonetickou abecedu (angl. International Phonetic Alphabet,

IPA)[2], která ale není kvůli kódování jednotlivých znaků vhodná pro zápis na běžné české klávesnici. Například znak ɛ se na české klávesnici nevyskytuje.

Pro zpracování na počítači je používána abeceda X-SAMPA[3] (The Extended Speech Assessment Methods Phonetic Alphabet), jenž je rozšířenou verzí abecedy SAMPA. Abeceda X-SAMPA má určený zápis pro jednotlivé znaky z abecedy IPA, nicméně každý takovýto zápis mnohdy obsahuje vícero znaků, což není vhodné pro přehlednost a další zpracování. Pro ukázku zápisu některých fonémů v X-SAMPA:

- foném t͡s se zapíše jako ts;
- foném d͡z se zapíše jako dz;
- foném t͡ʃ se zapíše jako tS;
- foném d͡ʒ se zapíše jako dZ.

Z ukázky je tedy jasné, že X-SAMPA není příliš vhodná pro zápis fonémů v této práci, výhodnější je možnost zápisu každého fonému jediným znakem.

Z praktických důvodů byl tedy pro tvorbu italské fonetické abecedy využit přístup navržený v [17][18] pro českou fonetickou abecedu, který je lehce pochopitelný pro českého editora. Tento přístup je výhodný, protože nabízí:

- možnost zápisu jednotlivých fonémů pomocí jediného znaku, což je praktické při převodu do fonetické podoby a při zpracování fonetického zápisu;
- jednoduchý zápis jednotlivých symbolů na české klávesnici, což je značně praktické pro českého editora.

Fonetické studie a tabulky (2.1, 2.2) v kapitole 2.1 slouží jako základ pro tvorbu fonetické abecedy, kterou znázorňuje tabulka 3.3. V tabulce je znázorněno namapování jednotlivých symbolů vlastní fonetické abecedy na fonetickou abecedu IPA.

[2] https://www.internationalphoneticalphabet.org/
[3] https://commons.wikimedia.org/wiki/X-SAMPA

Tabulka 3.3: Fonetická abeceda pro G2P s mapováním na IPA

G2P	IPA	G2P	IPA	G2P	IPA
a	a	g	g	p	p
b	b	j	j	r	r
c	t͡s	k	k	s	s
C	d͡z	l	l	š	ʃ
č	t͡ʃ	L	ʎ	t	t
Č	d͡ʒ	m	m	u	u
d	d	n	n	v	v
e	e/ɛ	ň	ɲ	z	z
f	f	o	o/ɔ	-	*ticho*

Tvorba výslovnostních pravidel pro převod slov ve slovníku na fonetickou podobu

Pro vytvoření výslovnostní části slovníku je potřeba zadefinovat pravidla pro převod z ortografické podoby slov na fonetickou. Po zadefinování těchto pravidel je možné využít tzv. grapeheme-to-phoneme (G2P) převodník pro samotný převod slov do fonetické podoby. Tabulka 3.4 znázorňuje podobu textu po převodu do fonetické podoby. Celou podobu výslovnostních pravidel popisovaných v této části lze nalézt v příloze A.1.

Pro nedostatek času, jenž byl vyhrazen pro tvorbu modulárního systému, byla pravidla pro převod z ortografické podoby slov na fonetickou podobu vytvořená v rámci této práce zjednodušena. To znamená, že pravidla tak zcela nepokrývají všechny možné výslovnostní detaily italštiny, ale pokrývají alespoň základ, který bylo možné popsat v rámci této bakalářské práce.

Popisu výslovnostních pravidel předcházela studie italštiny. Přestože existují zdroje popisující výslovnostní pravidla italštiny typu kurz italštiny [14], tak značnou část vytvořených pravidel bylo nutné "naposlouchat"ze zvukových záznamů s textovými přepisy. Bližší podoba italštiny z pohledu fonologie a ortografie je popsána v kapitole 2.

Jelikož je italština tzv. ortograficky mělká, což znamená, že často není žádný rozdíl mezi psanou podobou a výslovností slov, tak lze vytvořit pravidla pro fonetickou transkripci, která pouze sledují nejbližší okolí každého znaku. Transkripce pak probíhá tím způsobem, že ortografická podoba slova je sledována znak po znaku, a každý znak či řetězec znaků je v případě nalezení vhodného produkčního pravidla

nahrazován za řetězec znaků určený produkčními pravidly, která mají následující podobu:

$$A \rightarrow B/C\_D \tag{3.1}$$

Tento zápis znamená, že pokud řetězci A předchází řetězec C a je následován řetězcem D, tak je nahrazen řetězcem B. Pro jednodušší zápis mohou řetězce C a D obsahovat i zástupné skupiny obsahující více znaků (např. skupina pro samohlásky), které jsou následně automaticky rozděleny na více pravidel. Tyto zástupné skupiny jsou v souboru s pravidly popsány na úplném začátku textu. Definice zástupné skupiny pro znělé nepárové souhlásky a samohlásky, které jsou zastoupeny symboly *J* a *E*, vypadá následovně:

```
J:j,l,m,n,r
E:a,e,i,o,u
```

Po definování všech zástupných skupin mohou následovat samotná pravidla pro transkripci. Příklad několika pravidel vytvořených v rámci této práce je následující:

```
sciare => šiare / _
sci => š / _<a,e,o,u>
sc => š / _<i,e>
cia => čia / -_-
ci => č / _E
c => k~/ _<a,o,u,Q,W,J>
c => k~/ _c<a,o,u,Q,W,J>
c => č / _'<i,e>
c => č / _<i,e>
c => č / _c<i,e>
ch => k~/ _<i,e>
```

Zde lze pozorovat využití zástupných symbolů Q pro znělé souhlásky, W pro neznělé souhlásky a J pro nepárové znělé souhlásky.

