VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÉ HODNOCENÍ ANGLICKÉ VÝSLOVNOSTI NERODILÝCH MLUVČÍCH

AUTOMATIC PRONUNCIATION EVALUATION OF NON-NATIVE ENGLISH SPEAKERS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE **PETER GAZDÍK**
AUTHOR

VEDOUCÍ PRÁCE **Ing. KATEŘINA ŽMOLÍKOVÁ**
SUPERVISOR

BRNO 2019

Vysoké učení technické v Brně
Fakulta informačních technologií

Ústav počítačové grafiky a multimédií (UPGM) Akademický rok 2018/2019

Zadání diplomové práce

Student: **Gazdík Peter, Bc.**
Program: Informační technologie Obor: Počítačová grafika a multimédia
Název: **Automatické hodnocení anglické výslovnosti nerodilých mluvčích**
Automatic Pronunciation Evaluation of Non-Native English Speakers
Kategorie: Zpracování řeči a přirozeného jazyka

Zadání:

1. Seznamte se s problémem detekce špatné výslovnosti z řečových nahrávek založené na rozpoznávání řeči.
2. Seznamte se s potřebnou teorií automatického rozpoznávání řeči, neuronových sítí a dostupnými toolkity (např. Kaldi, PyTorch / Keras).
3. Navrhněte a implementujte základní systém pro daný úkol podle nastudované literatury.
4. Otestujte systém na vhodném datasetu a porovnejte s publikovanými výsledky.
5. Navrhněte a implementujte možné zlepšení systému (např. využití nahrávek řečníků v mateřském jazyce nebo použití silnější architektury neuronové sítě).

Literatura:

- dle doporučení vedoucího

Při obhajobě semestrální části projektu je požadováno:

- Body 1 až 3.

Podrobné závazné pokyny pro vypracování práce viz http://www.fit.vutbr.cz/info/szz/

Vedoucí práce: **Žmolíková Kateřina, Ing.**
Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.
Datum zadání: 1. listopadu 2018
Datum odevzdání: 22. května 2019
Datum schválení: 1. listopadu 2018

Abstrakt

Výuka anglickej výslovnosti s využitím počítača sa v súčasnej dobe stáva čoraz viac populárnejšou. Napriek tomu presnosť týchto systémov je stále pomerne nízka. Táto diplomová práca sa preto zameriava na zlepšenie existujúcich metód automatického hodnotenia výslovnosti. V prvej časti práce je uvedený prehľad v súčasnosti používaných techník v tejto oblasti. Následne bol navrhnutý systém využívajúci dva rôzne prístupy. Dosiahnuté výsledky ukazujú znateľné zlepšenie oproti referenčnému systému.

Abstract

Computer-Assisted Pronunciation Training (CAPT) is becoming more and more popular these days. However, the accuracy of existing CAPT systems is still quite low. Therefore, this diploma thesis focuses on improving existing methods for automatic pronunciation evaluation on the segmental level. The first part describes common techniques for this task. Afterwards, we proposed the system based on two approaches. Finally, performed experiments show significant improvement over the reference system.

Kľúčové slová

automatické hodnotenie výslovnosti, výuka výslovnosti s využitím počítača, automatické rozpoznávanie reči, hlboké neurónové siete, rekurentné neurónové siete

Keywords

automatic pronunciation evaluation, computer-aided pronunciation training, automatic speech recognition, deep neural networks, recurrent neural networks

Citácia

GAZDÍK, Peter. *Automatické hodnocení anglické výslovnosti nerodilých mluvčích.* Brno, 2019. Diplomová práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Ing. Kateřina Žmolíková

Automatické hodnocení anglické výslovnosti nerodilých mluvčích

Prehlásenie

Prehlasujem, že som túto diplomovú prácu vypracoval samostatne pod vedením Ing. Kateřiny Žmolíkové a uviedol som všetky literárne pramene a publikácie, z ktorých som čerpal.

.....
Peter Gazdík
31. júla 2019

Poďakovanie

Rád by som poďakoval mojej vedúcej práce Katke Žmolíkovej za cenné pripomienky a množstvo času, ktorý mi venovala pri konzultáciách.

Obsah

Kapitola 1

Úvod

Vplyvom globalizácie sa dnes učí cudzí jazyk výrazne viac ľudí ako kedykoľvek predtým. Dôležitou, ale často podceňovanou, súčasťou cudzieho jazyka je výslovnosť, ktorá je mnohokrát kľúčová pre správne dorozumenie. Efektívna výuka výslovnosti sa ale väčšinou nezaobíde bez individuálneho prístupu, čo si však väčšina študentov zväčša nemôže dovoliť. Z tohto dôvodu sa automatické hodnotenie výslovnosti javí ako vhodná alternatíva.

V tejto práci sa budeme zaoberať automatickým hodnotením segmentálnych chýb, ktoré spočívajú vo vkladaní, vypúšťaní alebo zámene foném. K tomuto účelu sa dnes výhradne využívajú systémy založené na rozpoznávaní reči, preto ani táto práca nebude výnimkou. Problém hodnotenia výslovnosti môžeme vnímať na dvoch úrovniach. Prvá z nich spočíva v detekcii nesprávnej výslovnosti, zatiaľ čo druhá sa snaží o presnejšiu diagnostiku vzniknutej chyby, napríklad aká fonéma bola v skutočnosti vyslovená.

Zameriame sa výlučne na detekciu nesprávnej výslovnosti, nakoľko jej správne fungovanie priamo ovplyvňuje prípadnú diagnostiku. Našim cieľom bude primárne snaha o zlepšenie existujúcich metód, ktoré sú k tomuto účelu používané. Pokúsime sa o to zavedením rôznych heuristík a na záver otestujeme vplyv použitia systémov rozpoznávania reči trénovaných na viacerých jazykoch.

Práca je štrukturovaná do niekoľkých kapitol. V kapitole 2 sú všeobecne popísané systémy rozpoznávania reči. Nasleduje kapitola 3, ktorá poskytuje základné informácie o neurónových sieťach a ich použití k rozpoznávaniu reči. V kapitole 4 sú rozobraté dôležité prístupy používané na detekciu nesprávnej výslovnosti. Kapitola 5 sa venuje charakteristike datasetov, ktoré budú použité pri našich experimentoch. Asi najdôležitejšou časťou práce je kapitola 6 zameraná na návrh systému a popis zmien zacielených na zlepšenie úspešnosti. Na koniec popíšeme priebeh experimentov a analýzu dosiahnutých výsledkov v kapitolách 7 a 8.

Kapitola 2

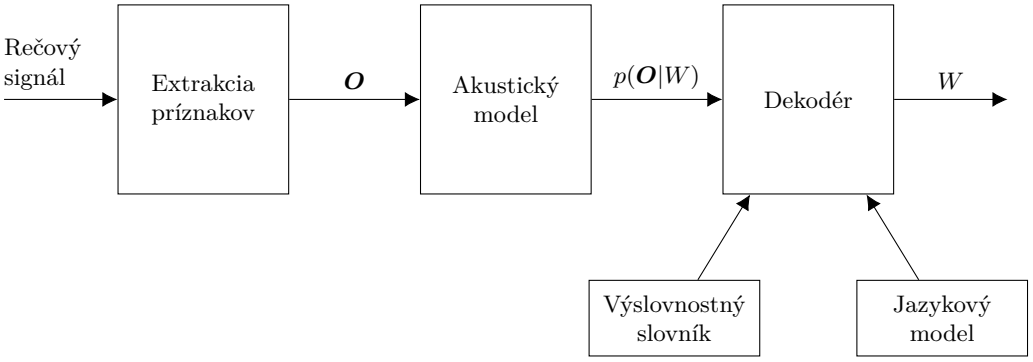
Systémy rozpoznávania reči

Nakoľko hodnotenie výslovnosti je vo väčšine prípadov založené na automatickom rozpoznávaní reči (ASR, z angl. *Automatic Speech Recognition*), v tejto kapitole si bližšie priblížime fungovanie systémov rozpoznávania reči a ich základné časti. Text tejto kapitoly vychádza z nasledovných publikácii [13, 34, 43, 44].

Pod problémom rozpoznávania reči rozumieme hľadanie najpravdepodobnejšej postupnosti slov, ktorá zodpovedá danému rečovému signálu. Naším cieľom je teda nájsť takú postupnosť slov $\hat{W} = (w_1, w_2, \ldots, w_K)$, pre ktorú je aposteriórna pravdepodobnosť $P(W|\boldsymbol{O})$ maximálna a kde $\mathbf{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ je rečový signál vo forme vektoru príznakov, viď sekciu 2.1. Túto úlohu je však obtiažne riešiť priamo a preto s využitím Bayesovho teorému definujme ekvivalentný problém

$$\hat{W} = \arg\max_{W} p(\boldsymbol{O}|W)P(W), \tag{2.1}$$

kde $\hat{W}$ je najpravdepodobnejšia postupnosť slov. K určeniu vierohodnosti $p(\boldsymbol{O}|W)$ sa používa tzv. akustický model, ktorý nesie informáciu o tom, ako znejú jednotlivé fonémy. Pravdepodobnosť $P(W)$ sa stanovuje pomocou jazykového modelu, ktorý nám poskytuje informáciu o tom, ktoré postupnosti slov sú v danom jazyku pravdepodobné. Výsledný systém rozpoznávania reči je potom možné rozdeliť do niekoľkých častí ako je to znázornené na obr. 2.1. V nasledujúcich sekciách si fungovanie jednotlivých častí bližšie popíšeme.



Obr. 2.1: Základné časti systému pre rozpoznávanie reči.

2.1 Extrakcia príznakov

Úlohou extrakcie príznakov je previesť rečový signál do podoby, ktorá je vhodná pre následné rozpoznávanie. Primárnou snahou je odstrániť prebytočnú informáciu, ktorú nesie pôvodný signál. Ďalšie požiadavky na vlastnosti príznakov vyplývajú z použitého akustického modelu. Samotný výpočet príznakov potom prebieha nad krátkymi, čiastočne sa prekrývajúcimi úsekmi reči.

Medzi najpoužívanejší typ príznakov patria melovské kepstrálne koeficienty (MFCC) [8]. Tieto príznaky sa snažia zohľadňovať nelineárne vnímanie frekvencií ľudským uchom s využitím tzv. melovej frekvenčnej škály. Algoritmus výpočtu pozostáva z určenia spektrálnych energií pomocou diskrétnej Fourierovej transformácie, váhovania pomocou banky filtrov rozmiestnených v melovej frekvenčnej škále, logaritmovania získaných hodnôt a nakoniec aplikovania diskrétnej kosínusovej transformácie (*Discrete Cosine Transform*, DCT).

Na podobnom princípe sú založené aj perceptívne lineárne prediktívne (*Perceptual Linear Predictive*, PLP) koeficienty [18], ktoré v porovnaní s MFCC dosahujú v niektorých prípadoch nepatrne lepšie výsledky [42].

Pri akustických modeloch založených na neurónových sieťach sa okrem vyššie zmienených príznakov zvyknú používať aj tzv. Fbank príznaky, ktoré sa líšia od MFCC len vo vynechaní diskrétnej kosínusovej transformácie. Mohamed a kol. [26] dosiahli pri ich použití značné zlepšenie v porovnaní s MFCC.

2.2 Akustický model

Základnou jednotkou, ktorú reprezentuje akustický model je fonéma. Preto je pre určenie vierohodnosti $p(\boldsymbol{O}|W)$ postupnosti slov $W = (w_1, w_2, \ldots, w_K)$ potrebné každé slovo w_k dekomponovať na postupnosť foném. K tomuto účelu slúži výslovnostný slovník, ktorý pre každé slovo jazyka obsahuje informáciu o jeho výslovnosti vo forme fonetického prepisu na fonémy. Jedno slovo však môže mať viacej variant výslovnosti, čo vedie na k nasledovnému výpočtu vierohodnosti

$$p(\boldsymbol{O}|W) = \sum_{Q} p(\boldsymbol{O}|Q)p(Q|W), \tag{2.2}$$

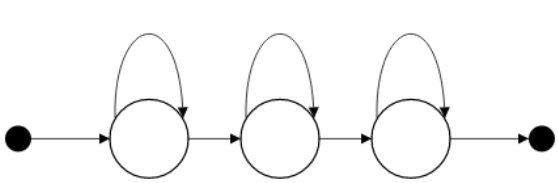
kde suma je nad všetkými možnými výslovnosťami reťazca W, t.j. každé Q je postupnosť výslovností slov $Q = (Q_1, Q_2, \ldots, Q_K)$, pričom každé Q_k je postupnosť foném $Q_k = (q_1, q_2, \ldots, q_m)$. Potom

$$p(Q|W) = \prod_{k=1}^{K} p(Q_k|w_k), \tag{2.3}$$

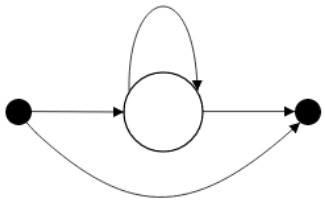
kde $p(Q_k|w_k)$ je pravdepodobnosť, že slovo w_k je vyslovené ako postupnosť foném Q_k. V praxi sa suma v rovnici (2.2) často aproximuje pomocou maxima.

Skryté Markovove modely

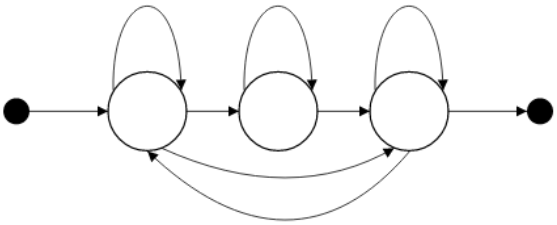
Akustické modelovanie s využitím skrytých Markovových modelov (*Hidden Markov Model*, HMM) je založené na predpoklade, že postupnosť príznakových vektorov $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ je generovaná pomocou nejakého skrytého Markovovho modelu. Skrytý Markovov model môžeme chápať ako konečný automat pozostávajúci zo stavov s_i a pravdepodobnostných



(a) model fonémy



(b) model krátkej pauzy



(c) model ticha

Obr. 2.2: Topológie rôznych HMM modelov, ktoré pozostávajú z neemitujúcich počiatočných a koncových stavov, emitujúcich stavov medzi nimi a orientovaných prechodov medzi týmito stavmi.

prechodov a_{ij} medzi jednotlivými stavmi s_i a s_j. Pri prechode do stavu s_j v čase t je zároveň generovaný príznakový vektor $\boldsymbol{o}_t$ pomocou rozdelenia pravdepodobnosti $b_j(\boldsymbol{o}_t)$, ktorá zodpovedá tomuto stavu.

Ako sme už spomenuli, základnou jednotku, ktorú modelujeme, je fonéma. Jedna fonéma býva zvyčajne reprezentovaná pomocou topológie znázornenej na obrázku 2.2a. Okrem foném je však zväčša potrebné modelovať aj krátke medzislovné pauzy, obr. 2.2b, a dlhšie trvajúce ticho, obr. 2.2c, ktoré je typické zväčša pre začiatok a koniec viet. Model odpovedajúci nejakej postupnosti slov W je potom možné skonštruovať zreťazením takýchto topológii za seba.

Vierohodnosť sekvencie príznakov $\boldsymbol{O}$ pre daný model M a danú postupnosť stavov S zodpovedá vzťahu

$$p(\boldsymbol{O}|M,S) = a_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\boldsymbol{o}_t) a_{s(t)s(t+1)}, \tag{2.4}$$

kde stavy $s(0)$, resp. $s(T+1)$, predstavujú počiatočný a koncový stav, ktoré negenerujú žiadny výstupný vektor. Nakoľko je však postupnosť stavov S skrytá, pre určenie vierohodnosti $p(\boldsymbol{O}|M)$ musíme uvažovať všetky možné sekvencie stavov S. Dostávame teda

$$p(\boldsymbol{O}|M) = \sum_{S} p(\boldsymbol{O}|M,S). \tag{2.5}$$

Nakoľko nás pri rozpoznávaní reči zaujíma zväčša len najpravdepodobnejšia postupnosť stavov pre danú sekvencia príznakov, môžeme zaviesť nasledovnú aproximáciu

$$\hat{p}(\boldsymbol{O}|M) = \max_{S} \left\{ a_{s(0)s(1)} \prod_{t=1}^{T} b_{s(t)}(\boldsymbol{o}_t) a_{s(t)s(t+1)} \right\}. \tag{2.6}$$

Takýto problém je možné riešiť pomocou Viterbiho algoritmu [35], ktorý je schopný určiť najpravdepodobnejšiu postupnosť S a teda aj maximálnu hodnotu vierohodnosti.

Posledným detailom, ktorý sme doposiaľ prehliadali je určenie prechodových pravdepodobností a_{ij} a rozdelenia výstupných pravdepodobností b_j. Pre modelovanie b_j sa v minulosti často používal model zmesí normálnych rozložení (*Gaussian Mixture Model*, GMM). V súčasnosti však býva nahrádzaný neurónovými sieťami, ktoré sú schopné dosahovať výrazne lepších výsledkov. GMM sa potom využíva len na určenie zarovnaní, na ktorých trénujeme neurónovú sieť, ako tomu bude aj v tejto práci. Preto nebudeme popisovať samotný algoritmus trénovania a_{ij} a b_j, ktorý je pomerne rozsiahly. Dobrý popis je možné nájsť napr. v [35].

2.3 Jazykový model

Ďalšou dôležitou časťou pri rozpoznávaní reči je jazykový model, ktorého úlohou je pre ľubovoľnú postupnosť slov $W = (w_1, w_2, \ldots, w_K)$ určiť aposteriórnu pravdepodobnosť $P(W)$, ktorá je daná vzťahom

$$P(W) = \prod_{k=1}^{K} P(w_k | w_{k-1}, \ldots, w_1). \tag{2.7}$$

Jednotlivé podmienené pravdepodobnosti v rovnici (2.7) je potrebné odhadnúť z trénovacích dát, čo bude s narastajúcou dĺžkou K postupnosti W čím ďalej tým náročnejšie, až takmer nemožné. Tento problém sa snažia riešiť N-gramové jazykové modely, ktoré sú zároveň najpoužívanejšími modelmi používanými k jazykovému modelovaniu.

