VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky
a komunikačních technologií

DIPLOMOVÁ PRÁCE

Brno, 2020 Bc. Jiří Nezval

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ
DEPARTMENT OF BIOMEDICAL ENGINEERING

ODHAD PŘESNOSTI ŘEČOVÝCH TECHNOLOGIÍ NA ZÁKLADĚ MĚŘENÍ SIGNÁLOVÉ KVALITY A OBSAHOVÉ BOHATOSTI AUDIA

ESTIMATION OF ACCURACY OF SPEECH TECHNOLOGIES BASED ON SIGNAL QUALITY AND AUDIO CONTENT RICHNESS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — **Bc. Jiří Nezval**
AUTHOR

VEDOUCÍ PRÁCE — **Ing. Petr Schwarz, Ph.D.**
SUPERVISOR

BRNO 2020

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Diplomová práce

magisterský navazující studijní obor **Biomedicínské a ekologické inženýrství**

Ústav biomedicínského inženýrství

Student: Bc. Jiří Nezval ***ID:*** 184255

Ročník: 2 ***Akademický rok:*** 2019/20

NÁZEV TÉMATU:

Odhad přesnosti řečových technologií na základě měření signálové kvality a obsahové bohatosti audia

POKYNY PRO VYPRACOVÁNÍ:

1) Prostudujte různé metody hodnocení signálové kvality zvukového záznamu. Mezi metrikami může být například poměr signálu a šumu, množství přebuzeného signálu, množství neřečového signálu, minimální úroveň signálu, počet kvantizačních úrovní atd. 2) Prostudujte metody pro převod zvukového záznamu na fonetický přepis a navrhnete metriku, která z fonetického přepisu odhadne obsahovou bohatost zvukového záznamu (například entropie). 3) Set řečových nahrávek s referenčními přepisy zpracujte automatickým přepisem a pomocí standardních metriky WER (word error rate) vyhodnoťte přesnost přepisu. Zároveň pro nahrávky vygenerujte metriky hodnotící signálovou kvalitu a obsahovou bohatost zvukového záznamu. 4) Vytvořte statistický model, který z dílčích metrik predikuje přesnost automatického přepisu. 5) Vyhodnoťte přesnost statistického modelu a diskutujte, které dílčí metriky jsou pro predikci přesnosti automatického přepisu nejdůležitější.

DOPORUČENÁ LITERATURA:

[1] I. TASHEV. Sound Capture and Processing: Practical Approaches. Wiley, 2009.

[2] SCHWARZ P. Phoneme Recognition based on Long Temporal Context. PhD thesis, Brno University of Technology, 2008.

Termín zadání: 3.2.2020 ***Termín odevzdání:*** 29.5.2020

Vedoucí práce: Ing. Petr Schwarz, Ph.D.

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

ABSTRAKT

Práce se zabývá teoretickým rozborem vzniku řeči, představuje možnosti využití řečových technologií a vysvětluje současný přístup k fonetickému přepisu řečových nahrávek. Jsou v ní popsány metriky hodnocení kvality audionahrávek, které rozdělujeme do dvou oddělených skupin. První skupinou jsou metriky signálové kvality, druhou metriky obsahové bohatosti. Prvním cílem praktické části je poté vytvořit statistický model pro predikci přesnosti strojového přepisu řečové nahrávky na základě měření její kvality. Druhým cílem je posoudit, které dílčí metriky jsou pro predikci přesnosti strojového přepisu nejdůležitější.

KLÍČOVÁ SLOVA

řeč, řečové technologie, fonetický přepis, signálová kvalita, obsahová bohatost, predikce přesnosti strojového přepisu, regrese

ABSTRACT

This thesis discusses theoretical analysis of the origin of speech, introduces applications of speech technologies and explains the contemporary approach to phonetical transcription of speech recordings. Furthermore, it describes the metrics of audio recordings quality assessment, which are split into two discrete classes. The first one groups signal quality metrics, while the other one groups content richness metrics. The first goal of the practical section is to create a statistical model for accuracy prediction of machine transcription of speech recordings based on a measurement of their quality. The second goal is to assess which partial metrics are the most essential for accuracy prediction of machine transcription.