Na konec jsou definována obecná pravidla, která řeší spodoby znělosti. Tato pravidla existují proto, že v italštině existují některá slova, která mohou mít více přípustných výslovností. To se týká hlavně podoby znělosti na konci slov. Znak *A* v těchto pravidlech značí změnu znělosti, znak *O* zase značí ponechání znělosti. Příklad několika obecných pravidel je následující:

```
W => O~/ _-W
W => A~/ _-Q
W => A~/ _<Q,W>Q
```

```
W => O~/ _<Q,W>W
Q => A~/ _<Q,W>Q-
Q => O~/ _<Q,W>-Q
Q => A~/ _Q-
Q => O~/ _<Q,W>Q
Q => A~/ _<Q,W>W
Q => A~/ _W
```

Takto definovaná pravidla následně slouží k vygenerování několika přípustných výslovností.

Tabulka 3.4: Ukázka převodu všeobecné deklarace lidských práv do fonetické podoby

Původní text	Tutti gli esseri umani nascono liberi ed eguali in dignità e diritti. Essi sonodotati di ragione e di coscienza e devono agire gli uni verso gli altri in spirito difratellanza.
Fonetický přepis	tutti Li esseri umani naskono liberi ed egvali in diňita e diritti essi sono dotati di raČone e di košenca e devono aČire Li uni verso Li altri in spirito di fratellanca
Český překlad	Všichni lidé rodí se svobodní a sobě rovní co do důstojnosti a práv. Jsou nadáni rozumem a svědomím a mají spolu jednat v duchu bratrství.

3.1.3 Tvorba jazykového modelu

Po vytvoření korpusu a slovníku bylo možné vytvořit bi-gramový jazykový model. Ten byl natrénován ve spolupráci s vedoucím této bakalářské práce. Souhrnný popis statistik pro korpus, výslovnostní slovník a jazykový model lze nalézt v tabulce 3.5

Bi-gramový jazykový model je vytvořen tak, že se spočítají četnosti slov a dvojic slov v korpusu. Pro spočítání podmíněné pravděpodobnosti pro každý bi-gram je použit vztah 3.2:

$$P(w_i|w_{i-1}) = \frac{C(w_i, w_{i-1})}{C(w_i)} \tag{3.2}$$

kde C je četnost výskytu slova či dvojice slov v korpusu.

Pro neviděné sekvence slov, pro které by vycházela četnost 0, je použit Witten-Bellův algoritmus [19] pro vyhlazování.

Tabulka 3.5: Souhrnná statistika pro korpus, slovník a jazykový model

Velikost korpusu	2,65 GB
Celkový počet slov v korpusu	419499868
Počet unikátních slov v korpusu	1759440
Celkový počet slov ve slovníku	154513
Celkový počet výslovností ve slovníku	167383
Míra OOV	4,3 %
Celkový počet bigramů	410571486
Počet unikátních bigramů	37389046

3.1.4 Tvorba akustického modelu

Pro trénování akustického modelu byla využita řečová databáze Common Voice (další popis databáze v 3.2.2), která je validována v rámci projektu Common Voice, a nepotřebuje tak další validaci v rámci této práce. Textové přepisy pro akustická data v této databázi byla převedena do fonetické podoby pomocí vytvořeného modelu G2P. Zároveň byla upravena akustická data přidáním ticha na začátek a na konec každé z použitých nahrávek.

Pro akustické modelování byla používána architektura typu feed-forward sequential memory network [20]. Ta využívá tzv. paměťové bloky, které umožňují eliminovat neomezenou zpětnou vazbu přítomnou v klasické rekurentní neuronové síti a dávají přitom podobně dobré výsledky v mnoha různých úlohách. Vstup do FSMN neuronové sítě byl tvořen 39-dimenzionálními spektrálními koeficienty, které byly počítány ze segmentů (též framů) dlouhých 25 ms s překryvem o délce 12.5 ms. Konkrétní použitá FSMN síť měla celkem 11 skrytých vrstev a každá skrytá vrstva obsahovala vždy 512 neuronů s ReLU aktivační funkcí.

3.1.5 Experimenty a porovnání systémů

Po vytvoření všech dílčích modelů generativního systému ASR bylo možné provést první experimenty. Bylo provedeno srovnání zde popisovaného vytvořeného systému a vytvořeného systému E2E, jehož vývoj je popsán v další podkapitole, na řečové databázi Common Voice, která obsahovala 234 hodin validovaných dat. Tato databáze byla využita jak pro trénování, tak pro evaluaci obou systémů.

Jak je vidět v tabulce 3.6, tak systém E2E dostává lepšímu výsledku. Kvůli této skutečnosti bylo rozhodnuto, že další prováděný vývoj v rámci této práce měl pokračovat pouze pro tento systém. Přispívajícím faktorem k tomuto rozhodnutí byla náročnost vytvoření a odladění modulárního systému ASR. Jak dokazuje [17], tak vytvoření takového systému je úkol vhodný spíše pro tým lidí a optimalizace jednotlivých částí může trvat i mnoho měsíců nebo až několik let. Pro vytvoření modelu

G2P je totiž potřeba buď pokročilá znalost jazyka určeného k rozpoznávání, nebo mít alespoň vhodná trénovací data při využití některého z algoritmů strojového učení. Na druhou stranu systém E2E vyžaduje téměř nulovou znalost jazyka, model se učí generovat výstupní text přímo ze vstupních akustických dat a přiřazené transkripce.

Tabulka 3.6: Výsledky a porovnání systémů na databázi Common Voice

Architektura	ACC [%]
Modulární ASR	73,6
End-to-end ASR	81,1

3.2 End-to-end architektura

V této podkapitole je nejprve popsána struktura použitého E2E systému. Následuje popis veřejně dostupných řečových databází využitých k jeho trénování. Na závěr je popsán navržený postup tvorby vlastní řečové databáze, který byl i v rámci této práce využit.

3.2.1 Struktura použitého E2E systému

Pro trénování E2E systému byl využíván open-source toolkit ESPnet [21], který využívá kombinaci obou dvou klíčových konceptů popsaných v kapitole 1.3, tedy CTC mechanismu a enkodér-dekodér architektury s attention mechanismem. V rámci trénování bylo využíváno modelování na úrovni „podslovních“ tokenů. Pro jejich tvorbu z textových anotací vstupních textových nahrávek byl využíván nástroj Sentencepiece[4] [22]. Počet tokenů byl nastaven na hodnotu 5000. Vstupní řečové nahrávky byly parametrizovány po segmentech o délce o 25ms s překryvem o 10 ms. Pro každý segment bylo přitom spočítáno 80 mel-spektrálních koeficientů. Celý výsledný E2E model měl celkem cca 136 mil. parametrů pro učení a tedy velikost cca 500 MB.