N-gramové jazykové modely riešia uvedený problém aproximáciou, pri ktorej sú pravdepodobnosti v (2.7) podmienené len $N-1$ poslednými slovami, t.j. dostávame

$$P(W) = \prod_{k=1}^{K} p(w_k | w_{k-1}, w_{k-2}, \ldots, w_{k-N+1}), \tag{2.8}$$

pričom hodnota N je zvyčajne v rozmedzí $2-3$. K odhadu podmienených pravdepodobností potom postačuje nad trénovacími dátami určovať počty výskytov postupností slov, napr. pre $N = 3$ je odhad daný vzťahom

$$p(w_k | w_{k-1}, w_{k-2}) \approx \frac{C(w_{k-2}, w_{k-1}, w_k)}{C(w_{k-2}, w_{k-1})}, \tag{2.9}$$

kde $C(w_{k-2}, w_{k-1}, w_k)$ je počet trigramov $w_{k-2}w_{k-1}w_k$ a $C(w_{k-2}, w_{k-1})$ je počet bigramov $w_{k-2}w_{k-1}$ nachádzajúcich sa v trénovacích dátach.

Napriek obmedzenej histórii dochádza pri týchto modeloch k problému s nedostatkom trénovacích dát, a to napriek tomu, že k trénovaniu postačujú dáta len v textovej podobe. K jeho riešeniu je možné použiť napr. ústupové (*backing-off*) vyhladzovanie. Pre viac informácii viď [34].

2.4 Dekódovanie

Ako sme si už uviedli v úvode, cieľom rozpoznávania je nájdenie najpravdepodobnejšej postupnosti slov $\hat{W} = (w_1, w_2, \ldots, w_K)$ v súlade so vzťahom (2.1). Ak by sme však postupovali

spôsobom uvedeným vyššie, t.j. určovali by sme hodnoty $p(\boldsymbol{O}|W)$ a $P(W)$ pre každú prípustnú postupnosť slov W, už pri malom slovníku by bol výpočet nerealizovateľný. Z tohto dôvodu bola od počiatku snaha o hľadanie efektívnejšieho spôsobu dekódovania.

Rozšíreným spôsobom dekódovania je v dnešných systémoch rozpoznávania reči používanie ohodnotených konečných transducerov (*Weighted Finite-State Transducer*, WFST), čo je v podstate konečný automat s ohodnotenými hranami, ktorý umožňuje nielen čítanie symbolov, ale aj ich generovanie. Veľkou výhodou tohto výpočetného modelu je možnosť reprezentácie všetkých spomenutých častí potrebných pre dekódovanie, t.j.

- skrytý Markovov model transducerom H,
- výslovnostný slovník transducerom L,
- jazykový model transducerom G.

Potom s využitím operácií kompozície, determinizácie a minimizácie sme schopný jednotlivé transducery skomponovať do výsledného transduceru HLG. Takouto kombináciou sme dosiahli výraznej redukcie stavového priestoru, ktorý je potrebné počas dekódovania prehľadávať. Pre samotné dekódovanie je potom možné použiť napr. Viterbiho paprskové (*beam*) dekódovanie. Pre viac informácii o použití WFST pre rozpoznávanie reči viď [27].

Kapitola 3

Neurónové siete

Táto kapitola popisuje obecne princíp neurónových sietí, ktoré využijeme v našej práci na viacero úloh – odhad pravdepodobností v akustickom modeli, odhad fonologických rysov, a taktiež na klasifikáciu výslovnosti. Informácie v tejto kapitole sú primárne založené na publikácii [5].

3.1 Dopredné neurónové siete

Neurónové siete sú dnes dominantnou metódou využívanou na strojové učenie, kde dosahujú výrazne lepšie výsledky ako iné konvenčné metódy. Neurónová sieť je biologicky inšpirovaný matematický model, ktorý mapuje vstupné vektory na zodpovedajúce výstupné vektory. Dimenzie týchto vektorov môžu byť obecne rôzne.

Najrozšírenejším typom neurónovej siete je neurónová sieť s dopredným šírením (*feed-forward neural network*), v ktorej sa informácia šíri jedným smerom zo vstupu na výstup. To znamená, že v nej neexistujú žiadne spätné väzby.

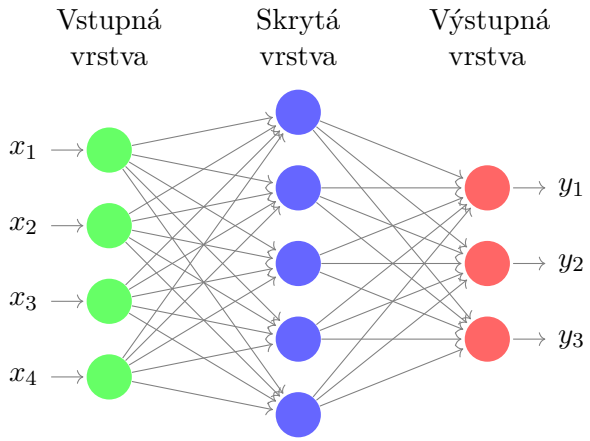
Neuróny v doprednej neurónovej sieti sa zvyčajne organizujú do vrstiev, kde prvá vrstva sa označuje ako vstupná, za ňou nasleduje skrytá vrstva a posledná vrstva je vrstvou výstupnou. Skrytých vrstiev môže sieť obsahovať niekoľko, kedy takéto siete zvykneme označovať ako hlboké neurónové siete (*Deep Neural Network*, DNN), ale nie sú výnimkou ani siete bez skrytej vrstvy. Najčastejšie sú jednotlivé vrstvy medzi sebou plne prepojené, čo znamená, že všetky výstupy z jednej vrstvy sú privedené na všetky vstupy tej nasledujúcej. Príklad neurónovej siete s jednou skrytou vrstvou je možné vidieť na obrázku 3.1.

Ako už bolo naznačené, základnou jednotkou neurónovej siete je neurón, ktorý transformuje vstupný vektor $x = (x_1, \ldots, x_N)$ na hodnotu y pomocou vzťahu

$$y = f\left(\sum_{i=1}^{N} w_i x_i + b\right), \tag{3.1}$$

kde $w = (w_1, \ldots, w_N)$ predstavuje vektor váh, b prahovú hodnotu, a f je nejaká vhodne zvolená funcia. Ak rozšírime vstupný vektor x o hodnotu $x_0 = 1$, môžeme prahovú hodnotu b zakomponovať do vektoru w, čiže $w_0 = b$. Potom môžeme vyššie uvedený vzťah upraviť na

$$y = f\left(\sum_{i=0}^{N} w_i x_i\right). \tag{3.2}$$



Obr. 3.1: Príklad neurónovej siete s jednou skrytou vrstvou.

Podstatnou časťou, ktorá má vplyv na správne fungovanie neurónovej siete, je funkcia $f(.)$, zväčša označovaná ako aktivačná funkcia. Aby bola neurónová sieť schopná realizovať nelineárne mapovanie vstupov na výstup, musí byť aj táto funkcia nelineárna. Okrem toho k nej musí existovať derivácia, aby bolo možné trénovanie s využitím spätného šírenia chyby. Vhodnými funkciami sú napr. logistická sigmoida, hyperbolický tangens alebo ReLU, viď tabuľku 3.1. Posledná menovaná je vhodná najmä pre použitie v skrytých vrstvách DNN, nakoľko umožňuje lepšie propagovanie gradientov pri trénovaní [16]. V prípade použitia neurónovej siete na klasifikáciu sa pri výstupnej vrstve využíva výhradne logistická sigmoida alebo softmax funkcia. Výstupom oboch funkcií sú hodnoty od 0 do 1, ktoré je možné interpretovať ako pravdepodobnosti. Softmax funkcia naviac realizuje normalizáciu cez všetky neuróny danej vrstvy, vďaka čomu je suma výstupných hodnôt rovná 1. To sa hodí najmä pri klasifikácii 1 z N, kedy potrebujeme, aby výstupy reprezentovali príslušnosť do nejakej triedy.

Názov	Funkcia	Derivácia
Logistická sigmoida	$f(x) = \frac{1}{1+e^{-x}}$	$f'(x) = f(x)(1-f(x))$
Hyperbolický tangens	$f(x) = \frac{e^x - e^{-x}}{e^x - e^{-x}}$	$f'(x) = 1 - f(x)^2$
ReLU	$f(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$	$f'(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases}$

Tabuľka 3.1: Najčastejšie používané aktivačné funkcie a ich derivácie.

Trénovanie

Cieľom trénovania je nájdenie vhodných parametrov w u jednotlivých neurónov. K tomu sa používa tzv. iteratívne trénovanie s učiteľom, kedy sa v každej iterácii vypočíta chyba medzi predikciami neurónovej siete a známym výstupom. Na základe tejto chyby sa potom upravia váhy jednotlivých neurónov. Najčastejšie sa k tomu používa algoritmus *stochas-*

tic gradient-descent v kombinácii so spätnou propagáciou chyby, ktorý aktualizuje váhy neurónov v jednotlivých vrstvách výpočtom gradientov chybovej funkcie.

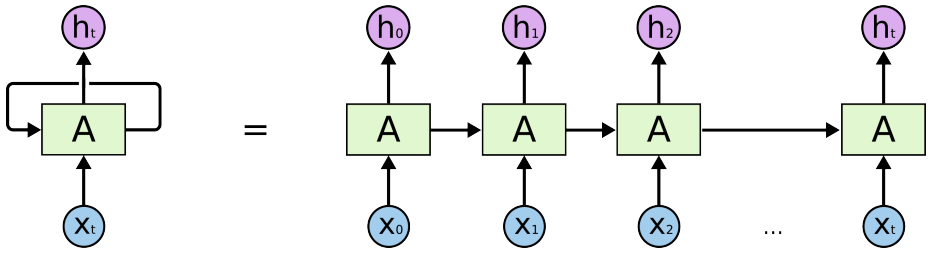
Dôležitá je aj voľba samotnej chybovej funkcie. Najpoužívanejšími sú stredná kvadratická chyba, krížová entropia a kategorická krížová entropia. Posledné dve menované sú vhodné pre klasifikačné úlohy, pričom kategorická krížová entropia sa využíva výhradne v kombinácii so softmax výstupnou vrstvou.

V tejto sekcii sme uviedli len základný popis dopredných neurónových sietí a ich fungovania. Pre ďalšie podrobnosti viď [5].

3.2 Rekurentné neurónové siete

Rekurentné neurónové siete (*Recurent Neural Networks*, RNN) narozdiel od dopredných sietí umožňujú zavádzanie spätných väzieb. To spôsobí, že aktuálny výstup siete nie je závislý len na aktuálnom vstupe, ale aj na vstupoch predchádzajúcich. RNN teda umožujú zachytiť časové závislosti medzi dátami, ktoré sú dôležité pre mnoho úloh, ako je aj spracovanie reči.

Obrázok 3.2 znázorňuje jednoduchú RNN pozostávajúcu z jedinej vrstvy H. Na jej vstup je okrem aktuálneho vzorku x_t privedený aj predchádzajúci výstup h_{t-1} odpovedajúci vzorke x_{t-1}. Na takúto sieť môžeme nazerať ako na niekoľko kópii tej istej siete, ktoré predávajú svoje výstupy svojim následovníkom, tak ako je to znázornené na obr. 3.2.



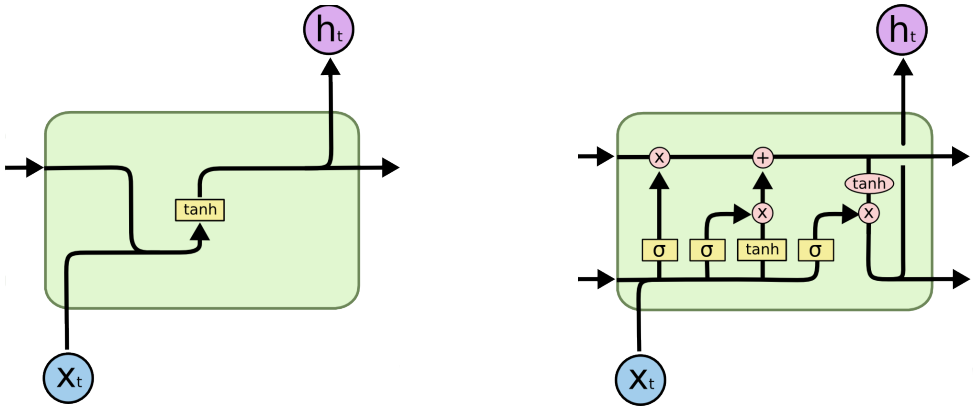
Obr. 3.2: Schéma základnej RNN s jednou skrytou vrstvou. Vľavo je znázornená spätná väzba pomocou cyklu, vpravo rozbalená sieť pre jednotlivé vstupné vzorky $x_0, \ldots, x_t$. Prevzaté z [31].

Siete s takouto architektúrou však majú v praxi problém zachytiť závislosti medzi vzorkami vzdialenými ďaleko od seba. Preto sa v súčasnosti používajú špeciálny typ RNN, tzv. *Long short-term memory* (LSTM) siete, u ktorých tento problém nie je tak výrazný.

LSTM neurónové siete

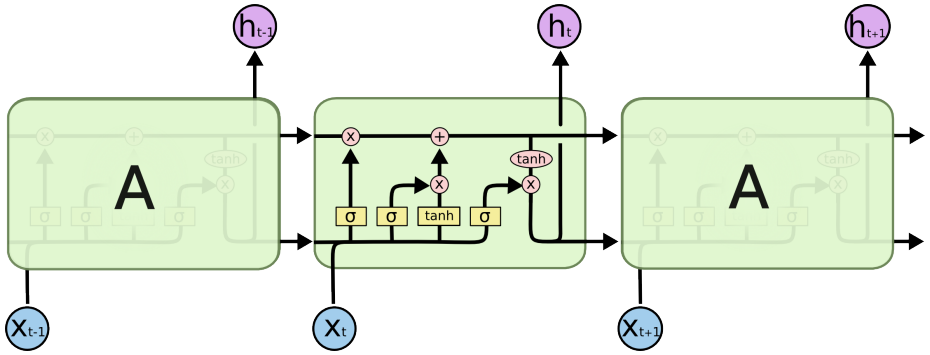
LSTM neurónové siete nie sú organizované do vrstiev, ako tomu bolo doteraz, ale sú tvorené komplexnými jednotkami realizujúcimi viacero operácii nad vstupnými dátami. Porovnanie jednoduchej RNN a LSTM je možné vidieť na obrázku 3.3. Zatiaľ čo RNN tvorí zväčša jediná vrstva, v tomto prípade s aktivačnou funkciou hyperbolický tangens, LSTM jednotka pozostáva z niekoľkých vrstiev s aktivačnými funkciami logistická sigmoida a hyperbolický tangens. Výstupy týchto vrstiev sú následne kombinované operáciami násobenia a sčítania, čím produkujú dva výstupy h_t a C_t. Druhý menovaný výstup slúži len na predanie pomocnej informácie do ďalšieho kroku výpočtu, viď obr. 3.4. Vďaka tejto dodatočnej komunikačnej linke sú jednotky schopné pridať váhu dôležitým vstupným vzorkám v sekvencii, a naopak

potlačiť menej významné vzorky. Pre detailnejší popis fungovania LSTM sietí odporúčame článok [31].



(a) RNN (b) LSTM

Obr. 3.3: Porovnanie architektúry jednoduchej RNN a LSTM neurónovej siete. Prevzaté z [31].



Obr. 3.4: Znázornenie spätných väzieb pri LSTM jednotkách spracovávajúcich vzorky x_{t-1}, x_t, x_{t+1}. Prevzaté z [31].

3.3 Použitie neurónových sietí k rozpoznávaniu reči

V súčasnosti bývajú neurónové siete čoraz častejšie využívané v systémoch rozpoznávania reči. Zväčša bývajú použité na realizáciu tzv. DNN-HMM akustických modelov, kde DNN nahrádza v minulosti často používané modely zmesí normálnych rozložení v GMM-HMM akustických modeloch.

Ako už bolo spomenuté, pri akustickom modeli potrebujeme určiť pre rámce reči $\boldsymbol{o}_t$ výstupné pravdepodobnosti $b_j(\boldsymbol{o}_t)$ pre jednotlivé HMM stavy s_j. Avšak aby sme k tomuto účelu mohli natrénovať neurónovú sieť, musíme najskôr získať zarovnanie reči na jednotlivé stavy s_j. K tomuto účelu sa využíva vopred natrénovaný GMM-HMM model.

Na vstup DNN sa privádzajú príznaky zodpovedajúce rámcu $\boldsymbol{o}_t$. Keďže sa jedná o klasifikáciu 1 z N, výstupná DNN vrstva je tvorená softmax aktivačnou funkciou. Takto natrénovaná sieť potom určuje aposteriórne pravdepodobnosti $p(s_j|\boldsymbol{o}_t)$ stavov s_j. Takto získaná pravdepodobnosť však obsahuje aj informáciu o apriórnej pravdepodobnosti stavov $P(s_j)$,

ktorá je nadbytočná, nakoľko túto informáciu v inej podobe a lepšie modeluje jazykový model. Preto je na záver ešte získané aposteriórne pravdepodobnosti $p(s_j|\boldsymbol{o}_t)$ potrebné previesť na vierodnosti, čo je možné pomocou vzťahu

$$p(\boldsymbol{o}_t|s_j) = \frac{p(s_j|\boldsymbol{o}_t)}{P(s_j)} p(\boldsymbol{o}_t), \tag{3.3}$$

kde $P(s_j)$ reprezentuje apriórne pravdepodobnosti $P(s_j)$, ktoré sa stanovia frekvenčnou analýzou. Čo ale nedokážeme určiť, je apriórna pravdepodobnosť príznakov rámca $p(\boldsymbol{o}_t)$. Našťastie to však ani nie je potrebné, nakoľko nie je závislá na stave s_j, takže ju je možné zo vzťahu úplne vypustiť, čím dostaneme

$$p(\boldsymbol{o}_t|s_j) = \frac{p(s_j|\boldsymbol{o}_t)}{P(s_j)}. \tag{3.4}$$

Kapitola 4

Hodnotenie výslovnosti

V rámci tejto kapitoly si priblížime problematiku hodnotenia výslovnosti nenatívnej reči. Na začiatok si obecne popíšeme charakteristiky nenatívnej reči a vymedzíme si dôležité pojmy, ktoré tvoria teoretický základ pre dostatočné pochopenie daných metód. V ďalšej časti sa zameriame na hodnotenie výslovnosti z pohľadu výuky cudzích jazykov, čo nám pomôže identifikovať dôležité aspekty, ktoré budú kľúčové pre návrh vlastného systému. Vo zvyšku kapitoly sa potom budeme venovať výlučne automatickému hodnoteniu výslovnosti a popíšeme si jednotlivé prístupy, ktoré sa v tejto oblasti používajú.