KEYWORDS

speech, speech technologies, phonetic transcription, signal quality, content richness, accuracy of automatic transcription, regression

NEZVAL, Jiří. *Odhad přesnosti řečových technologií na základě měření signálové kvality a obsahové bohatosti audia*. Brno, 2020, 77 s. Diplomová práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav biomedicínského inženýrství. Vedoucí práce: Ing. Petr Schwarz, Ph.D.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Odhad přesnosti řečových technologií na základě měření signálové kvality a obsahové bohatosti audia“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno ....

podpis autora

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Petru Schwarzovi, Ph.D. za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Úvod **10**

1 Řeč, řečové technologie **12**
1.1 Vznik řeči . . . 12
1.2 Foném . . . 13
1.3 Řečové technologie . . . 13
1.4 Použité řečové technologie . . . 14

2 Fonetický přepis **16**
2.1 Fonetický rozpoznávač . . . 16
2.2 Model řeči . . . 16
2.3 Cepstrum . . . 17
2.4 Mel banka filtrů . . . 18

3 Metriky signálové kvality audia **21**
3.1 Vzorkovací frekvence . . . 22
3.2 Počet bitů . . . 22
3.3 Množství přebuzeného signálu . . . 23
3.4 Střední hodnota signálu . . . 24
3.5 Směrodatná odchylka . . . 24
3.6 Koeficient šikmosti . . . 25
3.7 Koeficient špičatosti . . . 25
3.8 Poměr signálu a šumu . . . 26
3.9 Minimální a maximální absolutní hodnota . . . 28
3.10 Minimální a maximální hodnota . . . 28
3.11 Délka ticha . . . 29
3.12 Délka úseků obsahující impulsní šum . . . 29
3.13 Délka technického signálu . . . 30
3.14 Délka signálu určeného k odfiltrování . . . 31
3.15 Délka nahrávky . . . 31

4 Návrh metriky obsahové bohatosti audia **32**
4.1 Histogram fonémů . . . 32
4.2 Entropie . . . 33
4.3 Křížová entropie . . . 35
4.4 Počet unikátních fonémů . . . 37

5 Statistický model **39**
5.1 Použité datové sady 39
5.2 Vyhodnocení přesnosti přepisu řeči na text 40
5.3 Příprava dat pro statistický model 42
5.4 Lineární regrese 52
5.5 Logistická regrese 54
5.6 K-křížová validace 58

6 Programová realizace **59**

7 Vyhodnocení statistického modelu **66**

Závěr **72**

Literatura **73**

Seznam obrázků

2.1 Blokové schéma převodu řeči na text 16
2.2 Blokové schéma převodu řeči na text 17
2.3 Převodní funkce frekvence v Hz na frekvence v Mel 19
2.4 Ilustrace Mel banky filtrů, převzato z [29] 19
3.1 Ilustrace procesu vzorkování a kvantizace 22
3.2 Přebuzení signálu 23
3.3 Příklad nesymetrie rozložení 25
3.4 Ukázka vlivu koeficientu špičatosti 26
3.5 Ilustrace významu SNR 27
3.6 Příklad impulsního rušení signálu 29
3.7 Příklad detekce tónu prahováním spektra 30
4.1 Příklad histogramů výskytu fonémů v nahrávkách 33
4.2 Entropie pro dva fonémy 34
4.3 Shodná pravděpodobnostní rozložení 36
4.4 Rozdílná pravděpodobnostní rozložení 36
4.5 Histogramy fonémů nahrávek nevhodných pro přepis 38
5.1 Poměr signálu a šumu pro obě nové datové sady 44
5.2 Poměr signálu a šumu pro obě datové sady 45
5.3 Délky nahrávek pro obě datové sady 45
5.4 Množství přebuzeného signálu pro obě datové sady 46
5.5 Směrodatné odchylky pro obě datové sady 46
5.6 Koeficient šikmosti pro obě datové sady 47
5.7 Koeficient špičatosti pro obě datové sady 47
5.8 Maximální absolutní hodnoty pro obě datové sady 48
5.9 Maximální hodnoty pro obě datové sady 48
5.10 Střední hodnoty pro obě datové sady 49
5.11 Minimální absolutní hodnoty pro obě datové sady 50
5.12 Minimální hodnoty pro obě datové sady 50
5.13 Poměr signálu určeného k filtrování pro obě datové sady 51
5.14 Poměr ticha pro obě datové sady 51
5.15 Poměr technického signálu pro obě datové sady 52
5.16 Závislost hodnoty WER na poměru signálu a šumu 55
5.17 Proložení logistickou regresí pro závislost z Obr. 5.16 56
6.1 Zjednodušené blokové schéma programové realizace 59
6.2 Výstup SQE - textový soubor s metrikami a jejich specifikacemi . . . 60
6.3 Textový soubor - ukázka výstupu z PHR 61
6.4 Ukázka výstupu strojového přepisu nahrávky na text 62