3.2.2 Volně dostupné řečové databáze

Pro trénování modelu E2E byly shromážděny nalezené volně dostupné řečové databáze pro italštinu. Tyto databáze byly využity pro natrénování základního systémů ASR, který byl následně využit pro těžení dalších dat za účelem vytvoření vlastní řečové databáze.

[4] https://github.com/google/sentencepiece

Databáze Common Voice[5] je stále ve vývoji v rámci projektu Common Voice společnosti Mozzila. Projekt obsahuje databáze pro většinu světových jazyků. Na webové stránce projektu může kdokoliv přispět vlastním hlasem a rozšířit tak libovolnou databázi podle zvoleného jazyka. Nahraná řečová data jsou následně validována kýmkoliv, validace tak funguje na stejném principu jako nahrávání řeči.

Řečový korpus Multilingual TEDx [23] obsahuje datové sady pro 8 jazyků (španělština, francouzština, portugalština, italština, ruština, řečtina, arabština, němčina). Jedná se o záznamy z přednášek TEDx talks[6], které jsou zaměřené na různá témata, a manuální transkripce. Mimo dat určených pro trénování systémů ASR obsahuje i datové sady určené pro speech translation (ST).

Multilingual LibriSpeech (MLS) [24] je řečový korpus obsahující datasety v různých jazycích, který byly odvozeny z audioknih dostupných v rámci projektu LibriVox[7]. MLS obsahuje datasety pro 8 jazyků (angličtina, němčina, nizozemština, španělština, francouzština, italština, portugalština, polština). LibriVox je projekt, v rámci kterého jsou shromažďovány audioknihy v různých jazycích, které byly namluveny dobrovolníky.

VoxForge[8] obsahuje nahrávky od dobrovolníků v několika různých jazycích. Většina dostupných nahrávek je špatné kvality. Příčinou je pravděpodobně komprese či nekvalitní mikrofon použitý při nahrávání. Tento fakt ale může být přínosný, takováto data totiž pomáhají zvýšit robustnost modelu.

Zpracování volně dostupných řečových databází

Všechny nahrávky v dané řečové databázi, které byly příliš dlouhé, byly pomocí časových značek u přiřazených titulků rozřezány spolu se samotnými titulky na kratší úseky o délce trvání kratší než 25 sekund.

Kratší úseky bylo snadné profiltrovat a odstranit úseky, které obsahovaly neřečové elementy jako je hudba. K tomu bylo využito přiřazených transkripcí, které označovaly například úsek s hudbou jako „[musica]“ či „[music]“. Všechny transkripce obsahující hranaté či kulaté závorky byly tedy odstraněny spolu s odpovídající nahrávkou.

Přiřazená textová data bylo nutné normalizovat. U všech textů bylo sjednoceno kódování a byly sjednoceny různé znaky jako jsou pomlčky, apostrofy atp. Zároveň byly z textů odstraněny znaky neřečového charakteru. Dále ještě byla čísla zapsaná číslicemi přepsána na slovní podobu a různé zkratky a symboly byly nahrazeny celými slovy.

[5] https://commonvoice.mozilla.org/
[6] https://www.ted.com/watch/tedx-talks
[7] https://librivox.org/
[8] http://www.voxforge.org/it

Nahrávky z databází Multilingual TEDx, MLS a VoxForge byly rozpoznány rozpoznávačem pro italštinu na platformě Microsoft Azure, která nabízí přibližně 200 hodin rozpoznávání zdarma. Výstup byl následně porovnán s původními transkripcemi a na základě podobnosti dle metriky character error rate (CER) byly shodné nahrávky použity pro trénování. CER je definováno jako:

$$CER = (S\ + D + I)/N = (S\ + D + I)/(S\ + D + C) \tag{3.3}$$

Kde S je počet substituovaných znaků, D je počet smazaných znaků, I je počet vložených znaků, C je počet shodných znaků, N je počet znaků v předloze ($N = S + D + C$).

Tabulka 3.7 znázorňuje výsledky těžení jednotlivých databází.

Tabulka 3.7: Přehled volně dostupných řečových databází

Databáze	Celkem hodin	Použito hodin
Common Voice	234	234
Multilingual LibriSpeech	279	144
VoxForge	20	13
Multilingual TEDx	107	39

3.2.3 Tvorba vlastní databáze pro E2E trénování

Vzhledem k účelu vytvářeného systému ASR bylo potřeba doplnit trénovcí data o záznamy řeči zpravodajského charakteru. Volně dostupné databáze popsané výše obsahují spíše záznamy řeči, kde v jednu chvíli hovoří pouze jeden člověk a jsou bez příliš velkého ruchu na pozadí. Pro tvorbu vlastní trénovací databáze byly využity videa z platformy YouTube (YT)[9]. Jedná se hlavně o zpravodajské kanály. Tabulka 3.8 nabízí ucelený přehled jednotlivých zdrojů využitých pro tvorbu vlastní databáze.

La Republicca[10] je zpravodajský kanál na YT. Obsah kanálu tvoří videa pokrývající různé dění ve světě.

Marco Montemagno[11] je YT kanál Marca Montemagna, jenž je úspěšným podnikatelem. Jeho videa jsou směsí rozhovorů s různými celebritami či jinými vlivnými osobnostmi a různých videí, jenž pokrývají hlavně technologická témata.

Metro News Italia[12] je zpravodajský kanál na YT, který pokrývá různé lokální dění.

[9] https://www.youtube.com/
[10] https://www.youtube.com/c/repubblica/
[11] https://www.youtube.com/user/montymonty/
[12] https://www.youtube.com/c/MetroNewsItalia/

Palazzo Chigi[13] je oficiální kanál vlády Itálie na YT. Obsahem kanálu jsou záznamy z poslanecké sněmovny Itálie, projevy, záznamy z konferencí apod.