4.1 Problematika chybnej výslovnosti

Chyby vo výslovnosti v cudzom jazyku (L2) bývajú zapríčinené viacerými faktormi, avšak najviac ovplyvňuje výslovnosť rečníka jeho materinský jazyk (L1) [28]. Rozlišujeme dva druhy chýb, segmentálne, označované aj ako fonémické, a prozodické chyby [40]. Segmentálne chyby predstavujú substitúcie za iné fonémy, vynechávanie alebo vkladanie foném. Okrem toho môžeme do týchto chýb radiť aj menej závažné chyby, kedy je správna fonéma ako tak vyslovená, avšak je tam stále určitá odlišnosť od natívnej výslovnosti. Napriek tomu, že sú tieto chyby patrné pre natívneho rečníka, nemajú vplyv na porozumenie. Vo výuke jazykov preto prevláda názor, že sú prirodzenou súčasťou prejavu väčšiny rečníkov, a nekladie sa na ne veľký dôraz [6].

Prozodickými chybami rozumieme napr. nesprávny dôraz, rytmus, intonáciu a pod. Dôležitosť prozódie sa líši jazyk od jazyka, zatiaľ čo v napr. v slovenčine prozodické chyby nemajú na porozumenie zásadný vplyv, v tónových jazykoch, ako napr. čínština, môžu viesť až k zmene významu. V našej práci sa budeme zameriavať výhradne na segmentálne chyby. Pre viac informácii o prozodických chybách a ich detekcii viď [40].

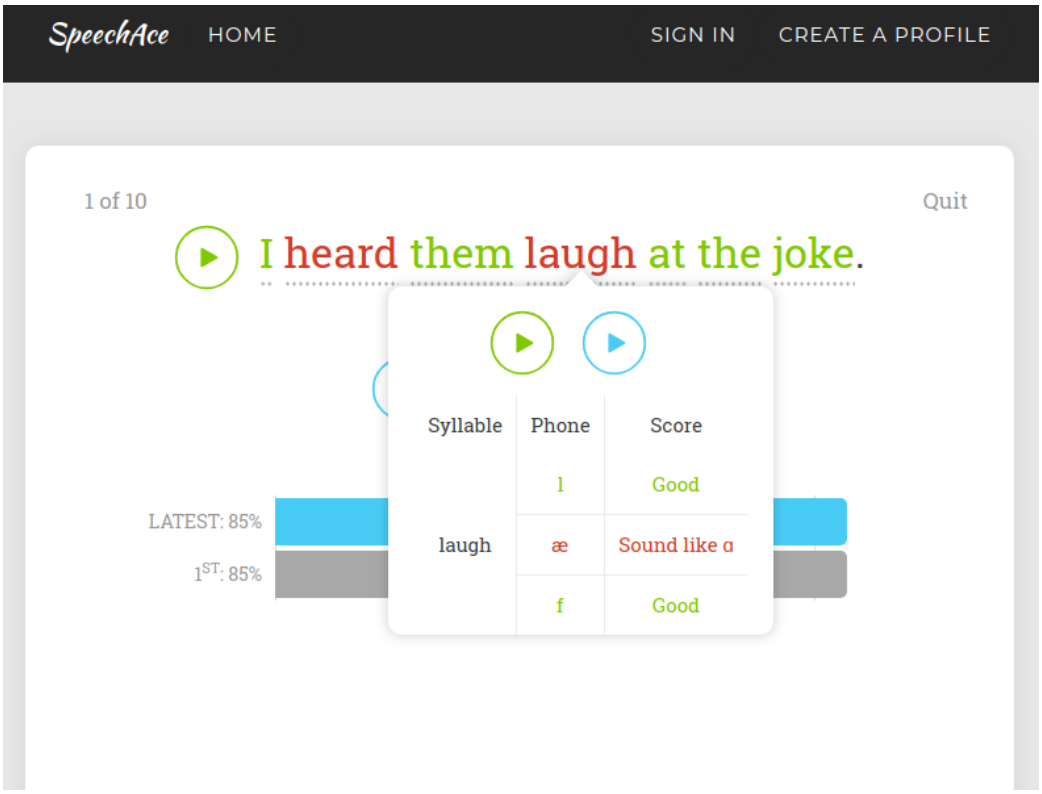
4.2 Výuka výslovnosti s využitím počítača

Výuka výslovnosti s využitím počítača (*Computer-assisted Pronunciation Training*, CAPT) môže mať rôzne podoby, od najjednoduchších programov, ktoré užívateľovi ponúkajú audio materiály, cvičenia a pod., až po komplexné systémy, kde používateľ číta zadaný text a systém automaticky ohodnotí jeho výslovnosť. Príklad existujúceho nástroja[1] slúžiaceho na výuku anglickej výslovnosti je možné vidieť na obr. 4.1.

[1] Nástroj SpeechAce dostupný online na adrese https://app.speechace.co.

Poskytovanie spätnej väzby je pre menej pokročilých študentov veľmi dôležité, nakoľko majú značné problémy s identifikovaním vlastných chýb [9]. Neri. a kol. [30] pozorovali výrazné zlepšenie výslovnosti u študentov používajúcich nástroj, ktorý vyznačoval zle vyslovené fonémy, v porovnaní so študentmi, ktorý si mali možnosť iba porovnávať svoju nahrávku s referenčnou. K podobným záverom dospeli aj ďalšie práce [33, 38]. Ďalšie zlepšenie výuky by mohlo priniesť diagnostikovanie druhu chyby, napr. že bola vyslovená fonéma /ʌ/ namiesto fonémy /æ/. Takáto spätná väzba by však už vyžadovala oboznámenie študentov so základmi fonetiky a fonológie daného jazyka, aby mohli takúto informáciu využiť vo svoj prospech. Preto by bolo asi vhodnejšie túto informáciu štylizovať skôr do podoby nápovedy, napr. požiadať užívateľa, aby pri vyslovení tejto fonémy výraznejšie otvoril ústa.

Nemenej dôležitým aspektom týchto systémov je ich presnosť, nakoľko mylné hodnotenie výslovnosti by mohlo užívateľov miasť a čoskoro ich od používania takýchto nástrojov celkom odradiť [30].



Obr. 4.1: Webové rozhranie nástroja SpeechAce, ktorý slúži na výuku anglickej výslovnosti. Nástroj umožňuje detekciu a diagnostiku segmentálnych chýb. Zle vyslovené fonémy sú označené a je poskytnutá informácia, za aký foném boli substituované. Chýba podpora detekcie chýb založených na vkladaní a vypúšťaní foném.

4.3 Metódy automatického hodnotenia výslovnosti

Pod pojmom automatické hodnotenie výslovnosti rozumieme nejaký algoritmus, ktorý priradí každému segmentu reči číselné skóre, ktoré hovorí, do akej miery je daný segment správne vyslovený. Následne sme aplikovaním určitého prahu schopný rozhodnúť, že daný segment je alebo nie je správne vyslovený. K rozdeleniu nahrávky na segmenty potrebujeme

poznať text, ktorý by mal v danej nahrávke odznieť, pomocou čoho je možné vytvoriť kanonický fonémový prepis, čo je prepis na úrovni foném získaný s využitím výslovnostného slovníka. Samotné rozdelenie na segmenty je potom realizované núteným zarovnaním voči tomuto prepisu s využitím ASR systému. Akustický model takéhoto ASR trénujeme s využitím nahrávok od natívnych resp. nenatívnych rečníkov, v závislosti na použitej metóde. V prípade nenatívnych nahrávok je niekedy potrebný aj skutočný prepis, čo je prepis na fonémy, ktoré v nahrávke rečník vyslovil. Zvyčajne sa na jeho tvorbe podieľajú fonetici, ktorý majú skúsenosť s daným jazykom.

Ako poznamenali v [39], problém hodnotenia výslovnosti je podobný určovaniu miery dôveryhodnosti rozpoznávania (*Confidence Measures*, CM) u ASR, kde sa určuje miera istoty, že rozpoznaný výsledok je správny. Analogicky ako v prípade CM [22], môžeme rozdeliť metódy hodnotenia výslovnosti do troch kategórii:

1. metódy založené na teste pomerom vierohodností,
2. metódy založené na aposteriórnej pravdepodobnosti foném,
3. metódy založené na priamej klasifikácii výslovnosti.

V tejto sekcii si podrobnejšie priblížime jednotlivé metódy.

4.3.1 Metódy založené na teste pomerom vierohodností

Test pomerom vierohodností (*Likelihood-ratio test*) je v štatistike využívaný na porovnanie dvoch modelov z hľadiska toho, ako dobre dané modely vyhovujú pozorovaným dátam. Na detekciu nesprávnej výslovnosti ho prvýkrát použil Franco a kol. [12] zavedením tzv. *log-likelihood ratio* (LLR) skóre, ktoré je založené na pomere vierohodností od dvoch rôznych GMM-HMM modelov λ_M a λ_C. Obidva modely sú natrénované na foneticky anotovanom datasete nenatívnej reči, avšak pri trénovaní modelu λ_M boli použité len nahrávky od rečníkov s veľmi dobrou výslovnosťou blížiacej sa natívnej úrovni, a model λ_C na nahrávkach s nesprávnou, akcentovanou výslovnosťou. Výpočet pre určitý segment reči, ktorý bol získaný núteným zarovnaním, je potom nasledovný

$$\mathrm{LLR}(p) = \frac{1}{d} \sum_{t=t_0}^{t_0+d-1} \left(\log p(y_t|p, \lambda_M) - \log p(y_t|p, \lambda_C) \right), \tag{4.1}$$

kde d je počet rámcov v danom segmente, t_0 predstavuje počiatočný index segmentu, $p(y_t|p, \lambda_M)$ je vierohodnosť určená modelom λ_M, že rámec reči y_t zodpovedá foneme p a analogicky $p(y_t|p, \lambda_C)$ je vierohodnosť určená modelom λ_C pre ten istý rámec y_t a fonému p. Normalizácia dĺžkou d zabezpečuje konzistentné skóre nezávislé na dĺžke segmentu. Čím je hodnota tohto skóre vyššia, tým je vyššia pravdepodobnosť, že segment je nesprávne vyslovený.

Franco a kol. vo svojej práci [12] dosiahli pri použití tohto skóre značné zlepšenie v porovnaní so skóre založeného na aposteriórnej pravdepodobnosti foném, ktoré je popísané v nasledujúcej sekcii. Veľkou nevýhodou tejto metódy je potreba nenatívneho datasetu, ktorého obstaranie je značne náročné.

4.3.2 Metódy založené na aposteriórnej pravdepodobnosti foném

Spoločným menovateľom metód založených na aposteriórnej pravdepodobnosti foném je snaha o čo najlepšiu aproximáciu pri výpočte aposteriórnej pravdepodobnosti $P(p|O^{(p)})$, že

foném p odpovedá segmentu reči $O^{(p)}$. Pre získanie vierohodností potrebných k výpočtu je použitý akustický model, ktorý je natrénovaný na natívnom datasete, čo je veľkou výhodou týchto metód. Objavujú sa ale aj práce, v ktorých sa rozhodli pre použitie nenatívneho datasetu [2] s fonetickým prepisom od fonetikov.

Tieto metódy sa zvyknú označovať aj ako *Goodness of pronunciation* (GOP) skóre. Hoci tento názov zodpovedal konkrétnej metóde zavedenej v práci [41], postupnými modifikáciami pôvodného výpočtu sa rozšíril na celú triedu týchto metód.

Na tomto mieste si popíšeme originálne GOP skóre zavedené v [41], ktoré sa dodnes v určitých podobách stále používa. Ako už bolo spomenuté, cieľom GOP je teda určenie normalizovanej aposteriórnej pravdepodobnosti, že foném p odpovedá akustickému segmentu $O^{(p)}$, t.j.

$$\mathrm{GOP}(p) \equiv \log P(p|O^{(p)})/d \tag{4.2}$$

$$= \log \left(\frac{p(O^{(p)}|p)P(p)}{\sum_{q \in Q} p(O^{(p)}|q)P(q)} \right) \Big/ d, \tag{4.3}$$

kde Q je množina všetkých foném a d je počet rámcov akustického segmentu $O^{(p)}$, $P(p)$, resp. $P(q)$, je apriórna pravdepodobnosť fonémy p, resp. q, určené jazykovým modelom. Ešte pred tým, než si vysvetlíme význam $p(O^{(p)})$, resp. $p(O^{(q)})$, tak si zavedieme dve zjednodušenia zavedením predpokladov, že všetky fonémy sa vyskytujú s rovnakou pravdepodobnosťou ($P(p) = P(q)$) a sumu v menovateli je možné aproximovať maximom. Dostávame teda

$$\mathrm{GOP}(p) = \log \left(\frac{p(O^{(p)}|p)}{\max_{q \in Q} p(O^{(p)}|q)} \right) \Big/ d. \tag{4.4}$$

Určenie vierohodnosti $p(O^{(p)}|p)$ v čitateli prebieha obvyklým spôsobom, t.j. súčinom vierohodností $p(y_t|p)$ po jednotlivých rámcoch y_t daného segmentu

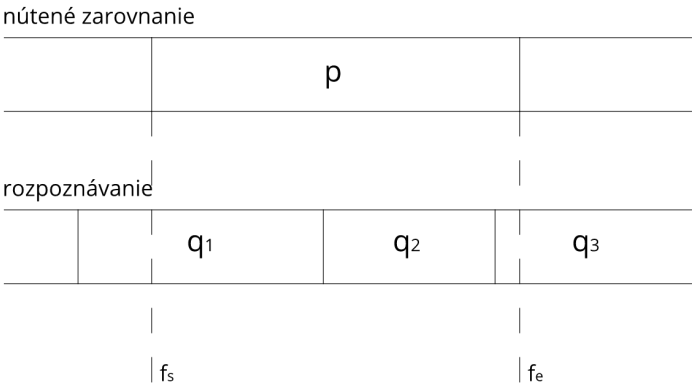
$$p(O^{(p)}|p) = \prod_{t=t_0}^{t_0+d-1} p(y_t|p), \tag{4.5}$$

kde t_0 je index prvého rámca segmentu $O^{(p)}$ a $p(y_t|p)$ je vierohodnosť určená akustickým modelom.

V prípade menovateľa však postupujeme odlišne. Namiesto výpočtu maxima totiž vykonáme nad danou nahrávkou fonémové rozpoznávanie, čím dostaneme najpravdepodobnejší foném q, ktorý zodpovedá segmentu $O^{(p)}$. V prípade zlej výslovnosti sa však stáva, že zarovnanie získané pri rozpoznávaní sa líši od núteného zarovnania. To má za následok, že segmentu $O^{(p)}$ zodpovedá hneď niekoľko foném $q_1, \ldots, q_N$, napr. ako je tomu na obrázku 4.2. Z toho dôvodu určíme výslednú vierohodnosť $p(O^{(p)}|q)$ ako

$$p(O^{(p)}|q) = \prod_{i=1}^{N} p(O^{(p)}|q_i). \tag{4.6}$$

kde jednotlivé vierohodnosti $p(O^{(p)}|q_i)$ získame analogickým spôsobom ako v prípade čitateľa, t.j. súčinom po jednotlivých rámcoch, avšak s ohľadom na hranice dané núteným zarovnaním a rozpoznávaním, viď už spomínaný obrázok 4.2.



Obr. 4.2: Výsledok zarovnania získaného rozpoznávaním

Tento odlišný výpočet však spôsobí, že skóre môže nadobúdať kladných aj záporných hodnôt, čo by vyžadovalo aj dve prahové hodnoty. Preto vo vzťahu 4.7 aplikujeme na výsledok absolútnu hodnotu, čím dostaneme výsledný vzťah

$$\mathrm{GOP}(p) = \left|\log\left(\frac{p(O^{(p)}|p)}{\max_{q\in Q} p(O^{(p)}|q)}\right)\right| \Big/ d. \tag{4.7}$$

Odteraz nám bude postačovať len jedna hodnota prahu. Narozdiel od predchádzajúceho vzťahu však vyššia hodnota GOP skóre predstavuje horšiu výslovnosť, nie lepšiu.

Toto skóre sa na dlhú dobu stalo de facto štandardom v hodnotení výslovnosti a v mnohých prácach sa dodnes používa ako referenčné hodnotenie. S rozmachom používania hlbokých neurónových sietí *(DNN – Deep Neural Network)* bola prirodzene snaha o ich použitie aj pri výpočte GOP skóre [19], čo podľa očakávania viedlo k obdobnému zlepšeniu ako pri ich použití v ASR systémoch.

4.3.3 Metódy založené na priamej klasifikácii

Veľkou výhodou doposiaľ zmieňovaných metód bola ich jednoduchosť a ľahká realizovateľnosť, nakoľko si vystačili s existujúcimi časťami bežných ASR systémov. Na druhú stranu však dosahujú pomerne nízku mieru presnosti, čo môže spôsobovať napr. rovnaký výpočet skóre pre všetky fonémy. Toto je síce možné čiastočne zlepšiť použitím rôznych prahov pre jednotlivé fonémy [41], ale výrazne lepšie dokážu takúto variabilitu pokryť metódy založené na klasifikácii. Tie totiž okrem pravdepodobností z akustického modelu, ako to bolo u doterajších metód, využívajú množstvo ďalších príznakov, ako napr. dĺžky segmentov, fonetické vlastnosti a pod. Na základe takýchto informácii sú niektoré klasifikátory potom schopné chyby nielen detekovať, ale aj diagnostikovať, t.j. určiť, o akú chybu sa jedná.

Prvé použitie klasifikátorov pri hodnotení výslovnosti sa objavilo krátko po zavedení vyššie uvedených metód, kde bola snaha o ich zlepšenie skombinovaním s ďalšími druhmi metrík, ktoré samostatne nemali veľkú výpovednú hodnotu, ako napr. skóre založené na dĺžke segmentov [23] a pod. Franco a kol. [11] porovnávali niekoľko prístupov, ako takúto kombináciu docieliť. Všetky nimi použité klasifikačné metódy dosiahli výrazne lepšie výsledky, ako len pri použití metódy založenej na aposteriornej pravdepodobnosti. K najlepšej

korelácii s ručne anotovanými hodnoteniami viedlo použitie neurónovej siete alebo rozhodovacieho stromu, kde neurónová sieť dosiahla mierne lepších výsledkov, ale jej trénovanie vyžadovalo ďaleko viac úsilia.