6.5 Ukázka části výstupu skórovacího nástroje 63
7.1 Závislost hodnoty ztrátové funkce na parametru α 69
7.2 Závislost hodnoty ztrátové funkce na počtu iterací 69

Úvod

Automatický přepis řeči do textové podoby je v posledních letech velkým fenoménem a jeho využití sahá do nejrůznějších oblastí života. Bezpečnostní složky jej používají pro hledání informací v obrovských archivech telefonních hovorů a call-centra pro zajištění kvality služeb. Zasahuje však i do každodenních činností běžných lidí při hlasovém diktování zpráv či komunikací se sofistikovanými systémy, jako jsou např. Siri od společnosti Apple, Alexa od společnosti Amazon a jinými život usnadňujícími zařízeními.

Použití těchto technologií je ovšem výpočetně i časově náročné. Zejména při strojovém přepisu velké databáze audionahrávek je žádoucí omezit přepis nekvalitních nahrávek, ze kterých dostaneme špatné či dokonce žádné výsledky. Cílem této práce je najít rychlý způsob odstranění těchto nevyhovujících nahrávek.

V první kapitole začnu s procesem vzniku řeči v lidském těle, zavedu velmi důležitý pojem foném jakožto základní řečovou jednotku a popíši aktuální způsoby využití a důležitost řečových technologií. Na závěr uvedu všechny technologie používané v této diplomové práci.

Dále ve druhé kapitole uvedu současný přístup k převodu řečového signálu na fonetický přepis a vysvětlím jednotlivé kroky tohoto procesu. Těchto poznatků později využiji ve čtvrté kapitole, kde představím jiný pohled na kvalitu audionahrávek vhodných pro strojový přepis, konkrétně metriky posuzující obsahovou bohatost.

Ve třetí kapitole popisuji metriky signálové kvality a jejich použití pro nalezení nekvalitních nahrávek. Pro samotný výpočet těchto metrik využívám Speech Quality Estimator od brněnské společnosti Phonexia. Zde pracujeme s časovou, frekvenční i časově-frekvenční analýzou řečových signálů. Navážu následně čtvrtou kapitolou, kde zavádím metriky obsahové bohatosti. Tyto metriky se dívají např. na fonetickou skladbu nahrávky či statistické rozložení fonémů v nahrávce.

Pátá část této diplomové práce se zabývá procesem vytváření statistického modelu. Zde popíši použité datové sady, vysvětlím standardní postupy pro vyhodnocení přesnosti strojového přepisu a vysvětlím, jak můžeme díky metodám strojového učení tyto přesnosti predikovat.

Cílem této diplomové práce je vytvoření statistického modelu k odhadu vhodnosti audionahrávek pro strojový přepis a vyhodnocení důležitosti jednotlivých metrik. V následující kapitole ukáži programovou realizaci vytvoření statistického modelu a vysvětlím princip použití a výstupy všech mnou využívaných nástrojů. Poslední kapitolu věnuji statistickému vyhodnocení vytvořených modelů.