Presidenza della Repubblica Italiana Quirinale[14] je oficiální YT kanál tiskového a komunikačního úřadu předsednictví Italské republiky.

Tabulka 3.8: Přehled zdrojů použitých pro tvorbu vlastní řečové databáze

Zdroj	Popis zdroje	Hodin Staženo
La Republicca	YT zpravodajství	2437
Marco Montemagno	Rozhovory, technologie	305
Metro News Italia	Různé lokální zprávy	20
Palazzo Chigi	Oficiální kanál vlády Itálie	796
Presidenza della Repubblica Italiana Quirinale	Kanál tiskového a komunikačního úřadu předsednictví Italské republiky	128

Stahování dat z platformy YouTube

Platforma YouTube je významným zdrojem akustických dat s textovými přepisy v podobě titulků. Díky své rozšířenosti ve světě totiž obsahuje obrovské množství videí, které mnohdy obsahují ručně přepisy. Tato skutečnost tedy umožňuje využití jediného nástroje pro stažení velkého množství dat. Pro tyto účely byl tak vytvořen program v programovacím jazyce Python, který umožňuje stažení celého obsahu zadaného kanálu. Knihovna pytube[15] umožňuje jednoduché stažení a zpracování YT videí s titulky, proto byla využita pro vytvoření programu pro stahování videí z YT. Proces stahování videí z daného kanálu je znázorněn na obrázku 3.2.

Metoda pro stažení zvukového záznamu a titulků v italštině z videa vypadá následovně:

```
def download_video(video, output_path):
    stream = video.streams.get_audio_only()
    if 'it' in video.captions.keys() and stream is not None:
        stream.download(output_path=output_path)
        caption = video.captions['it']
        caption.download(title=video.title,
            srt=True,
            output_path=output_path)
```

[13] https://www.youtube.com/c/palazzochigi/
[14] https://www.youtube.com/user/presidenzarepubblica
[15] https://pytube.io/

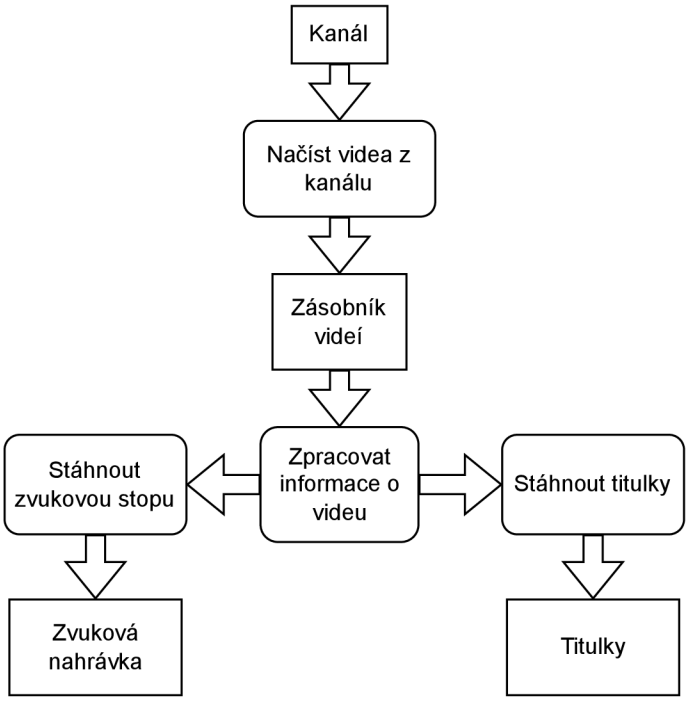
```
    else:
        print('failed')
```

Tato funkce pomocí knihovny pytube získá audio ze zadaného videa pouze audio, dále zkontroluje, zda video obsahuje titulky v italštině. Pokud video titulky v italštině obsahuje, tak se audio z videa spolu s italskými titulky stáhne a uloží do zvolené cesty.

Pomocí funkce knihovny pytube je v hlavním těle programu získán seznam jednotlivých videí zadaného kanálu, která jsou následně postupně předávána metodě download_video. Samotný kód vypadá následovně:

```
from pytube import Channel
from pytube import YouTube

url = 'https://www.youtube.com/c/repubblica'
channel_name = 'La Repubblica/'
c = Channel(url)
for i, yt in enumerate(c.videos):
    print('downloading', i+1, 'of', len(c.videos))
    download_video(yt, channel_name)
```



Obrázek 3.2: Proces stahování dat z kanálu na YouTube

Zpracování a těžení dat

Stažená data bylo nutné zpracovat podobně jako v případě zpracování volně dostupných řečových databází. Všechna textová data byla normalizována, to znamená že se sjednotilo kódování a různé znaky jako je apostrof či uvozovky, zkratky, různé symboly, a čísla byla nahrazeny za celá slova.

Po natrénování modelu E2E pomocí volně dostupných řečových databází ho bylo možné využít pro těžení dat za účelem vytvoření vlastní databáze. Vlastní natrénovaný systém pro rozpoznávání italštiny byl využit k přepisu stažených dat pro získání samotné textové transkripce a časových značek jednotlivých slov. Pro účely vývoje v rámci této práci byl využit systém ASR pro přepis češtiny zprovozněný na TUL, který na rozdíl od systému vyvinutého v rámce této práce anotuje i hluky a jiné neřečové elementy.

Po transkripci audio nahrávky vlastním modelem byl výstup z něj spárován s původními titulky za využití Levensteinovy vzdálenosti [25]. Zároveň byl výstup z vlastního modelu spárován podle časových značek s výstupem z modelu pro češtinu. A nakonec úseky, které byly shodně spárovány s původními titulky a ležely mezi anotovanými hluky, byly vystřiženy spolu s odpovídajícím textovým přepisem.