Širšie využitie však nachádzajú klasifikátory pri úplne odlišnom prístupe, kedy sa trénujú nad špecifickými fonémovými pármi, ktoré je v danom jazyku potrebné rozlišovať, prípadne stoja za častými chybami vo výslovnosti. V angličtine môže ísť napr. o dvojicu foném /ʌ/ a /æ/. Pre segment obsahujúci jednu z takýchto foném je klasifikátor schopný určiť, či bola naozaj vyslovená správna fonéma, alebo došlo k zámene za druhú fonému z fonémového páru. Príkladom takejto práce je napr. [37], kde autori trénovali klasifikátory založené na lineárnej diskriminačnej analýze (*Linear Discriminant Analysis*, LDA) s využitím niekoľkých akusticko-fonetických vlastností, ktoré boli počítané nad celými segmentmi. Výsledky, ktoré s takýmito klasifikátormi dosiahli, výrazne predčili GOP skóre. K podobným záverom dospeli aj Doremalen a kol. [10], ktorý použili SVM klasifikátory pre jednotlivé fonémové páry natrénované nad celou škálou príznakov: logaritmicko-aposteriórne skóre, MFCC príznaky a niekoľko fonetických vlastností.

Výskum sa v posledných rokoch sa začal orientovať na používanie neurónových sietí, či už na samotnú klasifikáciu alebo aj na extrakciu dodatočných príznakov, ako sú napr. fonetické vlastnosti. Pre ilustráciu spomenieme jednu z týchto prací [2], ktorej autori využívajú k detekcii nesprávnej výslovnosti fonetické vlastnosti, ktorých prítomnosť pre jednotlivé rámce reči určuje hlboká neurónová sieť (*Deep Neural Network*, DNN). Výstupy tejto neurónovej siete vo forme aposteriórnych pravdepodobností sa potom spriemerujú pre celý segment a na základe týchto hodnôt jednouchá neurónová sieť rozhoduje, či je daný segment správne vyslovený. Veľkou výhodou použitia fonetických vlastností je, že v prípade detekcie nesprávnej výslovnosti máme zároveň detailnú informáciu o vzniknutej chybe, na základe ktorej je možné rečníka usmerniť, ako danú chybu odstrániť.

4.4 Rozšírené dekódovacie siete

Rozšírená dekódovacia sieť (*Extended Recognition Network*, ERN) vzniká rozšírením štandardnej dekódovacej siete využívanej v ASR systémoch o množinu prechodov zodpovedajúcich chybným výslovnostiam. Takáto sieť môže byť používaná na detekciu nesprávnej výslovnosti bez potreby žiadnej z uvedených metód v sekcii 4.3 [17, 24]. Zároveň umožňuje okrem detekcie zlej výslovnosti aj jej diagnostiku rozpoznaním chybnej fonémy. Pri tomto prístupe však nemáme informáciu o miere istoty, že daný foném je zle vyslovený a nemôžeme tak výslednú chybu nijak ovplyvniť. Najvhodnejšie sa teda ukazuje jej použitie v kombinácii s predchádzajúcimi metódami, kedy sa ERN využije len na detekciu chýb spôsobených vložením foném. Na tento typ chýb totiž predchádzajúce metódy nie sú príliš vhodné.

K definovaniu množiny prechodov, o ktorú sa pôvodná dekódovacia sieť rozšíri, využijeme fonologické pravidlá v nasledovnom tvare

$$\phi \rightarrow \psi / \lambda \_ \rho, \tag{4.8}$$

ktoré hovoria, že fonéma ϕ môže byť zamenená za fonému ψ v prípade, že sa pred ňou nachádza fonéma λ a za ňou fonéma ρ. Zavedením prázdneho symbolu ε sme schopní definovať pravidlá chýb spočívajúcich vo vkladaní foném $\varepsilon \rightarrow \psi$ alebo chýb založených na vypúšťaní foném $\phi \rightarrow \varepsilon$.

Takéto pravidlá je možné pripraviť ručne [17] na základe dobrej lingvistickej znalosti oboch jazykov, L1 aj L2. Výsledok takéhoto postupu však bude závislý na autorových zna-

lostiach a bude sa teda výrazne líšiť od autora k autorovi. Preto vhodnejší spôsob získania týchto pravidiel je pomocou automatického porovnávania prepisov L1 a L2 jazykov [24], pričom pre konštrukciu ERN sa použijú len najčastejšie pravidlá, ktoré pokrývajú určitý počet chýb. Takýto prístup dosahuje výrazne lepšiu úspešnosť pri detekcii zlej výslovnosti, avšak zaobstaranie dostatočne veľkého nenatívneho datasetu je značne náročné.

Kapitola 5

Datasety

Táto kapitola sa venuje popisu jednotlivých datasetov reči, ktoré využijeme k realizácii systému pre detekeciu nesprávnej výslovnosti. K tomu budeme primárne potrebovať nenatívny dataset reči, ktorého popis je uvedený v sekcii 5.1. Nakoľko však máme v pláne aj experimentovanie s multiligválnymi akustickými modelmi, nezaobídeme sa bez ďalších, tentokrát natívnych datasetov, ktoré sú popísané v sekciách 5.2 – 5.4.

5.1 ISLE

ISLE (Interactive Spoken Language Education) dataset [25] bol vytvorený za účelom automatického hodnotenia výslovnosti. Pozostáva z nahrávok nenatívnej angličtiny, ktoré pochádzajú od 23 talianskych a 23 nemeckých rečníkov. Každý rečník pri nahrávaní čítal krátke úryvky textov, ktoré boli zostavné tak, aby pokrývali širokú škálu bežných výslovnostných chýb. Celkový dĺžka nahrávok je 9 hodín a 27 minút.

Každá nahrávka obsahuje okrem kanonického aj skutočný fonémový prepis s vyznačenými segmentálnymi a prozodickými chybami. Anotáciu zabezpečovalo niekoľko lingvistov, ktorý sa primárne snažili o využívanie anglických fonetických symbolov (viď tabuľku 5.1). Napriek tomu bolo v niektorých prípadoch nutné použiť aj fonetické symboly z iných jazykov. Celkovo tak prepis tvorí 41 anglických foném a 8 foném, ktoré boli prevzaté z nemčiny alebo taliančiny.

Hoci bola pri vytváraní datasetu snaha o rovnomerné zastúpenie rečníkov podľa pohlavia a úrovne angličtiny, nebol tento cieľ úplne naplnený. Výsledné zastúpenie rečníkov je možné nájsť v tabuľke 5.2.

5.2 TIMIT

TIMIT [14] je korpus pozostávajúci z nahrávok angličtiny a ich fonetických prepisov. Celkovo sa jedná o 6300 nahrávok s dĺžkou 5 hodín a 24 minút, pričom jeden rečník nahovoril vždy presne 10 nahrávok. Vo všetkých prípadoch sa jedná o americkú angličtinu. Rečníci pochádzajú z 8 vybraných regiónov, pričom pre každú oblasť je typický určitý dialekt. Nahrávanie prebiehalo v tichej miestnosti s jedným typom mikrofónu na frekvencii 16 kHz a 16 bitovým rozlíšením.

Prepisy sú zostavené zo 61 rôznych foném tzv. TIMITBET abecedy. V praxi sa však pre rozpoznávanie používa redukovaná abeceda so 48 fonémami, čo bude aj náš prípad.

Symbol	IPA	Príklad	Symbol	IPA	Príklad	Symbol	IPA	Príklad
aa	ɑː	b**a**lm	oy	ɔɪ	b**oy**	dh	ð	**th**at
ae	æ	b**a**t	uh	ʊ	b**oo**k	th	θ	**th**in
ah	ʌ	b**u**t	uw	uː	b**oo**t	f	f	**f**an
ao	ɔː	b**ough**t	l	l	**l**ed	u	v	**v**an
aw	aʊ	b**ou**t	r	r	**r**ed	s	s	**s**ue
ax	ə	**a**bout	w	w	**w**ed	sh	ʃ	**sh**oe
ay	aɪ	b**i**te	y	j	**y**et	z	z	**z**oo
eh	e	b**e**t	hh	h	**h**at	zh	ʒ	mea**s**ure
er	ɜː	b**ir**d	b	b	**b**et	ch	tʃ	**ch**eap
ey	eɪ	b**ai**t	d	d	**d**ebt	jh	dʒ	**j**eep
ih	ɪ	b**i**t	g	ɡ	**g**et	m	m	**m**et
iy	iː	b**ee**t	k	k	**c**at	n	n	**n**et
oh	ɒ	b**o**x	p	p	**p**et	ng	ŋ	thi**ng**
ow	əʊ	b**oa**t	t	t	**t**at			

Tabuľka 5.1: Anglická fonémová sada použitá v datasete ISLE s odpovedajúcimi fonetickými symbolmi IPA fonetickej abecedy.

	Pohlavie		Úroveň angličtiny				Spolu
L1	M	Ž	1	2	3	4	
nemčina	13	10	-	-	8	15	23
taliančina	19	4	27	11	4	1	23
Spolu	32	14	27	11	12	16	46

Tabuľka 5.2: Zastúpenie rečníkov podľa pohlavia a ich úrovne angličtiny.

5.3 Voxforge DE

Voxforge [1] je open source projekt zameraný na získanie prepísanej reči, ktorá je potom šírená pod GPL licenciou. Na tvorbe datasetu sa podieľali dobrovoľníci z niekoľkých krajín, a v súčasnej dobe je publikovaných 17 datasetov v rozličných jazykoch. Keďže nahrávanie prebiehalo na rôznych miestach pomocou rôznych zariadení, kvalita nahrávok medzi jednotlivými rečníkmi sa značne líši. Ich správnosť bola ručne validovaná.

Nemecká verzia datasetu pozostáva z nahrávok od 322 rečníkov s celkovou dĺžkou 32 hodín a 14 minút. Z povahy získavania datasetu je možné predpokladať zastúpenie širokého množstva rôznych dialektov nemčiny. To však nie je možné overiť, nakoľko metadáta túto informáciu vo väčšine prípadov neobsahujú. Keďže prepisy sú len na úrovni slov, je nevyhnutné použitie výslovnostného slovníka. My sme použili slovník, ktorý je súčasťou nástroja CMUSphinx[1] a bol zostavený pre použitie s nemeckým Voxforge datasetom. Keďže však nie je garantované, že obsahuje záznamy o všetkých slovách v prepise, na získanie fonémové prepisu neznámych slov využijeme k tomu určený nástroj *Sequitur G2P* [4] natrénovaný nad použitým slovníkom.

5.4 Voxforge IT

Taliansky dataset Voxforge pozostáva z menšieho počtu nahrávok, ktorých dĺžka je v tomto prípade 19 hodín a 56 minút. Na jeho zostavení sa podieľalo 347 rečníkov. Ako v predchádzajúcom prípade, nahrávky sú opatrené len slovnými prepismi. K prevodu na fonémový prepis sme preto opäť využili slovník distribuovaný s nástrojom CMUSphinx[1], resp. nástroj *Sequitur G2P* na prevod slov, ktoré sa v slovníku nenachádzajú.

[1] Výslovnostné slovníky k nástroju CMUSphinx sú dostupné z https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/

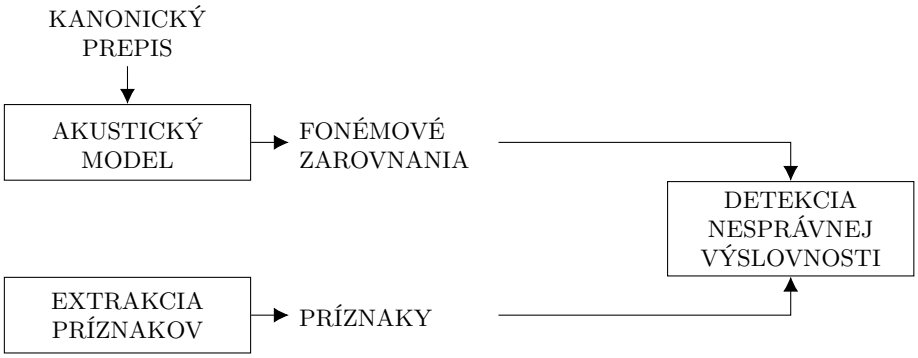
Kapitola 6

Návrh systému

V rámci tejto kapitoly popíšeme návrh systému určeného na automatickú detekciu nesprávnej výslovnosti. Systém je do veľkej miery inšpirovaný prácou [2], aby sme si mohli overiť a porovnať nami dosiahnuté výsledky. Všetky použité metódy a algoritmy boli popísané v kapitolách 2 – 4.

6.1 Časti systému

Systém pozostáva z niekoľkých častí, ktoré sú schematicky znázornené na obr. 6.1. Nakoľko v rámci experimentov budeme testovať niekoľko odlišných prístupov, schéma je do veľkej miery abstraktná, aby postihla všetky tieto varianty. Dôležitou časťou celého systému je akustický model, ktorý realizuje zarovnanie reči podľa kanonického prepisu. Ďalšou podstatnou súčasťou sú príznaky, pričom ich typ závisí na použitej metóde detekcie. V prípade GOP metódy sú nimi vierohodnosti jednotlivých HMM stavov. Získanie týchto príznakov zabezpečuje neurónová sieť, ktorá je zároveň súčasťou DNN-HMM akustického modelu. Pri hodnotení výslovnosti pomocou klasifikátora je možné použiť celú radu príznakov. V našej práci využijeme dva typy, a to jednak už spomínané vierohodnosti HMM stavov, ale taktiež pravdepodobnosti fonologických rysov určené samostatne natrénovanou neurónovou sieťou.



Obr. 6.1: Schematické znázornenie systému pre detekciu nesprávnej výslovnosti.

6.2 Akustický model

Pre modelovanie reči využijeme HMM v kombinácii s hlbokou neurónovou sieťou (*Deep Neural Network*, DNN). Akustický model v systéme bude slúžiť na získanie fonémových zarovnaní a extrakciu príznakov z nenatívnej reči. Z tohto dôvodu musí byť aj trénovaný nad nenatívnym datasetom, konkrétne nad skutočnými fonémovými prepismi.

Jednotlivé fonémy budeme reprezentovať jednak nezávisle (monofónový model), ale aj ako kontextovo závislé na susedných fonémach (trifónový model). Modely foném v oboch prípadoch pozostávajú z troch stavov, tak ako na obr. 2.2a. Rovnaký počet stavov, len s dodatočnými prechodmi (viď obr. 2.2c), bude použitý aj pri modelovaní ticha.

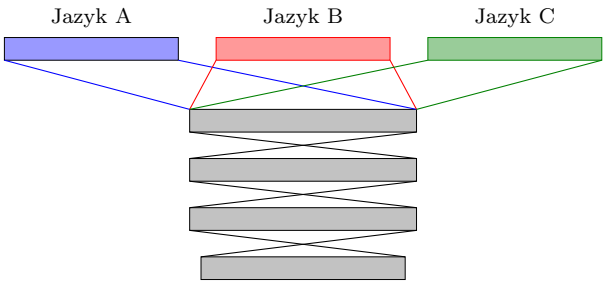
Pred samotným trénovaním DNN je potrebné zarovnať trénovacie dáta pomocou skôr natrénovaného GMM-HMM modelu. V prípade GMM sú vstupom modelu MFCC príznaky. Napriek tomu, že tieto príznaky by bolo možné použiť aj v prípade trénovania DNN, vhodnejšie je použitie filter bank (označované aj fbank) príznakov, nakoľko u nich bývajú dosahované lepšie výsledky [26]. Pre dosiahnutie dobrej konvergencie neurónovej siete je však nevyhnutná ich normalizácia na nulovú strednú hodnotu a jednotkový rozptyl. Okrem toho zahrnieme do príznakového vektora odpovedajúceho nejakému rámcu aj hodnoty pre istý počet okolitých rámcov, vďaka čomu získa neurónová sieť určitú informáciu o kontexte.

Multilingválne akustické modely

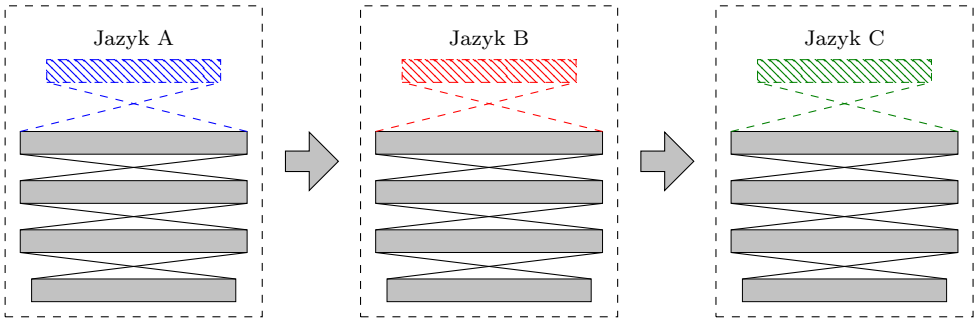
V rámci našich experimentov budeme skúmať, aký má vplyv použitie DNN-HMM akustického modelu natrénovaného na viacerých jazykoch. Presnejšie povedané, na viacerých jazykoch je trénovaná len DNN a HMM pochádza z GMM-HMM modelu natrénovaného len na príslušnom jazyku. Predpokladáme totiž, že keďže je nenatívny dataset tvorený nenatívnymi rečníkmi, mohlo by trénovanie akustického modelu na ich natívnych jazykoch prispieť k vyššej presnosti. Naviac pri trénovaní na viacerých jazykoch dochádza k lepšej generalizácii, takže takého modely dosahujú lepšie výsledky nezávisle na použitých jazykoch [15, 21].

Multilingválne trénovanie je špecifický prípad viacúlohového učenia (angl. *Multi-Task Learning*), ktorého snahou je zdieľanie parametrov medzi úlohami. V prípade hlbokých neurónových sietí má takéto učenie najčastejšie podobu zdieľania skrytých vrstiev, nad ktorými sa nachádza niekoľko vrstiev, ktoré sú špecifické pre dané úlohy. V prípade učenia nad viacerými jazykmi hovoríme o výstupných vrstvách špecifických pre každý uvažovaný jazyk, keďže fonémové sady sa medzi jazykmi obecne líšia, viď obr. 6.2. Vstupná vrstva je spoločná pre všetky jazyky, pričom pri trénovaní sa parametre neurónovej siete aktualizujú podľa chyby propagovanej zo zodpovedajúcej výstupnej vrstvy. Kľúčovým faktorom pri takejto topológii je simultánne trénovanie na všetkých jazykoch súčasne. To môže byť dosiahnuté rovnomerným výberom trénovacích vzoriek v každom kroku trénovania. V praxi však majú datasety pre jednotlivé jazyky odlišnú veľkosť. Možným riešením je napr. opätovné používanie trénovacích vzoriek u menších datasetov.