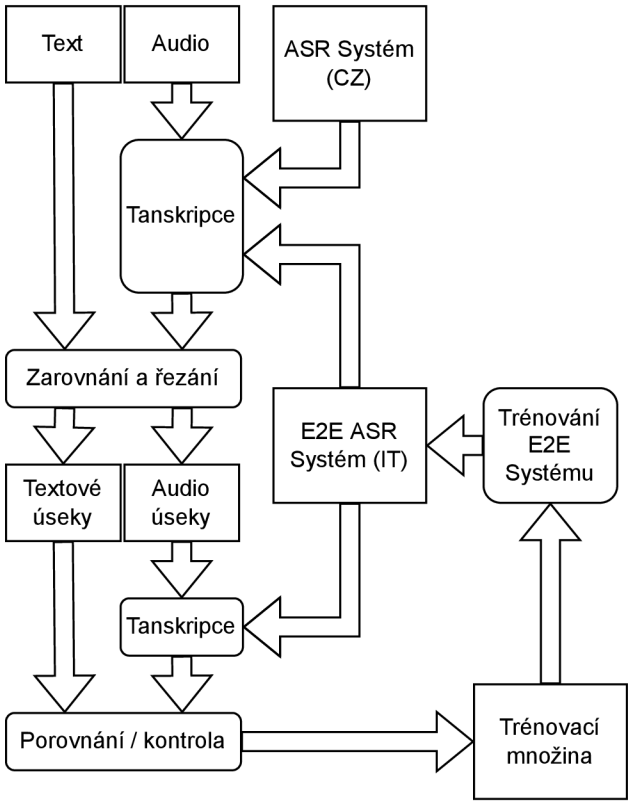
Vzhledem k tomu, že časové značky, podle kterých bylo stříháno, nemusí být naprosto přesné, byly pro zvýšení přesnosti přidány další tři varianty textu, kdy se k porovnávanému úseku titulků přidá:

- jedno slovo z předcházejícího úseku titulků na začátek;
- jedno slovo z nadcházejícího úseku titulků na konec;
- jedno slovo z předcházejícího úseku titulků na začátek a jedno slovo z nadcházejícího úseku titulků na konec.

Aby se ověřilo, že byly úseky správně vystřiženy, byly ještě jednou přepsány systémem pro přepis italštiny, jehož výstup byl porovnán se čtyřmi variantami přepisu z předchozího kroku. Zároveň bylo u nahrávek kontrolováno, zda odpovídá prvních pět a posledních pět znaků s titulky. Při výběru byla umožněna malá míra neshody pro zvýšení výtěžnosti. Otestováno bylo několik prahových hodnot shody, při čemž se ukázalo, že shoda 95 % a více s kontrolou krajních pěti znaků má podobnou výtěžnost jako shoda 98 % a více bez kontroly krajních znaků. Vybíráno bylo podle první zmíněné varianty. Tímto způsobem je zajištěná dostatečná kontrola a přesnost nově vytvořené množiny trénovacích dat.

Všechny ověřené úseky bylo nakonec možné přidat do trénovací množiny dat a byly tím způsobem využity pro další trénování systému. Výše popsaný postup lze iterativně opakovat a dále tak rozšiřovat trénovací množinu dat, čímž se postupně

zvyšuje přesnost a robustnost systému. Obrázek 3.3 celý popsaný postup znázorňuje. Tabulka 3.9 obsahuje výsledky těžení pro jednotlivé zdroje dat.



Obrázek 3.3: Proces iterativního těžení akustických dat

Tabulka 3.9: Výsledky těžení s porovnáním s výstupem ze systému ASR pro češtinu

		Shoda		
Zdroj	**Hodin vyřezáno**	**98 %**	**Okraje +95 %**	**Výtěžnost [%]**
La Republicca	348,2	17,3	17,1	4,9
Marco Montemagno	68,7	1	1,2	1,7
Metro News Italia	8,8	0,4	0,5	5,7
Palazzo Chigi	261,3	24,7	24,7	9,5
Presidenza della Repubblica Italiana Quirinale	74,4	7,6	8,2	11,0
	761,4	**51**	**51,7**	**6,7**

Vzhledem k nízké výtěžnosti dat výše popsaným postupem byl ještě vyzkoušen postup, který pro hledání shody nevyužívá transkripci z ASR systému pro češtinu, ale pouze porovnává výstup z vlastního systému pro přepis italštiny s původními titulky. Vyřezány jsou úseky, které se shodují alespoň ve třech po sobě jdoucích

slovech. Výsledky těžení tímto postupem jsou znázorněny v tabulce 3.10. Všechny ostatní kroky jsou identické postupu, který je popsaný výše a znázorněný na 3.3.

Po porovnání výsledků těžení shromážděných dat, které jsou znázorněny v tabulkách 3.9 a 3.10, je jasné, že druhý vyzkoušený postup nabízí mnohem větší výtěžnost záznamů řeči s textovými přepisy, při čemž je díky použitým postupům zajištěna stejná přesnost dat. Proto byl druhý uvedený postup využit pro tvorbu vlastní řečové databáze. Porovnání obou postupů je znázorněno v tabulce 3.11.

Tabulka 3.10: Výsledky těžení bez porovnání s výstupem ze systému ASR pro češtinu

		Shoda		
Zdroj	**Hodin vyřezáno**	**98 %**	**Okraje +95 %**	**Výtěžnost [%]**
La Republicca	577,9	192,4	180,9	31,3
Marco Montemagno	171,5	59,6	55,2	32,2
Metro News Italia	7	2,3	2	28,6
Palazzo Chigi	262,5	85,2	69,2	26,4
Presidenza della Repubblica Italiana Quirinale	66,4	19,9	16,8	25,3
	1085,3	**359,4**	**324,1**	**29,8**

Tabulka 3.11: Porovnání vyzkoušených postupů pro těžení dat

	Celk. hod. vyřezáno	**Shodných hodin**	**Výtěžnost [%]**
Shoda s hluky	768,4	51,9	6,8
Shoda bez hluků	1085,3	324,1	29,8

4 Dosažené výsledky a provedené experimenty

V této kapitole jsou popsány provedené experimenty a jsou shrnuty dosažené výsledky vyhodnocení vytvořených modelů pro přepis řeči v italštině. Tabulka 4.2 znázorňuje výsledky vyhodnocení vytvořeného modulární systému a postupný vývoj systému E2E. Pro přehlednost jsou jednotlivé datasety označeny zkratkami, které jsou následující:

- CV - Common Voice;
- VF - VoxForge;
- MLS - Multilingual LibriSpeech;
- TED - Multilingual TEDx;
- VD - vlastní trénovací databáze;
- BC - zpravodajské pořady z různých kanálů, ručně anotované firmou Newton Media.