V našom prípade je ale výrazný rozdiel medzi veľkosťami datasetov, čo by mohlo viesť na výrazné zhoršenie výsledkov. Preto použijeme v našej práci odlišný postup, kedy neurónovú sieť budeme trénovať sekvenčne jazyk po jazyku, tak ako je to znázornené na obr. 6.3. V prvom kroku sa náhodne inicializovaná neurónová sieť natrénuje nad jedným z uvažovaných jazykov. Následne sa odstráni výstupná vrstva a nahradí sa novou, náhodne inicializovanou vrstvou zodpovedajúcej ďalšiemu jazyku v poradí, a model sa natrénuje na tomto jazyku. Takto sa postupuje pre všetky jazyky. Nevýhodou tohto prístupu je, že vý-



Obr. 6.2: Paralelné trénovanie na viacerých jazykoch s využitím hlbokej neurónovej siete so zdieľanou skrytou vrstvou [21].



Obr. 6.3: Postupné, sekvenčné trénovania hlbokej neurónovej siete na viacerých jazykoch [15].

sledný model je značne ovplyvnený poradím jazykov, v ktorom sa postupovalo. V našom prípade to môže byť ale aj výhodou, nakoľko výsledný model bude viacej „zaujatý“ voči nenatívnemu jazyku.

6.3 Extrakcia príznakov

V závislosti na použitej metóde detekcie nesprávnej výslovnosti budeme využívať dva druhy príznakov – vierohodnosti HMM stavov alebo pravdepodobnosti fonologických rysov.

6.3.1 Vierohodnosti HMM stavov

K získaniu vierohodností HMM stavov, využijeme DNN z akustického modelu, ktorý sme popisovali v predchádzajúcej sekcii. Výstupom neurónovej siete však nie sú vierohodnosti HMM stavov, ale ich aposteriórne pravdepodobnosti. K prevodu pravdepodobností na vierohodnosti nám však postačuje poznať apriórnu pravdepodobnosť $P(s)$ stavu s, ktorá sa určí z trénovacích dát frekvenčnou analýzou. Samotný prevod pravdepodobností $p(s|\boldsymbol{o})$ na vierohodnosti $p(\boldsymbol{o}|s)$ je daný vzťahom

$$p(\boldsymbol{o}|s) = \frac{p(s|\boldsymbol{o})}{P(s)} p(\boldsymbol{o}), \tag{6.1}$$

kde $p(\boldsymbol{o})$ je apriórna pravdepodobnosť príznakov rámca $\boldsymbol{o}$. Tú je však možné zo vzťahu vypustiť, nakoľko nie je závislá na s.

6.3.2 Pravdepodobnosti fonologických rysov

Fonologické rysy, alebo tiež označované ako dištinktívne rysy, sú najmenšou fonologickou jednotkou. Ako už názov napovedá, umožňujú nám popísať každú fonému jedinečnou sadou rysov. V minulosti sa objavili pokusy o využitie fonologických rysov aj k rozpoznávaniu reči [36], ale pri súčasných ASR systémoch je vhodnejšie priamo použiť DNN určujúcu pravdepodobnosti HMM stavov. Pri ich použití k detekcii nesprávnej výslovnosti je však výhoda, že môžu byť zároveň aplikované na poskytnutie spätnej väzby používateľovi systému, nakoľko každý rys zároveň popisuje, ako je zodpovedajúca fonéma tvorená v rečovom ústrojenstve.

Rozlišujeme niekoľko fonologických rysov v závislosti na type uvažovaného fonologického modelu. V súlade s referenčnou prácou [2] použijeme fonologický model pozostávajúci z 18 binárnych rysov. Definíciu jednotlivých rysov pre každý foném ISLE datasetu je možné nájsť v tabuľke 6.1.

Fonologické rysy	Fonémy
VOC	aa ae ah ao aw ax ay eh er ey ih iy oh ow oy uh uw
CONS	b ch d dh f g hh jh k p s sh t th v z zh l m n ng r
CONT	dh f hh l s sh th v z zh
OBSTR	b ch d dh f g jh k p hh s sh t th v z zh
STR	ch s sh th z zh
VOICE	b d dh g jh v z zh
SON	aa ae ah ao aw ax ay eh er ey ih iy l m n ng oh ow oy r uh uw w y
STOP	b ch d g jh k p t
LOW	aa ae aw ay
HIGH	ch ih iy jh sh uh uw w y zh
LAB	ao b f m oh ow oy p uh uw v w
COR	ae ch d dh eh ey ih iy jh l n r s sh t th y z zh
DOR	aa ao aw ay g k ng oh ow oy uh uw w
RTR	ah ax eh er ih uh w
NAS	m n ng
LAT	l
RHO	er r
RAD	hh

Tabuľka 6.1: Fonologické rysy s odpovedajúcimi fonémami ISLE datasetu.

Určovanie pravdepodobností fonologických rysov prebieha obdobne, ako je tomu u vierohodností HMM stavov. Rovnako ako v predchádzajúcom prípade sa k tomu využíva DNN, ktorá sa trénuje na základe zarovnaní na fonémy získanými pomocou GMM-HMM modelu. Vstup tejto neurónovej siete sú fbank príznaky a výstupom je 19 hodnôt, kde 18 zodpovedá fonologickým rysom a jedna je vyhradená pre detekciu ticha. Keďže jednému rámcu môže zodpovedať niekoľko rysov, výstupnú vrstvu tvorí softmax aktivačná funkcia.

6.4 Detekcia nesprávnej výslovnosti

Detekciu nesprávnej výslovnosti založíme na dvoch prístupoch. Prvým z nich bude hodnotenie výslovnosti pomocou metód založených na aposteriórnej pravdepodobnosti foném.

Okrem štandarného GOP skóre popísaného v kapitole 4 zavedieme niekoľko modifikácií, ktoré majú potenciál dosiahnuť lepšie výsledky. Druhým uvažovaným prístupom bude detekcia nesprávnej výslovnosti s využitím priamej klasifikácie, kde porovnáme rôzne druhy klasifikátorov v kombinácii s rôznymi príznakmi.

6.4.1 Metódy založené na aposteriórnej pravdepobnosti foném

Štandardné GOP skóre

Ako už bolo uvedené v kapitole 4, rovnica pre výpočet štandardného GOP skóre má tvar

$$\text{STD GOP}(p) = \left| \log \left(\frac{p(\boldsymbol{O}^{(p)}|p)}{\max_{q \in Q} p(\boldsymbol{O}^{(p)}|q)} \right) \right| \Big/ d, \tag{6.2}$$

kde hodnota v čitateli je určená súčinom vierohodností po jednotlivých rámcoch, ktoré sú dané núteným zarovnaním. Hodnotu v menovateli počítame obdobne, avšak v tomto prípade uvažujeme vierohodnosti zodpovedajúce foneme q, ktorú určíme pomocou fonémového rozpoznávania. Tento postup vychádza z predpokladu, že fonémovým rozpoznávaním obdržíme fonému, ktorá bola s najväčšou pravdepodobnosťou vyslovená, a to aj s ohľadom na jazykový model. Pre viac detailov viď kapitolu 4.

Ako už bolo naznačené, výpočet vierohodnosti $p(\boldsymbol{O}^{(p)}|p)$ pre celý segment $\boldsymbol{O}^{(p)}$ je daný súčinom vierohodností $p(\boldsymbol{o}_t|p)$ po jednotlivých rámcoch $\boldsymbol{o}_t$ segmentu $\boldsymbol{O}^{(p)}$, t.j.

$$p(\boldsymbol{O}^{(p)}|p) = \prod_{t=t_0}^{t_0+d-1} p(\boldsymbol{o}_t|p), \tag{6.3}$$

kde t_0 je index prvého rámca segmentu $\boldsymbol{O}^{(p)}$. Vierohodnosti $p(\boldsymbol{o}_t|p)$ však vzhľadom na použitý akustický model nie je možné získať priamo. Jedna fonéma p je totiž reprezentovaná niekoľkými HMM stavmi s, označovanými tiež ako sonémy. To znamená, že jednej foneme p zodpovedá niekoľko soném s, t.j. aj niekoľko vierohodností $p(\boldsymbol{o}_t|s)$. Preto vierohodnosť $p(\boldsymbol{o}_t|p)$ vypočítame ako sumu zodpovedajúcich vierohodností $p(\boldsymbol{o}_t|s)$, t.j.

$$p(\boldsymbol{o}_t|p) = \sum_{s \in \mathscr{P}} p(\boldsymbol{o}_t|s), \tag{6.4}$$

kde množina $\mathscr{P}$ obsahuje všetky sonémy, ktoré odpovedajú foneme p. Takéto riešenie zároveň lepšie pokrýva variabilitu medzi nenatívnymi rečníkmi, čo dokazujú aj výsledky v práci [20].

Likelihood-ratio GOP skóre

Likelihood-ratio (LR) GOP skóre [19] sa významne nelíši od štandardného GOP skóre popísaného vyššie. Jediný rozdiel spočíva v určení hodnoty v menovateli, ktorá sa nestanovuje na základe fonémového rozpoznávania. Namiesto toho sa sa zvolí maximálna hodnota vierohodnosti pre nejakú inú fonému q, ktorá je odlišná od fonémy p danej núteným zarovnaním. Vzťah pre výpočet LR GOP skóre je teda

$$\text{LR GOP}(p) = \left| \log \left(\frac{p(\boldsymbol{O}^{(p)}|p)}{\max_{q \in Q, q \neq p} p(\boldsymbol{O}^{(p)}|q)} \right) \right| \Big/ d. \tag{6.5}$$

Výhodou v tomto prípade je, že výsledné skóre nie je do takej miery ovplyvnené jazykovým modelom. V predchádzajúcom prípade bol totiž výsledok výrazne závislý na fonémovom rozpoznávaní, ktoré môže byť problematické na nenatívnej reči. Jazykový model totiž často nepostihuje širokú variabilitu, ktorá je pre nenatívnych rečníkov tak typická.

Normalizované aposteriórne pravdepodobnosti foném

Ako už bolo zmienené v kapitole 4, snahou metód založených na aposteriórnej pravdepodobnosti foném je určenie normalizovanej (dĺžkou d) aposteriórnej pravdepodobnosti, že fonéma p zodpovedá akustickému segmentu $\boldsymbol{O}^{(p)}$, čiže stanovanie hodnoty $p(p|\boldsymbol{O}^{(p)})/d$.

Keďže v našej práci využívame DNN-HMM akustický model, ponúka sa, aby sme namiesto štandardného výpočtu s využitím vierohodností použili priamo aposteriórne pravdepodobnosti určené pomocou DNN. To znamená, že aposteriórne pravdepodobnosti soném s nebudeme prevádzať na vierohodnosti spôsobom uvedeným v sekcii 6.3.1, ale využijeme ich priamo k výpočtu normalizovanej aposteriórnej pravdepodobnosti $p(p|\boldsymbol{O}^{(p)})/d$.

Analogicky ako v prípade výpočtu vierohodnosti $p(\boldsymbol{O}^{(p)}|p)$ stanovíme pravdepodobnosť $p(p|\boldsymbol{O}^{(p)})$ ako súčin pravdepodobností $p(p|\boldsymbol{o}_t)$ po jednotlivých rámcoch $\boldsymbol{o}_t$ segmentu $\boldsymbol{O}^{(p)}$, čiže aplikáciou vzťahu

$$p(p|\boldsymbol{O}^{(p)}) = \prod_{t=t_0}^{t_0+d-1} p(p|\boldsymbol{o}_t), \tag{6.6}$$

kde pravdepodobnosť $p(p|\boldsymbol{o}_t)$ je určená z pravdepodobností soném $s \in \mathscr{P}$ odpovedajúcich fonéme p ako

$$p(p|\boldsymbol{o}_t) = \sum_{s \in \mathscr{P}} p(s|\boldsymbol{o}_t). \tag{6.7}$$

Výsledné skóre označíme ako *Averaged Posteriors* (AP) skóre, a bude dané vzťahom

$$\mathrm{AP}(p) = p(p|\boldsymbol{O}^{(p)})/d. \tag{6.8}$$

6.4.2 Metódy založené na priamej klasifikácii

Metóda priamej klasifikácie je založená na klasifikátore, ktorý pre každý segment $\boldsymbol{O}^{(p)}$ odpovedajúci kanonickej fonéme p určí, s akou pravdepodobnosťou je fonéma p správne vyslovená. Jednotlivé fonémové segmenty sú získané núteným zarovnaním podľa kanonického prepisu. Budeme experimentovať s dvoma typmi klasifikátorov – doprednou neurónovou sieťou (angl. *feedforward neural network*) a LSTM neurónovou sieťou.

Ako vstupy klasifikátorov budú použité príznaky extrahované v súlade s popisom v sekcii 6.3, t.j. budú nimi vierohodnosti HMM stavov alebo pravdepodobnosti fonologických rysov. Pre dosiahnutie dobrej konvergencie pri trénovaní budú logaritmické hodnoty príznakov normalizované na nulovú strednú hodnotu a jednotkový rozptyl. V prípade doprednej neurónovej siete nie je možné na vstup privádzať príznaky pre celý segment, nakoľko majú rôznu dĺžku. Preto na vstup privedieme len ich priemerné hodnoty pre celý segment. Pri LSTM neurónovej sieti rôzna dĺžka príznakov nepredstavuje problém, čiže vstupom budú príznaky všetkých rámcov daného segmentu.

Výstupné vrstvy oboch neurónových sietí sú tvorené neurónmi so sigmoidovou aktivačnou funkciou. Každý neurón zodpovedá jednej kanonickej fonéme, pričom jeho výstupom je pravdepodobnosť, že fonéma bola správne vyslovená.

Počas trénovania je výstupná hodnota neurónu, ktorý zodpovedá foném v kanonickej transkripcii, nastavená na hodnotu 1, resp. 0, ak bola fonéma správne, resp. nesprávne vyslovená. Táto hodnota je určená porovnaním kanonického a skutočného prepisu. Ostatné výstupné neuróny zostávajú nenastavené, a chyba je určená len z daného neurónu.

Kapitola 7

Experimenty

Táto kapitola je venovaná popisu a vyhodnoteniu experimentov navrhnutých za cieľom vyhodnotenia úspešnosti systému popísaného v predchádzajúcej kapitole. Na začiatok porovnáme výsledky nášho systému s referenčným článkom, aby sme overili jeho funkčnosť. Následne vyhodnotíme rôzne prístupy hodnotenia výslovnosti vrátane ich modifikácii a vyberieme metódy s najlepšími výsledkami. Na záver nad vybranými metódami overíme, aký bude mať vplyv použitie akustického modelu trénovaného na viacerých jazykoch na výslednú úspešnosť pri hodnotení výslovnosti.

7.1 Spôsob vyhodnotenia experimentov

Pri vyhodnotení experimentov budeme uvažovať len chyby výslovnosti spočívajúce vo vypúšťaní a zámene foném. Na detekciu chýb spôsobených vložením jednej alebo viacerých foném sa totiž používajú odlišné prístupy, napr. rozšírené rozpoznávacie siete (*Extended Recognition Networks*) [3]. K určeniu druhu chyby je potrebné porovnať kanonické a skutočné prepisy.

Úspešnosť uvažovaných metód vyhodnotíme pomocou tzv. miery chybného prijatia (*False Acceptance Rate*, FAR) a miery chybného odmietnutia (*False Rejection Rate*, FRR), kde FAR udáva pomer medzi počtom zle vyslovených foném klasifikovaných ako správne vyslovené k celkovému počtu zle vyslovených foném a FRR je naopak pomer medzi počtom správne vyslovených foném klasifikovaných ako nesprávne vyslovené k celkovému počtu správne vyslovených foném. Keďže sa tieto hodnoty líšia v závislosti na použitom prahu, vykreslíme si ich vzájomnú závislosť. Takto získame graf, ktorý sa zvykne označovať ako ROC krivka. Pre číselné porovnanie metód použijeme tzv. rovnakú mieru chyby (*Equal Error Rate*, ERR), ktorá je rovná hodnote FAR, resp. FRR, keď je FAR a FRR totožné, t.j. EER := FAR = FRR.

Výsledky v prípade použitia neurónovej siete bývajú do určitej miery závislé na jej počiatočnej inicializácii pri trénovaní. Nakoľko v našom prípade inicializujeme sieť náhodne, je vhodnejšie uvádzať strednú hodnotu spolu s rozptylom určenú z niekoľkých klasifikátorov s rôznou inicializáciou. Toto nám umožní objektívnejšie porovnať rôzne metódy využívajúce neurónové siete. Vo všetkých prípadoch bude táto hodnota určená z 10 klasifikátorov s náhodnou inicializáciou. Obdobný postup použijeme pri grafoch závislosti FAR a FRR, kde určíme tieto hodnoty zo zpriemerovaných rozhodnutí od všetkých 10 klasifikátorov.

7.2 Parametre experimentov

Navrhnutý systém realizujeme v toolkite Kaldi [32], čo je open-source nástroj vyvinutý k rozpoznávaniu reči. Napriek tomu, že obsahuje aj natívnu implementáciu neurónových sietí potrebných k akustickému modelovaniu, rozhodli sa pre ich implementáciu v nástroji Keras [7]. Ten totiž podporuje jednoduchšie a flexibilnejšie modifikácie realizovanej neurónovej siete, čo využijeme v niektorých experimentoch.

Dataset

K experimentom využijeme nenatívny dataset ISLE, ktorý rozdelíme na trénovaciu a testovaciu sadu, kde pre každý L1 jazyk v datasete použijeme nahrávky od 19 rečníkov na trénovanie a 4 rečníkov na testovanie. Ďalej ale už nahrávky podľa jazykov rozlišovať nebudeme, pretože samostatne by na trénovanie nebolo dostatok dát. Rozdelenie datasetu je v súlade s referenčnou prácou [2], čo nám umožní jednoduché porovnanie dosiahnutých výsledkov. Po rozdelení majú nahrávky v trénovacej sade dĺžku 8 hodín a 24 minút a v testovacej sade 1 hodinu a 34 minút. Zoznam rečníkov v jednotlivých sadách sa nachádza v tabuľke 7.1. Pri trénovaní neurónovej siete použijeme aj validačnú sadu, ktorá pozostáva z troch rečníkov trénovacej sady – SESS0131, SESS0138 a SESS0186.