Pro vyhodnocení modelu je použita metrika ACC [26]. Číslo u modelu znázorňuje vnitřní velikost modelu (E2E 256 - 256 neuronů v každé skryté vrstvě). Přehled počtu hodin testovacích dat vyčleněných z vybraných databází je znázorněn v tabulce 4.1.

Tabulka 4.1: Přehled počtu hodin v testovacích datasetech

Dataset	**CV**	**VF**	**MLS**	**BC**	**Celkem**
Test. hod.	1,68	1,33	5,20	1,15	9,36
Počet slov	9775	8310	40854	4074	63013

V rámci této práce byl nejprve vytvořen modulární systém pro automatický přepis řeči v italštině. Tvorba takového systému zahrnovala vytvoření textového korpusu a výslovnostního slovníku, a následně natrénování jazykového modelu a akustického modelu, pro jehož natrénování byla využita řečová databáze Common

Voice. Po vytvoření a sjednocení jednotlivých modulů bylo možné vyhodnotit systém na základě přesnosti přepisu italštiny.

Pro porovnání možných přístupů byl zároveň natrénován E2E model pro přepis řeči v italštině. Aby nedošlo ke zkreslení výsledků, tak byla ke trénování tohoto modelu též využita pouze řečová databáze Common Voice, na které byly následně oba systémy vyhodnoceny. Na základě porovnání přesnosti přepisu obou systémů bylo rozhodnuto, že by se další vývoj systému ASR v rámci této práce měl zabývat pouze E2E přístupem. K tomuto rozhodnutí přispěla i vysoká náročnost vytvoření kvalitního modulárního systému. Odladění jednotlivých modulů totiž může trvat měsíce a je vhodné spíše pro větší tým lidí. Výsledky porovnání těchto základních modelů jsou znázorněny na prvních dvou řádcích tabulky 4.2.

Pro další trénování E2E modelu byly využity volně dostupné řečové databáze (MLS, Multilingual TEDx, Voxforge), jejichž data bylo ale nutné zkontrolovat na to, zda obsahují přesné přepisy, a vytěžit je. Ke kontrole dat byla využita platforma Azure, která nabízí službu automatického přepisu. Důvod ke kontrole těchto databází byl takový, že přepisy obsažené v nich obsažené nebyly vždy přesné. Po přepisu pomocí platformy Azure byli postupně jednotlivé databáze vytěženy. Vytěžená data byla přidávána do trénovací množiny dat, a tím způsobem byla využita ke trénování nových E2E modelů.

Po natrénování základní verze systému E2E za použití všech zmíněných volně dostupných řečových databází bylo možné systém využít pro tvorbu vlastní řečové databáze. Proces tvorby vlastní databáze spočíval v těžení záznamů řeči s textovými přepisy, které byly staženy z různých zdrojů. Data byla rozpoznána vlastním systémem a následně se hledala shoda mezi výstupem ze systému a původními přepisy.

Tabulka 4.2: Vyhodnocení vytvořených modelů na testovacích sadách

			ACC [%]				
Model	**Dataset**	**Trén. hodin**	**CV**	**VF**	**MLS**	**BC**	**Prům.**
FSMN	CV	234	73,6	74,3	65,3	66,5	67,9
E2E 256	CV	234	81,1	79,9	66,8	68,7	70,9
E2E 256	CV, VF, MLS	391	81,4	93,7	82,1	79,3	83,3
E2E 256	CV, VF, MLS, TED	430	84,2	94,9	84,9	80,1	85,8
E2E 512	CV, VF, MLS, TED	430	85,5	95,8	86,9	81,9	87,5
E2E 512	CV, VF, MLS, TED, VD	651	85,9	95,9	87,3	87,3	**88,2**

4.1 Porovnání přesnosti přepisu vytvořeného systému s komerčním systémem

Pro objektivní pohled na výsledky přepisu vytvořeného E2E systému byla navíc některá testovací data použita pro porovnání přesnosti přepisu vlastního systému s přepisem systému, který funguje v rámci cloudové platformy MS Azure. Toto srovnání znázorňuje tabulka 4.3.

Výsledky z vlastního systému jsou celkově porovnatelné s výsledky z platformy Azure. Ta je lepší na datasetu BC o 3,6 % procent a naopak horší na datasetech CV,VF a MLS absolutně o 4,3 % až 7 %. Nutno ale podotknout, že systém vytvořený v rámci této práce byl trénován na trénovacích podmnožinách těchto databází, kdežto u systému Azure tyto databáze použity být nemusely, což může vést k horším výsledkům na odpovídajících testovacích podmnožinách.

Tabulka 4.3: Porovnání vlastního systému s MS Azure

	ACC [%]				
Systém	**CV**	**VF**	**MLS**	**BC**	**Prům.**
Azure	86,2	88,9	81,2	90,9	85,4
E2E	85,9	95,9	87,3	87,3	89,7

5 Závěr

Podle zadání práce byla provedena analýza současného stavu v oblasti automatického rozpoznávání řeči pro italštinu a dostupných zdrojů umožňujících tvorbu jazykového a akustického modelu. Cílem práce byla tvorba různých modelů pro automatický přepis italštiny na základě různých, i vlastnoručně vytěžených dat a následné vyhodnocení vytvořených modelů. Ty budou moci být na základě této práce dále vylepšovány až do podoby nasaditelné v rámci platformy pro nepřetržitý monitoring televizního a rádiového vysílání.

V rámci práce byly vytvořeny dva systémy s různou architekturou, a to konkrétně modulární systém a end-to-end model. Větší část prací však byla věnována sběru dat pro end-to-end model, což bylo rozhodnuto na základě porovnání náročnosti vytvoření a míry přesnosti přepisu italštiny základních verzí obou systémů. Tvorba modulárního systému s sebou však nesla nutnost studia italštiny, a to především z pohledu fonetiky. Poznatky z této části práce jsou tak zachyceny v kapitole 2.