Nakoľko nebudeme pri experimentoch uvažovať chyby založené na vkladaní foném, rozšírime kanonický prepis o vložené fonémy nachádzajúce sa v skutočnom prepise. Ak by sme tento krok vynechali, nútené zarovnania na kanonický prepis by mohli viesť na segmenty pozostávajúce z rámcov odpovedajúcich viacerým fonémam.

Testovacia sada	Trénovacia sada				
SESS0006	SESS0012	SESS0183	SESS0191	SESS0126	SESS0134
SESS0011	SESS0021	SESS0184	SESS0192	SESS0127	SESS0135
SESS0015	SESS0161	SESS0185	SESS0193	SESS0128	SESS0136
SESS0020	SESS0162	SESS0186	SESS0003	SESS0129	SESS0137
SESS0041	SESS0163	SESS0187	SESS0040	SESS0130	SESS0138
SESS0121	SESS0164	SESS0188	SESS0123	SESS0131	SESS0140
SESS0122	SESS0181	SESS0189	SESS0124	SESS0132	
SESS0139	SESS0182	SESS0190	SESS0125	SESS0133	

Tabuľka 7.1: Rozdelenie ISLE datasetu na testovaciu a trénovaciu sadu, pričom údaje v tabuľke predstavujú identifikátory jednotlivých rečníkov.

Akustický model

Vstupom DNN akustického modelu je 23 fbank príznakov s kontextom ±5 rámcov. Príznaky sú pred tým normalizované, aby mali nulovú strednú hodnotu a jednotkový rozptyl. DNN pozostáva z 3 skrytých vrstiev, ktoré sú tvorené 512 neurónmi s ReLU aktivačnou funckiou. Výstupné neuróny so soft-max prenosovou funkciou určujú pravdepodobnosť stavov monofónneho, resp. trifónneho modelu. Jednotlivé modely zodpovedajú 41 kanonickým fonémam, tichu a nakoniec špeciálne zavedenej foném, do ktorej je združených 8 foném pochádzajúcich z natívneho jazyka rečníkov.

Trénovanie prebieha na fonémových zarovnaniach vzhľadom na skutočný prepis. K zarovnaniu je použitý GMM-HMM model natrénovaný na tých istých dátach. Na vstup GMM-

HMM modelu je privedených 39 MFCC+Δ+ΔΔ príznakov. DNN je trénovaná s využitím Adam optimalizátoru nad trénovacími dávkami (mini-batches) o veľkosti 256 rámcov. Ako objektívna funkcia je zvolená kategorická krížová entropia. K zabráneniu pretrénovania je použitý 10 % dropout. Miera rýchlosti trénovania (*learning rate*) je 0,001 do doby, než sa hodnota objektívnej funkcie na validačnej dátovej sade prestane zlepšovať, maximálne však po dobu 15 epoch. Potom sa s každou epochou zmenšuje o polovicu a trénovanie končí, ak sa hodnota objektívnej funkcie nezlepšila počas 10 epoch.

Určovanie fonologických rysov

Extrakciu fonologických rysov zabezpečuje DNN s totožnou topológiou a parametrami ako v prípade akustického modelu. Jediný rozdiel je vo výstupnej vrstve, ktorá má softmax aktivačnú funkciu, a tvorí ju 19 neurónov, ktoré odpovedajú fonologickým rysom a tichu. Počas trénovania sa využívajú zarovnania na kanonický prepis získané pomocou monofónového GMM-HMM modelu, ktoré sa prevádzajú na fonologické rysy pomocou tabuľky 6.1.

Priama klasifikácia výslovnosti

Na priamu klasifikáciu výslovnosti použijeme jednak jednoduchú neurónovú sieť s doprednýým šírením a taktiež rekurentnú neurónovú sieť s LSTM architektúrou. Jednoduchá neurónová sieť sa skladá z jedinej skrytej vrstvy, ktorú tvorí 512 neurónov s ReLU aktivačnou funkciou. LSTM neurónovú sieť tvorí 512 jednotiek, pričom ich architektúra je v súlade s popisom uvedeným v kapitole 3.

V závislosti na experimentoch sú na vstup neurónových sieti privádzané logaritmické hodnoty vierohodností HMM stavov, alebo pravdepodobnosti fonologických rysov, normalizované na nulovú strednú hodnotu a jednotkový rozptyl. V prípade LSTM neurónovej siete tvoria príznakový vektor príznaky určené zo všetkých rámcov uvažovaného segmentu, pričom u jednoduchej neurónovej siete je vstupom len priemerná hodnota týchto príznakov. Pre obe neurónové siete platí, že aktivačná funkcia výstupnej vrstvy je logistická sigmoida.

Trénovanie prebieha nad segmentmi získanými núteným zarovnaním pomocou GMM-HMM modelu voči kanonickému prepisu. S trénovania sú vynechané segmenty, ktoré zodpovedajú vloženým fonémam, nakoľko tento druh chýb neuvažujeme. Riadenie trénovania zabezpečuje optimalizačný algoritmus Adam s veľkosť trénovacej dávky (*mini-batch*) 256 vzoriek. Objektívnou funkciou je krížová entropia a droput je na úrovni 10 %. Miera rýchlosti trénovania je 0,001 počas prvých troch epoch, potom sa s každou epochou znižuje o polovicu. Trénovanie končí, ak sa za posledné 4 epochy nezlepšila hodnota objektívnej funkcie na validačných dátach.

7.3 Porovnanie základných metód s referenčnou prácou

Za účeľom overenia funkčnosti nami navrhnutého systému porovnáme dosiahnuté výsledky s referenčným systémom od Arora a kol. [2]. Porovnávať budeme trojicu metód postavených na monofónovom akustickom modeli, ktoré boli použité aj v referenčnej práci. Sú nimi štandardné GOP skóre (STD GOP), dopredná neurónová sieť natrénovaná nad vierohodnosťami HMM stavov (NN HMM) a nakoniec dopredná neurónová sieť rozhodojúca na základe pravdepodobností fonologických rysov (NN PFeats).

V prípade metód priamej klasifikácie sme oproti referenčnému systému zaviedli niekoľko zmien. Namiesto strednej kvadratickej chyby v pri oboch neurónových sieťach využívame

kategorickú krížovú entropiu, ktorá je na tento typ problému vhodnejšia. V dôsledku tohto kroku je možné očakávať mierne lepšie výsledky. Ďalšia zmena spočíva v inom type použitej neurónovej siete určujúcej fonologické rysy. Sieť použitá v referenčnom systéme je súčasne trénovaná na dve úlohy – určovanie fonologických rysov aj HMM stavov. Keďže naša neurónová sieť je trénovaná len na prvej z nich, dá sa predpokladať, že dosiahneme horší výsledok. Rozdiel by však nemal byť výrazný. Poslednou odlišnosťou je normalizácia príznakov (vierohodností HMM stavov aj pravdepobností fonologických rysov) na nulovú strednú hodnotu a jednotkový rozptyl. Ukázalo sa totiž, že najmä v prípade nenormovaných vierohodností neurónová sieť horšie konverguje k riešeniu, alebo dokonca pri použití trifónového modelu nekonverguje vôbec.

Dosiahnuté výsledky je možné v podobe ROC kriviek vidieť na obrázku 7.1 a ich vyjadrenie pomocou miery rovnakej chyby (EER) v tabuľke 7.2, kde sa zároveň nachádzajú aj výsledky referenčného systému. Hodnoty ukazujú, že nami zostrojený systém je funkčný, aj keď sme pri metóde využívajúcej fonologické rysy dosiahli mierne horší výsledok. To je však v súlade s hore uvedeným predpokladom o vplyve viacúlohového učenia.

Výrazne nižšiu chybu vykázal systém u GOP skóre a klasifikátora využívajúceho pravdepodobnosti HMM stavov. Pri druhej menovanej metóde bolo zlepšenie až 4,23 percentuálnych bodov. Takýto výrazný rozdiel môže byť spôsobený dvojicou faktorov, a to použitím odlišnej objektívnej funkcie a normalizovaním príznakov. Oboje totiž výrazne prispieva ku konvergencii k optimálnemu riešeniu, čo nemuselo byť prípadom referenčného systému.

Poslednou metódou, ktorú sme porovnávali, je štandardné GOP skóre. Rozdiel medzi systémami je tentokrát ešte značnejší, so zlepšením na úrovni 5,65 percentuálnych bodov. K zdôvodneniu takéhoto výsledku nám však v prípade referenčného systému chýba dostatok informácii. Predpokladáme však, že nimi implementovaný akustický model mohol byť nesprávne optimalizovaný, čo by vysvetlovalo aj výsledok u NN HMM klasifikátora.

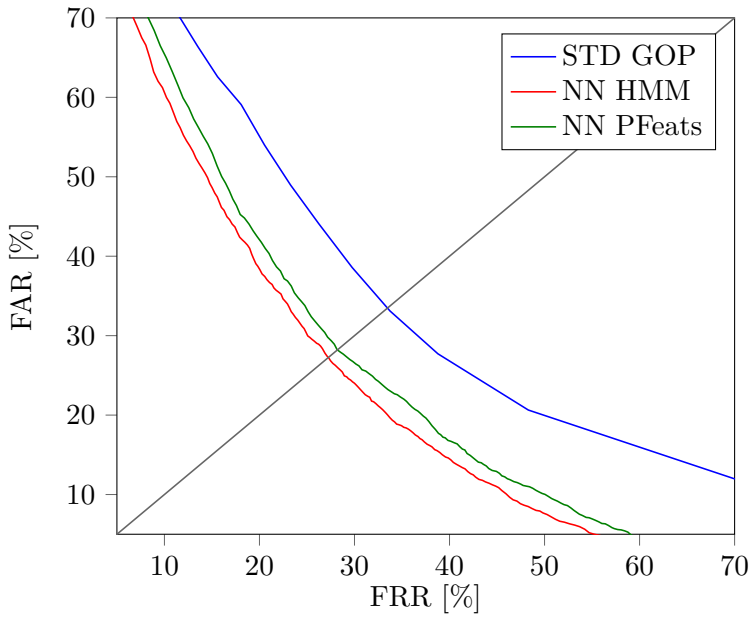
EER [%]	STD GOP	NN HMM	NN PFeats
Navrhnutý systém	**33,35**	**27,57** $\pm$ 0,16	28,49 $\pm$ 0,14
Referenčný systém [2]	39,00	31,80	**28,30**

Tabuľka 7.2: Dosiahnuté výsledky pomocou základných metód – štandardná GOP metóda (STD GOP), neurónové siete klasifikujúce na základe HMM vierohodností (NN HMM) a na základe pravdepodobností fonologických rysov (NN Pfeats).

7.4 Porovanie metód založených na aposteriórnej pravdepodobnosti foném

Táto sekcia sa venuje porovnaniu zavedených metód založených na aposteriórnej pravdepodobnosti foném – štandardného GOP skóre, likelihood ratio GOP skóre (LR GOP) a spriemerovaných aposteriórnych pravdepodobností (AP). Okrem toho sme pri každom uvedenom skúmali, aký vplyv má využitie trifónového akustického modelu na výslednú úspešnosť.

Ako je možné vidieť na obrázku 7.2, nezávisle na akustickom modeli dosiahla najlepšie výsledky LR GOP metóda, a to v celom skúmanom intervale, t.j. pre ľubovoľnú hodnotu prahu. Naopak najnižšia úspešnosť bola nameraná pri štandardnom GOP skóre. Pri porov-



Obr. 7.1: Graf závislosti FAR a FRR pre základné metódy – štandardná GOP metóda (STD GOP), neurónové siete klasifikujúce na základe HMM vierohodností (NN HMM) a pravdepodobností fonologických rysov (NN Pfeats).

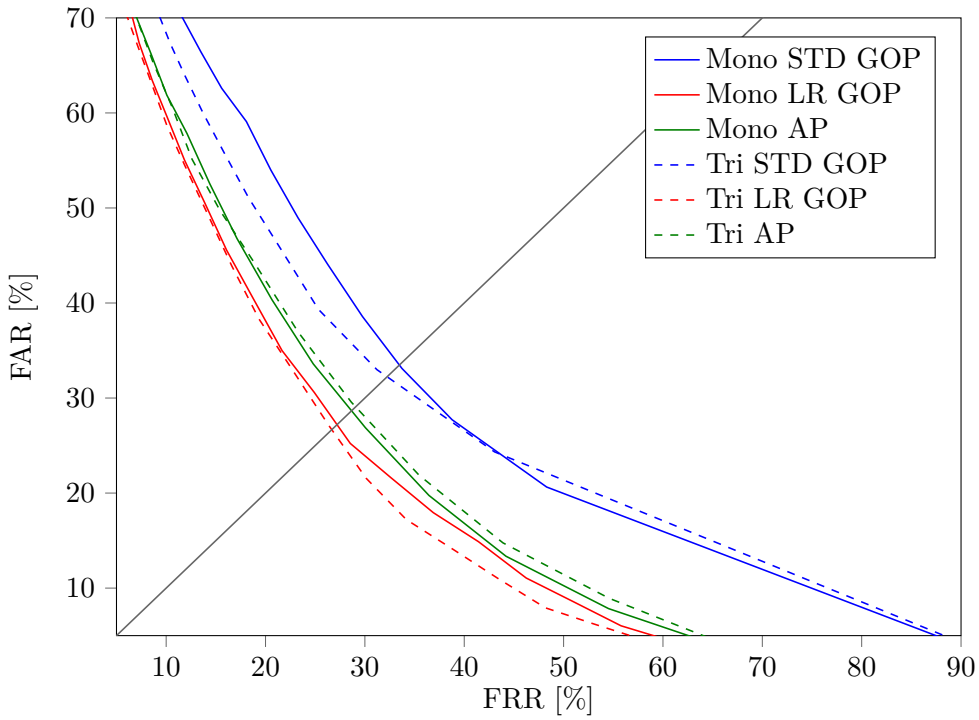
naní EER týchto dvoch metód, viď tab. 7.3, sa výsledok u monofónového modelu líšil o 6,12 percentuálnych bodov, čo je až 18,35 % relatívne zlepšenie pre LR GOP.

Domnievame sa, že za horší výsledok u štandardného GOP môže fonémové rozpoznávanie, ktoré je používané len pri tejto metóde. Výsledok rozpoznávania je totiž závislý na kvalite jazykového modelu, ktorý nemusí dostatočne pokrývať variabilitu nenatívnej reči, a to aj napriek tomu, že je nad nenatívnym prepisom trénovaný. Zlepšenie jazykového modelu by teda mohlo viesť v konečnom dôsledku k priaznivejším výsledkom, to však nie je náplňou tejto práce.

Použitie trifónového akustického modelu neviedlo k výrazne odlišej chybe. Pri porovnaní EER bolo najväčšie zlepšenie pozorované pri štandardnom GOP skóre na úrovni 1,65 percentuálnych bodov, čo je 4,8 % relatívne zlepšenie. Pri pohľade na graf však vidieť, že to neplatí pre celý interval, a pri hodnote FRR väčšej ako 40 % je výsledok dokonca horší. Nižšiu chybu v celom intervale dosiahla len LR GOP metóda, kde sa však EER znížila len o 0,32 percentuálneho bodu, čiže o 1,1 %.

EER [%]	STD GOP	LR GOP	AP
Mono AM	33,35	**27,23**	28,60
Tri AM	31,74	**26,91**	29,00

Tabuľka 7.3: Výsledky dosiahnuté pomocou štandardnej GOP metódy (STD GOP), modifikácie založenej na počítaní pomeru vierohodností (LR GOP) a spriemerovanými aposteriórnymi pravdepodobnostiami (AP).



Obr. 7.2: Graf závislosti FAR a FRR pre metódy založené na aposteriórnej pravdepodobnosti foném – štandardné GOP (STD GOP), likelihood-ratio GOP (LR GOP) a spriemerované aposteriórne pravdepodobnosti (AP).

7.5 Porovnanie metód založených na priamej klasifikácii

Obdobným spôsobom porovnáme rôzne metódy založené na priamej klasifikácii. Okrem dopredných neurónových sietí využívajúcich ako príznaky vierohodnosti HMM stavov (NN HMM) a pravdepobnosti fonologických rysov (NN PFeats) sa zameriame aj na použitie LSTM neurónových sietí nad rovnakými príznakmi (LSTM HMM, resp. LSTM PFeats). Pri klasifikátoroch pracujúcich s fonologickými rysmi využijeme len monofónový akustický model, nakoľko v tomto prípade slúži len na získanie nútených zarovnaní, ktoré sú pri trifónovom modeli takmer totožné.

Pri klasifikácii pomocou LSTM je možné očakávať mierne lepšie výsledky, nakoľko takáto neurónová sieť pracuje so všetkými rámcami daného segmentu. V prípade jednoduchej neurónovej siete je totiž tieto príznaky nutné spriemerovať, čím môže dôjsť k strate dôležitej informácie o kontexte.

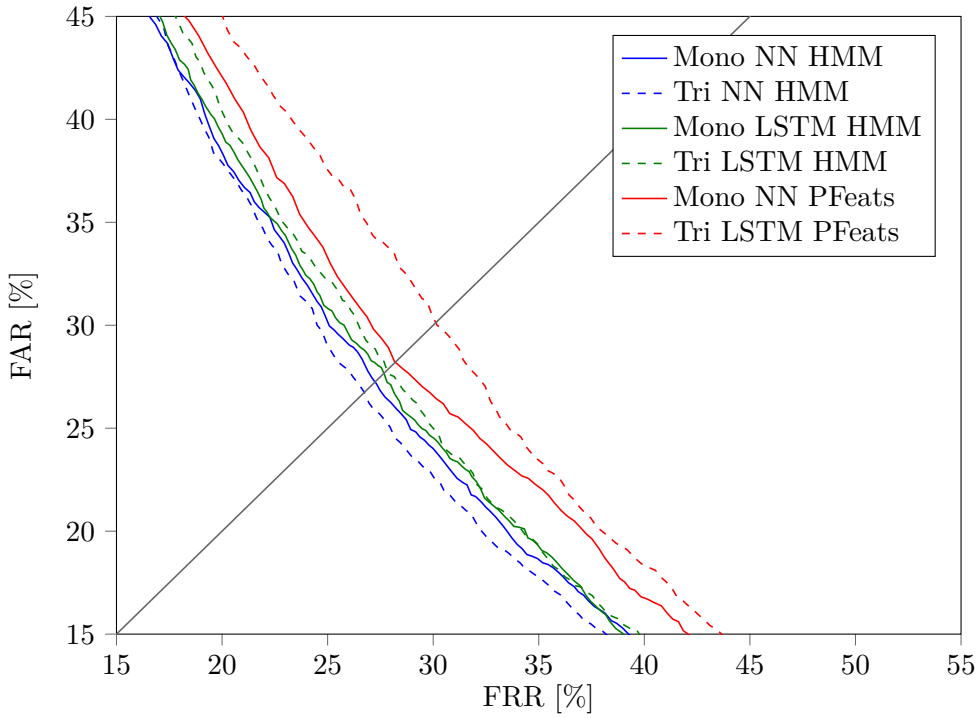
Už pri prvom pohľade na graf 7.3 je vidieť, že klasifikátory pracujúce s vierohodnosťami HMM stavov dosahujú nezanedbateľne lepšie výsledky ako je tomu pri klasifikátoroch založených na pravdepodobnostiach fonologických rysov. EER medzi NN HMM a NN PFeats metódami sa líši o 0,92 percentuálnych bodov, čo predstavuje zlepšenie o 3,2 % pri NN HMM klasifikátore. Takýto záver však nie je veľkým prekvapením. V prvom rade totiž vierohodností HMM stavov majú značne väčšiu dimenziu (150 vs. 19), takže sa dá očakávať, že nesú aj viacej informácie. Okrem toho sa môže veľké množstvo informácie stratiť už pri trénovaní neurónovej siete určujúcej fonologické rysy. Pri ňom sa totiž fonologické rysy stanovujú pomocou tabuľky prevodom z fonémového prepisu. Takýto prevod ale môže zaniesť do trénovania veľa chýb daných variabilitou reči. Napokon takýto prevod je len lineárnou

funkciou, ktorú sa môže naučiť aj samotná neurónová sieť stanovujúca HMM stavy. To dokazuje aj práca od Nagamine a kol. [29], ktorá vizualizáciou ukázala, že v skrytých vrstvách DNN akustického modelu sa formujú skupiny neurónov, ktoré pripomínajú fonologické rysy.

Očakávané zlepšenie sme pri použití LSTM nedosiahli. Pri oboch druhoch príznakov tento typ klasifikátora vykázal horšie výsledky ako jednoduchá neurónová sieť. Najmä v prípade fonologických rysov bola EER o 1,8 percentuálnych bodov vyššia než pri jednoduchej neurónovej sieti, čo znamená relatívne zhoršenie o 6,32 %. Príčinou je zrejme nepomer v počte parametrov LSTM modelu a veľkosti trénovacej sady. V porovnaní s jednoduchou neurónovou sieťou je totiž počet parametrov násobne vyšší (406 058 vs. 44 074 pri monofónovom AM, 2 215 466 vs. 496 426 pri trifónovom AM).

EER [%]	NN HMM	LSTM HMM	NN PFeats	LSTM PFeats
Mono AM	**27,57** $\pm$ 0,16	27,99 $\pm$ 0,24	28,49 $\pm$ 0,14	30,29 $\pm$ 0,10
Tri AM	**27,06** $\pm$ 0,18	28,00 $\pm$ 0,12	-	-

Tabuľka 7.4: Výsledky dosiahnuté metódami založenými na priamej klasifikácii pomocou neurónových sietí (NN), resp. LSTM neurónových sietí, ktoré boli trénované na vierohodnostiach HMM stavov (NN HMM, resp. LSTM HMM), a na pravdepodobnostiach fonologických rysov (NN PFeats, resp. LSTM PFeats).



Obr. 7.3: Graf závislosti FAR a FRR pre metódy založené na priamej klasifikáci pomocou neurónových sietí (NN), resp. LSTM neurónových sietí, ktoré boli trénované na vierohodnostiach HMM stavov (NN HMM, resp. LSTM HMM), a na pravdepodobnostiach fonologických rysov (NN PFeats, resp. LSTM PFeats).

7.6 Porovnanie GOP skóre a priamej klasifikácie

Na záver si porovnáme najúspešnejšie metódy z oboch prístupov, a to likelihood ratio (LR) GOP skóre a doprednú neurónovú sieť s HMM príznakmi na vstupe (NN HMM).

Výsledky pre obe metódy je možné vidieť v tabuľke 7.5 a obrázku 7.4. Hodnoty EER sa pri jednotlivých metódach líšia len nepatrne, s rozdielom 0,34, resp. 0,15, percentuálnych bodov u monofónového, resp. trifónového, modelu. Najmä druhá hodnota je dokonca nižšia, ako je štandardná odchýlka pri NN HMM. Pohľadom na ROC krivku zároveň zistíme, že žiadna z metód nedosahuje lepšie výsledky v celom intervale.

Tieto výsledky naznačujú, že použitím klasifikátora nie sme schopný zužitkovať z vierohodností HMM stavov viacej informácie, ako pri výrazne jednoduchšej LR GOP metóde, ktorá na detekciu nesprávnej výslovnosti využíva vierohodnosti odpovedajúce len 2 fonémam. Toto je však pozitívne zistenie, a to hneď z niekoľkých dôvodov. Prvým z nich je nízka výpočetná náročnosť LR GOP skóre, nakoľko celá detekcia spočíva len vo výpočte jednoduchého vzťahu, ktorý ani nevyžaduje fonémové rozpoznávanie, ako v prípade štandardného GOP. Okrem toho ale táto metóda striktne nevyžaduje vierohodnosti určené nenatívnym akustickým modelom. Namiesto neho postačuje model natrénovaný nad natívnym datasetom, ktorý nie je tak náročný na obstaranie. Pri jeho použití však môže drasticky klesnúť úspešnosť, preto je skôr vhodnejšie použitie kombinácie oboch datasetov, kde sa teraz už výrazne menší nenatívny dataset využije len k adaptácii.

EER [%]	LR GOP	NN HMM
Mono AM	**27,23**	$27{,}57 \pm 0{,}16$
Tri AM	**26,91**	$27{,}06 \pm 0{,}18$

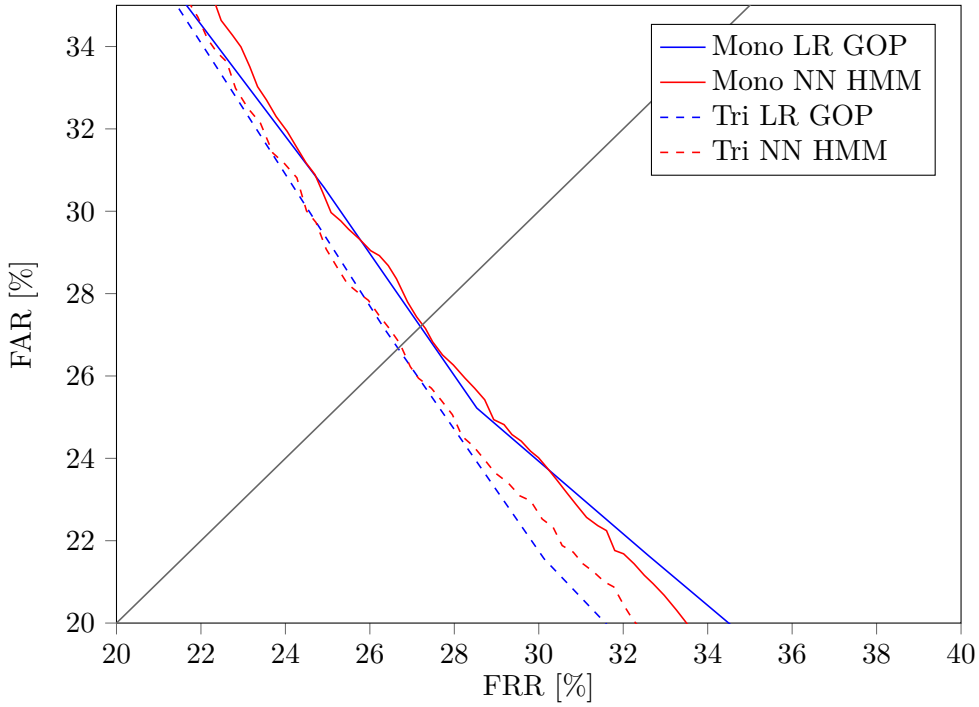
Tabuľka 7.5: Výsledky dosiahnuté najúspešnejšími metódami z každej kategórie – likelihood-ratio GOP (LR GOP) a jednoduchá neurónová sieť s HMM príznakmi (NN HMM).

7.7 Multilingválne akustické modely

Pri metódach, ktoré používajú k detekcii nesprávnej výslovnosti vierohodnosti HMM stavov, sa kvalita akustického modelu priamo odráža na výslednej úspešnosti. Ako sme už spomenuli v predchádzaúcej kapitole, jednou z možností zlepšenia akustického modelu je multilingválne trénovanie. S ohľadom na to, že nenatívny dataset pozostáva z rečníkov s nemeckým a talianským materinským jazykom, použijeme k trénovaniu multilingválneho akustického modelu celkovo 4 datasety – anglický (EN), nemecký (DE), taliansky (IT) a na koniec samotný dataset nenatívnej angličtiny (NN-EN).

Multilingválny akustický model zkonštruujeme pomocou sekvenčného spôsobu trénovania, ktorý bol popísaný v predchádzajúcej kapitole. Nakoľko má poradie jazykov pri tomto prístupe významný vplyv na kvalitu výsledného modelu, vyhodnotíme niekoľko modelov natrénovaných v rôznom poradí. Vo všetkých scenároch bude vždy použitá ako posledná nenatívna angličtina, nakoľko na nej bude prebiehať detekcia nesprávnej výslovnosti.

Topológia DNN a parametre trénovania budú totožné s monolingválnym akustickým modelom. Pre získanie zarovnaní sú použité monolingválne GMM-HMM modely. Uvažovať zároveň budeme monofónové aj trifónové akustické modely.



Obr. 7.4: Graf závislosti FAR a FRR pre najúspešnejšie metódy z každej kategórie – likelihood-ratio GOP (LR GOP) a jednoduchej neurónovej siete (NN HMM).

7.7.1 Porovnanie modelov na základe chyby pri rozpoznávaní

Ešte pred použitím zostrojených multilingválnych akustických modelov na detekciu nesprávnej výslovnosti vykonáme ich porovnanie vyhodnotením chyby rozpoznávania na úrovni foném, ktorá sa označuje aj ako *Phone Error Rate* (PER), ktorá je daná ako

$$\mathrm{PER} = \frac{S + D + I}{N_T}, \tag{7.1}$$

kde N_T je celkový počet foném v prepise a S, D a I sú počty chýb spôsobených substitúciou, vypustením alebo vložením foném. Tie získame porovnaním rozpoznaných foném so skutočným prepisom nenatívneho datasetu.

Výsledky pre jednotlivé akustické modely sa nachádzajú v tabuľke 7.6. V prípade monofónových aj trifónových akustických modelov dosiahli najnižšiu chybu modely trénované na sekvencii jazykov EN → DE → IT → NN-EN. Pri porovovaní s monolingválnymi akustickými modelmi ide o zlepšenie na úrovni 1,7 percentuálnych bodov, resp. 3 percentuálnych bodov u trifónového modelu.

V súlade s očakávaniami je chyba pri použití trifónových akustických modelov výrazne nižšia. Za povšimnutie stojí, že už len trénovanie na ľubovoľnej dvojici jazykov viedlo k podstatnému zlepšeniu pri rozpoznávaní reči.

7.7.2 Vplyv na detekeciu nesprávnej výslovnosti

V rámci tejto sekcie vyhodnotíme vplyv, ktorý má použitie multilingválneho akustického modelu na detekciu nesprávnej výslovnosti. Za týmto účeľom použijeme metódy, ktoré dosiahli v predchádzajúcich experimentoch najlepšie výsledky, t.j. likelihood-ratio (LR) GOP

Jazyky	Mono PER (%)	Tri1 PER (%)
NN-EN	42,4	35,7
EN → NN-EN	41,8	34,9
DE → NN-EN	41,3	34,1
IT → NN-EN	41,7	33,5
EN → DE → NN-EN	41,6	33,7
EN → IT → NN-EN	41,2	33,7
EN → DE → IT → NN-EN	**40,7**	**32,7**
EN → IT → DE → NN-EN	40,9	33,5

Tabuľka 7.6: Výsledky rozpoznávania pri použití akustického modelu trénovanom len na nenatívnej reči v porovnaní s multilingválnymi akustickými modelmi.

a doprednú neurónovú sieť s HMM príznakmi na vstupe (NN HMM). Vyhodnocovať budeme vplyv monofónového aj trifónového multilingválneho akustického modelu. Porovnanie vykonáme s referenčnými výsledkami získanými pri monolingválnom akustickom modeli, ktorý však bol pri tomto experimente reinicializovaný, takže aj jeho výsledky sa mierne líšia.

Ako vidieť v tabuľke 7.7, použitie multilingválneho AM viedlo vždy aspoň k nepatrnému zlepšeniu EER. Najväčšie zlepšenie sme zaznamenali pri monofónovom NN HMM, ktorého najlepší výsledok je o 5,11 % lepší než pri monolingválnom AM. Celkovo najnižšiu EER, 25,78 % dosiahla monofónová LR GOP metóda, čo predstavuje zlepšenie o 4,2 % oproti monolingválnemu systému.

EER [%]	Mono LR GOP	Tri LR GOP	Mono NN HMM	Tri NN HMM
NN-EN	27,75	26,91	27,57 ± 0,13	27,58 ± 0,22
EN → NN-EN	27,27	26,35	27,32 ± 0,10	27,35 ± 0,29
DE → NN-EN	**26,55**	26,63	27,10 ± 0,07	27,42 ± 0,19
IT → NN-EN	26,71	26,35	**26,16** ± 0,16	27,06 ± 0,13
EN → DE → NN-EN	**26,55**	26,43	26,56 ± 0,13	26,98 ± 0,21
EN → IT → NN-EN	26,83	26,47	26,77 ± 0,15	27,26 ± 0,04
EN → DE → IT →NN-EN	**26,55**	**25,78**	27,16 ± 0,11	**26,76** ± 0,19
EN → IT → DE →NN-EN	26,67	26,31	26,64 ± 0,09	26,89 ± 0,25

Tabuľka 7.7: Porovnanie výsledkov pri použití monolingválneho vs. multilingválneho akustického modelu.

7.8 Zhrnutie výsledkov

V tejto kapitole sme vykonali celú radu experimentov s cieľom nájsť najvhodnejšiu metódu na detekciu nesprávnej výslovnosti. Skúmali sme dve kategórie techník, ktoré sa za týmto účeľom používajú. Podarilo sa nám takto určiť dvojicu metód, LR GOP a NN HMM, ktoré dosahujú najlepších, navzájom podobných, výsledkov.

Nad týmito vybranými metódami sme sa ďalej pokúšali o ich zlepšenie využitím multilingválneho trénovania akustických modelov. V prípade najlepšieho výsledku sa nám podarilo dosiahnúť zníženie EER v porovnaní s referenčnou prácou o 2,52 percentuálnych bodov, čo predstavuje 8,9 % zlepšenie. Dosahujeme toho pri tom LR GOP metódou, ktorá je výrazne jednoduchšia na implementáciu, než metóda dosahujúca najlepší výsledok v referenčnom systéme. Výsledná chyba sa nám však javí stále vysoká, preto sa v ďalšej kapitole budeme venovať bližšej analýze dosiahnutých výsledkov.

Kapitola 8

Analýza výsledkov

Táto kapitola je zameraná na detailnejšiu analýzu doposiaľ dosiahnutých výsledkov. Našim cieľom je zdôvodniť relatívne vysokú chybu (EER = 25,78 %) ktorú dosahujú nami navrhnuté prístupy, a zistiť, či je v tejto oblasti priestor na potenciálne zlepšenie.

Analýzu vykonáme nad hodnoteniami výslovnosti určenými vybranou metódou, ktorej EER bola spomedzi ostatných najnižšia. Aby sme lepšie pochopili, čo je dôvodom nesprávneho hodnotenia pri tejto metóde, vyhodnotíme výsledky podľa jednotlivých foném, pri ktorých rečníci najčastejšie chybovali. Na záver podrobíme analýze aj samotný dataset nenatívnej reči, ktorý môže mať významný podiel na výslednej chybe.

8.1 Vyhodnotenie výsledkov podľa foném

Chyby u nenatívnych rečníkov sa pri niektorých kanonických fonémach vyskytujú častejšie ako u iných. Zároveň pri rôznych fonémach môžeme pozorovať chyby odlišného charakteru. Je pre to namieste, aby sme sa aj pri vyhodnocovaní detekcie nesprávnej výslovnosti zamerali na jednotlivé fonémy.

K tomuto experimentu využijeme LR GOP metódu využívajúcu vierohodnosti HMM stavov, ktoré sú určené multilingválnym (EN $\rightarrow$ DE $\rightarrow$ IT $\rightarrow$ NN-EN) trifónovým akustickým modelom. Rovnako ako v predchádzajúcej kapitole budeme vyhodnocovať úspešnosť pomocou EER a ROC krivky.