V rámci tvorby systému ASR s modulární architekturou se podařilo vytvořit všechny dílčí části nutné pro zprovoznění celého systému, konkrétně tedy:

- výslovnostní model a textový korpus, které byly využity k vytvoření výslovnostního slovníku se 150 tis. slovy a 170 tis. výslovnostmi;
- bigramový jazykový model, který obsahuje celkem 410 mil. bigramů;
- akustický model, pro jehož trénování byla využita řečová databáze Common Voice, s architekturou tvořenou neuronovu sítí typu FSMN.

Vytvořený systém dosáhl na testovací sadě vyčleněné z dat databáze Common Voice přesnosti 73,6 %, zatímco end-to-end model natrénovaný na stejné databázi dosáhl přesnosti 81,1 %.

Pro další trénování end-to-end systému byly shromážděny a vhodně upraveny volně dostupné databáze: VoxForge, Multilingual LibriSpeech, Multilingual TEDx. Posléze byl natrénovaný systém využit ke tvorbě vlastní řečové databáze těžením různých akustických dat s textovými přepisy. Samotný postup těžení lze shrnout následovně:

- přepis vlastním systémem ASR;
- nalezení shody mezi výstupem z vlastního systému a původním přepisem nahrávky;
- vyříznutí shodných úseků z nahrávky a přepisu;
- opětovný přepis vlastním systémem pro kontrolu správnosti vyříznutí;
- přidání shodných úseků do trénovací množiny.

Navržený postup je možné využít i v budoucnu pro další rozšíření vytvořených ASR modelů.

Výsledný vytvořený end-to-end model, jehož trénování prováděl vedoucí práce, dosahuje na množině televizních a rozhlasových nahrávek přesnosti přepisu 87,3 %. Tento výsledek byl porovnán s přesností komerčního systému pro přepis řeči na cloudové platformě MS Azure. Ze srovnání těchto systému lze vyvodit, že model vytvořený v rámci této práce je co do přesnosti na podobné úrovni jako komerční model rozpoznávání řeči, a je ho tak možné do budoucna dále rozvíjet nebo využít.

Literatura

1. HINTON, Geoffrey; DENG, Li; YU, Dong; DAHL, George E.; MOHAMED, Abdel-rahman; JAITLY, Navdeep; SENIOR, Andrew; VANHOUCKE, Vincent; NGUYEN, Patrick; SAINATH, Tara N.; KINGSBURY, Brian. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*. 2012, roč. 29, č. 6, s. 82–97. Dostupné z DOI: 10.1109/MSP.2012.2205597.
2. KILGOUR, Kevin; HECK, Michael; MÜLLER, Markus; SPERBER, Matthias; STÜKER, Sebastian; WAIBEL, Alex. The 2014 KIT IWSLT speech-to-text systems for English, German and Italian. In: *Proceedings of the 11th International Workshop on Spoken Language Translation: Evaluation Campaign*. Lake Tahoe, California, 2014, s. 73–79. Dostupné také z: https://aclanthology.org/2014.iwslt-evaluation.9.
3. WEIBIN, Chen. *Phoenix: Deep Speech Based Automatic Speech Recognition System for Italian Language*. Milano, 2019. Dostupné také z: https://www.politesi.polimi.it/bitstream/10589/152319/1/Tesi.pdf.
4. HANNUN, Awni Y.; CASE, Carl; CASPER, Jared; CATANZARO, Bryan; DIAMOS, Greg; ELSEN, Erich; PRENGER, Ryan; SATHEESH, Sanjeev; SENGUPTA, Shubho; COATES, Adam; NG, Andrew Y. Deep Speech: Scaling up end-to-end speech recognition. *CoRR*. 2014, roč. abs/1412.5567. Dostupné z arXiv: 1412.5567.
5. WANG, Dong; WANG, Xiaodong; LV, Shaohe. An Overview of End-to-End Automatic Speech Recognition. *Symmetry*. 2019, roč. 11, č. 8. ISSN 2073-8994. Dostupné z DOI: 10.3390/sym11081018.
6. ČERVA, Petr. *Adaptation of speech recognition systems to selected real-world deployment conditions*. Liberec, 2021.
7. WANG, Song; LI, Guanyu. Overview of end-to-end speech recognition. *Journal of Physics: Conference Series*. 2019, roč. 1187, č. 5, s. 052068. Dostupné z DOI: 10.1088/1742-6596/1187/5/052068.
8. GRAVES, Alex; FERNÁNDEZ, Santiago; GOMEZ, Faustino; SCHMIDHUBER, Jürgen. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In: *Proceedings of the*

23rd International Conference on Machine Learning. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 2006, s. 369–376. ICML '06. ISBN 1595933832. Dostupné z DOI: 10.1145/1143844.1143891.