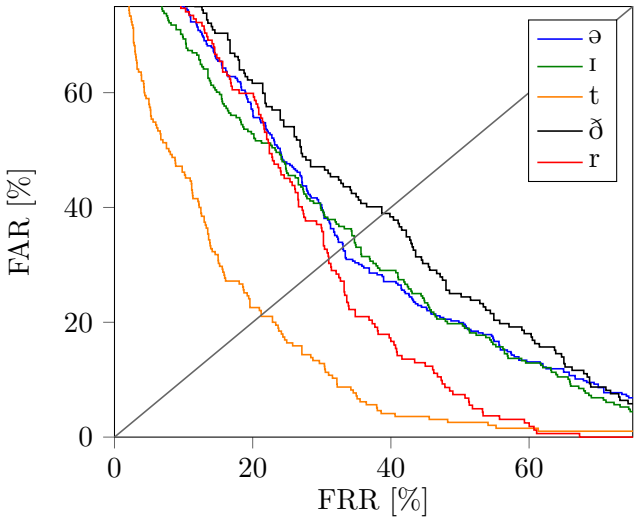
Výsledky pre päť foném, pri ktorých rečníci v testovacej sade najviac chybovali, sú uvedené v tabuľke 8.1 a obrázku 8.1. Už na prvý pohľad je zrejmé, že úspešnosť pri jednotlivých fonémach je výrazne odlišná. Pri foném /ə/ je EER na úrovni 34,52 %, čo je rozdiel 14,01 percentuálnych bodov oproti foném /t/. Tento výsledok je dosť prekvapivý, nakoľko bol v tomto prípade klasifikátor trénovaný na väčšom množstve dát ako pri foném /t/. Príčin takto nízkej úspešnosti môže byť celá rada. My sa domnievame, že najvýznamnejším faktorom môže byť široká variabilita vo výslovnosti fonémy /ə/ oproti foném /t/. Tá totiž môže spôsovať problematické rozoznanie nesprávnej výslovnosti od správnej. A to nie len klasifikátorom, ale aj samotnými anotátormi, ktorý zabezpečovali prepis nenatívnej reči. Preto sa v ďalšej sekcii zameriame na analýzu anotácii v nenatívnom datasete.

8.2 Konzistencia anotácii

Konzistencia anotácii je významným faktorom, ktorý vypovedá o kvalite datasetu. S pribúdajúcim množstvom detailov, ktoré sú anotované, dochádza aj k väčšiemu vzniku chýb. To

Fonéma	Výskyt [%]		EER [%]
	Testovacia sada	Trénovacia sada	
ə	13,52	16,88	34,52
ɪ	9,98	12,79	34,27
t	7,84	4,30	20,51
ð	6,92	6,53	37,21
r	6,52	3,58	32,10

Tabuľka 8.1: Výsledky dosiahnuté pre päť najviac chybovaných foném v testovacej sade pri použití LR GOP metódy s najlepšími výsledkami.



Obr. 8.1: Graf závislosti FAR a FRR pre päť najviac chybovaných foném v testovacej sade pri použití LR GOP metódy s najlepšími výsledkami.

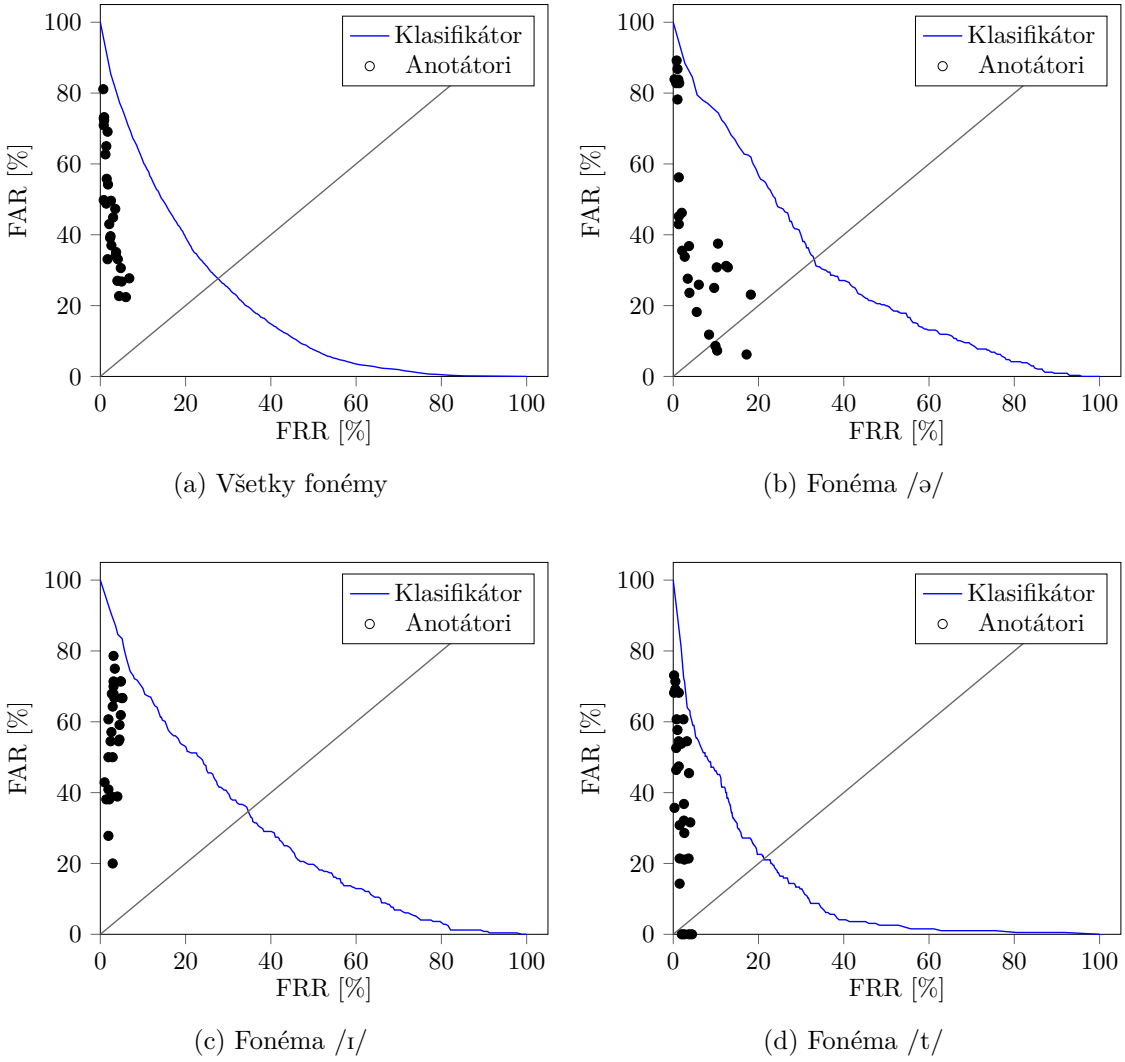
je aj prípadom fonémového prepisu, kde sa malé odlišnosti medzi fonémami stávajú pre ich krátke trvanie pre človeka ťažko rozoznateľné. Tento problém sa o to viac prejavuje u nenatívnej reči, kde je vysoká variabilita vo výslovnosti foném. Na takejto úrovni je preto ťažké stanoviť, či je fonéma ešte správne vyslovená alebo nie. Nekonzistencia sa potom objavuje nielen medzi rôznymi anotátormi, ale aj pri prepisoch od toho istého anotátora.

Za účelom overenia konzistencie ISLE datasetu je časť nahrávok opatrená prepismi od viacerých anotátorov. Keďže všetky prepisy sa do určitej miery líšia, nedokážeme rozhodnúť, aká je referečná anotácia danej nahrávky. Preto budeme uvažovať vždy dvojicu rôznych prepisov, kde prvý z nich budeme považovať za referenčný a druhý budeme voči nemu porovnávať. Takýmto spôsobom sme schopný určiť všetky druhy chýb ako v prípade klasifikátora, a teda aj vypočítať hodnoty FAR a FRR pre jednotlivých anotátorov. Pre porovnanie s výsledkami automatického hodnotenia zanesieme tieto hodnoty do grafu s ROC krivkou.

Chybu sme vyhodnocovali jednak pre všetky fonémy spoločne, ako aj po jednotlivých fonémach. Celkové výsledky spolu s tromi najfrekventovanejšími fonémami testovacej sady sa nachádzajú na obrázku 8.2. Ako sme očakávali, konzistentné anotácie na fonémovej úrovni je prakticky nemožné dosiahnuť. Určené hodnoty FAR a FRR majú skôr informatívny charakter a nebudeme sa pokúšať o ich konkrétnejšiu interpretáciu. Už aj bez toho nám však

mnohé napovedajú. Najmä pri porovnaní chýb u foném /ə/ a /t/ vidíme podobný trend, ako je tomu pri automatických hodnoteniach, kde je taktiež chyba u fonémy /ə/ vyššia ako u /t/. Je teda dosť pravdepodobné, že vyššia chyba u automatického hodnotenia je spôsobená práve nekonzistentnými anotáciami. Tie sa totiž na výsledku odrazia hneď niekoľkokrát – pri trénovaní akustického modelu, potom pri trénovaní klasifikátora výslovnosti a na záver aj pri samotnom vyhodnotení.

K totožnému záveru sa však nedá dospieť pri foném /ɪ/, pri ktorej má nekonzistencia podobný charakter ako pri foném /t/, aj tak klasifikátor dosahuje takmer rovnakú chybu, ako pri foném /ə/. Nekonzistencia anotácii teda zďaleka nebude jediným faktorom, ktorý ovplyvňuje výslednú úspešnosť. Napriek tomu však môžeme konštatovať, že jej vplyv je významný.



(a) Všetky fonémy

(b) Fonéma /ə/

(c) Fonéma /ɪ/

(d) Fonéma /t/

Obr. 8.2: FAR a FRR určené po dvojiciach anotátorov, kde anotácie jedného anotátora v dvojici sú považované za referenčné. Pre porovnanie je taktiež uvedená krivka LR GOP metódy s najlepšími výsledkami.

Kapitola 9

Záver

V tejto diplomovej práci sme sa venovali automatickému hodnoteniu výslovnosti na segmentálnej úrovni a to najmä z pohľadu detekcie chýb. Bližšie sme sa zamerali na dva prístupy hodnotenia výslovnosti, konkrétne na metódy založené na aposteriórnej pravdepodobnosti foném a metódy priamej klasifikácie pomocou neurónových sietí.

Na základe vykonaných experimentov sme ukázali, že nami navrhnutý systém dosahuje nad použitým nenatívnym datasetom výrazne lepšie výsledky ako referenčná práca, z ktorej sme vychádzali. Pri vybraných metódach sme sa ďalej pokúšali o zlepšenie navrhnutého systému multilingválnym trénovaním akustického modelu na niekoľkých rôznych jazykoch. Tento prístup priniesol taktiež významné zlepšenie, avšak výsledná chyba sa nám javila stále pomerne vysoká.

Preto sme na záver práce vykonali bližšiu analýzu výsledkov u vybranej metódy. Týmto postupom sme zistili, že výsledky automatického hodnotenia sa výrazne líšia v závislosti na hodnotenej foném. To nás viedlo aj na analýzu použitého nenatívneho datasetu, kde sme na jeho podčasti skúmali odlišnosti medzi prepismi pochádzajúcimi od rôznych anotátorov. Takto sme dospeli k záveru, že dataset trpí v určitých prípadoch významnou nekonzistenciou, ktorá má nezanedbateľný dopad na dosiahnuté výsledky. Zároveň sme však pri jednej z foném pozorovali vysokú chybu pri automatickýck hodnoteniach napriek tomu, že anotácie boli pri nej dostatočne konzistenté. Dá sa teda usudzovať, že v tejto oblasti je priestor na ďalšie zlepšenie.

V rámci ďalšej práce preto navrhujeme bližšie preskúmať charakter jednotlivých chýb v nenatívnom datasete. Okrem toho by bolo taktiež vhodné overiť dosiahnuté výsledky aj na nejakom inom nenatívnom datasete s vyššou konzistenciou anotácii.

Literatúra

[1] Voxforge.org, Free Speech Recognition. http://www.voxforge.org/, [online, cit. 29.7.2019].

[2] Arora, V.; Lahiri, A.; Reetz, H.: Phonological Feature Based Mispronunciation Detection and Diagnosis using Multi-Task DNNs and Active Learning. In *INTERSPEECH*, 2017.

[3] Arora, V.; Lahiri, A.; Reetz, H.: Phonological feature-based speech recognition system for pronunciation training in non-native language learning. 2018, s. 98–108.

[4] Bisani, M.; Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, ročník 50, č. 5, 2008: s. 434 – 451.

[5] Bishop, C.: *Pattern Recognition and Machine Learning*. Berlin, Heidelberg: Springer-Verlag, 2006, ISBN 0387310738.

[6] Celce-Murcia, M.; Brinton, D.; Goodwin, J.: *Teaching Pronunciation: A Reference for Teachers of English to Speakers of Other Languages*. Cambridge: Cambridge University Press, 1996.

[7] Chollet, F.; aj.: Keras. https://keras.io, 2015.

[8] Davis, S.; Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ročník 28, č. 4, 1980: s. 357–366.

[9] Dlaska, A.; Krekeler, C.: Self-assessment of Pronunciation. *System*, ročník 36, č. 4, 2008: s. 506 – 516.

[10] Doremalen, J.; Cucchiarini, C.; Strik, H.: Automatic Detection of Vowel Pronunciation Errors using Multiple Information Sources. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, s. 580–585.

[11] Franco, H.; Neumeyer, L.; Digalakis, V.; aj.: Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communication*, ročník 30, č. 2, 2000: s. 121 – 130.

[12] Franco, H.; Neumeyer, L.; Ramos, M.; aj.: Automatic Detection of Phone-level Mispronunciation for Language Learning. In *Learning, Proc. of Eurospeech 99*, 1999.

[13] Gales, M.; Young, S.: The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, ročník 1, 2007: s. 195–304.

[14] Garofolo, J.; Lamel, L.; Fisher, W.; aj.: TIMIT Acoustic-phonetic Continuous Speech Corpus. *Linguistic Data Consortium*, 1992.

[15] Ghoshal, A.; Swietojanski, P.; Renals, S.: Multilingual training of deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, s. 7319–7323.

[16] Glorot, X.; Bordes, A.; Bengio, Y.: Deep Sparse Rectifier Neural Networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ročník 15, Fort Lauderdale, FL, USA: PMLR, 2011, s. 315–323.

[17] Harrison, A. M.; Lo, W. K.; Qian, X.; aj.: Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-assisted Pronunciation Training. In *SLaTE*, 2009.

[18] Hermansky, H.: Perceptual Linear Predictive (PLP) Analysis of Speech. *The Journal of the Acoustical Society of America*, ročník 87, č. 4, 1990: s. 1738 – 1752.

[19] Hu, W.; Qian, Y.; Soong, F. K.: A New DNN-based High Quality Pronunciation Evaluation for Computer-aided Language Learning (CALL). In *INTERSPEECH*, 2013.

[20] Hu, W.; Qian, Y.; Soong, F. K.: An Improved DNN-based Approach to Mispronunciation Detection and Diagnosis of L2 Learners' Speech. In *SLaTE*, 2015.

[21] Huang, J.; Li, J.; Yu, D.; aj.: Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, s. 7304–7308.

[22] Jiang, H.: Confidence Measures for Speech Recognition: A Survey. *Speech Communication*, ročník 45, č. 4, 2005: s. 455 – 470.

[23] Kim, Y.; Franco, H.; Neumeyer, L.: Autoamtic Pronunciation Scoring of Specific Phone Segments for Language Instruction. In *EUROSPEECH 97*, Rhodes, Greece, 1997.

[24] Lo, W. K.; Zhang, S.; Meng, H. M.: Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System. In *INTERSPEECH*, 2010.

[25] Menzel, W.; Atwell, E.; Bonaventura, P.; aj.: The ISLE corpus of non-native spoken English. 2000.

[26] Mohamed, A.; Hinton, G. E.; Penn, G.: Understanding how Deep Belief Networks perform Acoustic Modelling. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012: s. 4273–4276.

[27] Mohri, M.; Pereira, F.; Riley, M.: *Speech Recognition with Weighted Finite-State Transducers*. Berlin, Heidelberg: Springer, 2008, s. 559 – 584.

[28] Moyer, A.: *Foreign Accent: The Phenomenom of Non-native Speech*. Cambridge: Cambridge University Press, 2013.

[29] Nagamine, T.; Seltzer, M. L.; Mesgarani, N.: Exploring how deep neural networks form phonemic categories. In *INTERSPEECH*, 2015.

[30] Neri, A.; Cucchariní, C.; Strik, H.: ASR-based Corrective Feedback on Pronunciation: Does It Really Work? In *INTERSPEECH*, Pittsburgh, PA, 2006.

[31] Olah, C.: Understanding LSTM Networks. [online] https://colah.github.io/posts/2015-08-Understanding-LSTMs/, 2015, [online, cit. 29.7.2019].

[32] Povey, D.; Gihostal, A.; Boulianne, G.; aj.: The Kaldi Speech Recognition Toolkit. 2011.

[33] Precoda, K.; Halverson, C. A.; Franco, H.: Effects of Speech Recognition-based Pronunciation Feedback on Second-Language Pronunciation Ability. In *Proceedings of InSTILL*, 2000, s. 102–105.

[34] Psutka, J.; Müller, L.; Matoušek, J.; aj.: *Mluvíme s počítačem česky*. Prague: Academia, 2006.

[35] Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, ročník 77, č. 2, 1989: s. 257–286.

[36] Stouten, F.; Martens, J.-P.: Speech recognition with phonological features: Some issues to attend. 2006.

[37] Strik, H.; Truong, K.; de Wet, F.; aj.: Comparing Different Approaches for Automatic Pronunciation Error Detection. *Speech Communication*, ročník 51, č. 10, 2009: s. 845 – 852.

[38] Tanner, M.; Landon, M.: The Effects of Computer-assisted Pronunciation Readings on ESL Learners' Use of Pausing, Stress, Intonation, and Overall Comprehensibility. *Language Learning and Technology*, ročník 13, 2001.

[39] Wei, S.; Hu, G.; Hu, Y.; aj.: A New Method for Mispronunciation Detection using Support Vector Machine Based on Pronunciation Space Models. *Speech Communication*, ročník 51, č. 10, 2009: s. 896 – 905.

[40] Witt, S.: Automatic Error Detection in Pronunciation Training: Where we are and where we need to go. 2012.

[41] Witt, S. M.; Young, S. J.: Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communication*, ročník 30, č. 2, 2000: s. 95–108.

[42] Woodland, P. C.; Gales, M. J. F.; Pye, D.; aj.: The Development of the 1996 Htk Broadcast News Transcription System. 1997.

[43] Young, S.; Evermann, G.; Gales, M.; aj.: *The HTK Book*. Cambridge: Cambridge University, 2002.

[44] Yu, D.; Deng, L.: *Automatic Speech Recognition: A Deep Learning Approach*. London: Springer, 2014.

Príloha A

Obsah priloženého pamäťového média

Na priloženom pamäťovom médiu sa nachádza

- elektronická verzia tohoto dokumentu spolu zo zdrojovými súbormi v jazyku LaTeX,
- archív obsahujúci skripty a zdrojové kódy, ktoré realizujú popísané experimenty.

Obsah bližšie popisuje súbor `README.txt`, ktorý sa nachádza v koreňovom adresári.