9. GRAVES, Alex; JAITLY, Navdeep. Towards end-to-end speech recognition with recurrent neural networks. In: *International conference on machine learning*. 2014, s. 1764–1772.
10. CHO, Kyunghyun; VAN MERRIËNBOER, Bart; GULCEHRE, Caglar; BAHDANAU, Dzmitry; BOUGARES, Fethi; SCHWENK, Holger; BENGIO, Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. 2014.
11. *Italian Speaking Countries 2022*. [B.r.]. Dostupné také z: https://worldpopulationreview.com/country-rankings/italian-speaking-countries.
12. BERTINETTO, Pier Marco; LOPORCARO, Michele. The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the International Phonetic Association*. 2005, roč. 35, č. 2, s. 131–151. ISSN 0025-1003. Dostupné z DOI: 10.1017/S0025100305002148.
13. VIETTI, Alessandro. Phonological Variation and Change in Italian. In: 2019. Dostupné z DOI: 10.1093/acrefore/9780199384655.013.494.
14. CAVANA, Alessandra Chiodelli Mc. *Italsky za 4 týdny: intenzivní kurz pro začátečníky a mírně pokročilé : A1-A2*. Dubicko: INFOA, [2017]. ISBN 978-80-7547-187-1.
15. *REGIONAL DIALECTS: HOW TUSCAN DIFFERS FROM STANDARD ITALIAN*. Giulio Simeone, 2021. Dostupné také z: https://emeraldnetwork.altervista.org/tuscan-dialect-main-differences-with-italian-language/.
16. *PIEMONTEIS ON THE INTERNET*. Vittorio Bertola, 1995. Dostupné také z: https://bertola.eu/piemonteis/.
17. ŠAFAŘÍK, Radek. *Multilingvální systémy rozpoznávání řeči a jejich efektivní učení*. Technická univerzita v Liberci, Studentská 1402/2, 461 17 Liberec 1, 2020.
18. NOUZA, J.; PSUTKA, J.; UHLIR, J. Phonetic Alphabet for Speech Recognition of Czech. *Radioengineering*. 1997, roč. 6, č. 4, s. 16–20. ISSN 1210-2512. Dostupné také z: http://hdl.handle.net/11012/58374.
19. WITTEN, Ian; BELL, Timothy. The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*. 1991, roč. 37, s. 1085–1094. Dostupné z DOI: 10.1109/18.87000.
20. ZHANG, Shiliang; LIU, Cong; JIANG, Hui; WEI, Si; DAI, Lirong; HU, Yu. *Feedforward Sequential Memory Networks: A New Structure to Learn Long-term Dependency*. arXiv, 2015. Dostupné z DOI: 10.48550/ARXIV.1512.08301.

21. WATANABE, Shinji; HORI, Takaaki; KARITA, Shigeki; HAYASHI, Tomoki; NISHITOBA, Jiro; UNNO, Yuya; SOPLIN, Nelson Enrique Yalta; HEYMANN, Jahn; WIESNER, Matthew; CHEN, Nanxin; RENDUCHINTALA, Adithya; OCHIAI, Tsubasa. ESPnet: End-to-End Speech Processing Toolkit. *CoRR*. 2018, roč. abs/1804.00015. Dostupné z arXiv: 1804.00015.

22. KUDO, Taku; RICHARDSON, John. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, 2018, s. 66–71. Dostupné z DOI: 10.18653/v1/D18-2012.

23. SALESKY, Elizabeth; WIESNER, Matthew; BREMERMAN, Jacob; CATTONI, Roldano; NEGRI, Matteo; TURCHI, Marco; OARD, Douglas W.; POST, Matt. Multilingual TEDx Corpus for Speech Recognition and Translation. In: *Proceedings of Interspeech*. 2021.

24. PRATAP, Vineel; XU, Qiantong; SRIRAM, Anuroop; SYNNAEVE, Gabriel; COLLOBERT, Ronan. MLS: A Large-Scale Multilingual Dataset for Speech Research. In: *Interspeech 2020*. ISCA, 2020. Dostupné z DOI: 10.21437/interspeech.2020-2826.

25. YUJIAN, Li; BO, Liu. A Normalized Levenshtein Distance Metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2007, roč. 29, č. 6, s. 1091–1095. Dostupné z DOI: 10.1109/TPAMI.2007.1078.

26. GAD, Ahmed Fawzy. *Evaluating Deep Learning Models: The Confusion Matrix, Accuracy, Precision, and Recall*. [B.r.]. Dostupné také z: https://blog.paperspace.com/deep-learning-metrics-precision-recall-accuracy/.

A Přílohy

A.1

Definovaná pravidla pro přepis z ortografické podoby na fonetickou pro italštinu

```
Q:b,d,g,v,z
W:f,k,p,s,t,q
J:j,l,m,n,r
E:a,e,i,o,u
A:sh-ž,b-p,d-t,f-v,g-k,k-g,p-b,s-z,t-d,z-s,v-f,q-g
sciare => šiare / _
sci => š / _<a,e,o,u>
sc => š / _<i,e>
cia => čia / -_-
ci => č / _E
c => k / _<a,o,u,Q,W,J>
c => k / _c<a,o,u,Q,W,J>
c => č / _'<i,e>
c => č / _<i,e>
c => č / _c<i,e>
ch => k / _<i,e>
c => k / _ch<i,e>
ch => č / _<a,o,u>
c => č / _ch<a,o,u>
gi => Č / _E
g => Č / _<i,e>
g => / _g<i,e>
gh => g / _<i,e>
gn => ň / _
gli => Li / _-
gli => L / _
gl => L / _
qu => ku / _-
```

```
qu => kv / _
q => k / _
u => u / -g_E
u => v / g_<a,e,i>
stasera => stasera / _
s => s / -_
s => z / E_E
lizzazio => liCacio / _
mezz => meC / _
azzur => aCur / _
zz => c / _
z => c / _
y = j / E_
x => ks / _
w => v / _
via => via / -_-
i => i / -_
i => i / _j
i => j / <b, f, p, ch>_o
i => i / <b, f, p, ch>_E-
i => j / <b, f, p, ch>_E
i => i / i_
i => i / u_
i => j / E_
i => j / gh_E
h => / _
’ = _ / _
W => O / _<v, Wv, -v, -Wv, W-v>
W => O / _<Q,W>-<Q,W>W
W => A / _<Q,W>-<Q,W>Q
W => O / _<Q,W>-W
W => A / _<Q,W>-Q
W => O / _-<Q,W>W
W => A / _-<Q,W>Q
W => O / _-W
W => A / _-Q
W => A / _<Q,W>Q
W => O / _<Q,W>W
W => O / _Q-
W => A / _Q
Q => A / _<Q,W>Wv
Q => A / _Wv
Q => O / _-<Q,W>Q
Q => A / _-<Q,W>W
Q => O / _-Q
```

```
Q => A / _-W
Q => O / _-<J,E>
Q => O / _-<Q,W>Q
Q => O / _<Q,W>Q-<Q,W>Q
Q => O / _<Q,W>-<Q,W>Q
Q => O / _<Q,W>Q-Q
Q => O / _<Q,W>-<Q,W>Q
Q => A / _<Q,W>Q-
Q => O / _<Q,W>-Q
Q => A / _Q-
Q => O / _<Q,W>Q
Q => A / _<Q,W>W
Q => A / _W
Q => A / _-
```