Tato diplomová práce vznikla ve spolupráci s brněnskou firmou Phonexia. Phonexia se zabývá řečovou analytikou a hlasovou biometrií. To zahrnuje například přepis řeči do textu, detekce klíčových slov, identifikace řečníka, identifikace jazyka, identifikace pohlaví apod. V práci využívám některé z jejich technologií, konkrétně Phonexia Speech Quality Estimator, Phonexia Phoneme Recognizer a Phonexia Speech to Text. Rád bych tímto poděkoval za cenné rady, znalosti a poskytnutí těchto technologií.

1 Řeč, řečové technologie

V první kapitole této práce nejdříve vysvětlíme základní principy tvorby lidské řeči, které později využijeme pro popis složitějších metod strojového rozpoznávání řeči. Následně zavedeme pojem foném jakožto velmi důležitou součást celé práce. Dále plynule přejdeme ke strojovému zpracování řeči a rozebereme využití řečových technologií v běžném životě. Některé z těchto moderních technologií budu v průběhu této diplomové práce používat i já, v poslední části první kapitoly je tedy v rychlosti představím a popíšu princip jejich použití.

1.1 Vznik řeči

Na vzniku řeči v lidském těle se podílí několik skupin orgánů, které dohromady tvoří hlasový trakt. Ten můžeme dále rozdělit na dechové, hlasové a artikulační ústrojí.

Dechové ústrojí slouží jako zdroj energie pro tvorbu řeči. Vzduch, který proudí dechovým ústrojím při nádechu je zdrojem energie pro výdechový proud vzduchu, díky kterému může docházet ke tvorbě řeči. Tento výdechový proud má vliv na sílu i výšku hlasu (výšku ale více ovlivňuje nastavení hlasivek) [28].

V hlasovém ústrojí, které je uloženo v hrtanu, se nachází nejdůležitější orgán pro tvorbu řeči, a to hlasivky. Zde se nachází hlasivková štěrbina, která vzniká mezi hlasovými vazy. Rozlišujeme dvě základní postavení hlasivek - klidové, které umožňuje volný průchod vzduchu při dýchání s odkrytou štěrbinou a fonační, které se uplatňuje při tvorbě hlasu. V případě fonačního postavení dochází během výdechu k průchodu vzduchu hlasivkovou štěrbinou, což způsobuje rozkmitání hlasových vazů za vzniku znělých zvuků řeči. Délka, tloušťka a napětí hlasových vazů ovlivňují výšku tónu, tj. frekvenci. Tímto způsobem vzniká základní (hlasivkový) tón. Při klidovém postavení hlasivek dochází k tvorbě neznělých zvuků řeči. Znělé i neznělé zvuky řeči jsou dále modifikovány v artikulačním ústrojí [7][28].

Artikulační ústrojí tvoří nadhrtanové dutiny (dutina hrdelní, ústní, nosní) a artikulační orgány (měkké patro, jazyk, rty, zuby). Nadhrtanové dutiny se podílí na vytváření tónové struktury. V těchto dutinách dochází k rezonanci výdechového proudu vzduchu z hlasového ústrojí, což má za následek změnu rozložení akustické energie. Tímto způsobem vzniká základ pro samohlásky - složený zvuk tónového charakteru. Nadhrtanové dutiny se na tvorbě řeči podílí pasivně, naopak artikulační orgány aktivně. Pohybem např. rtů nebo jazyka mohou změnit průchod vzduchu dutinami a také frekvenční vlastnosti. Podílí se na vytváření šumové složky, která tvoří základ pro souhlásky [8].

1.2 Foném

V artikulačním ústrojí člověka tedy vzniká řeč jako posloupnost vydaných zvuků. Soubor tímto způsobem vytvořených zvuků nazýváme hláska, kterou považujeme za základní řečovou jednotku. Při zkoumání hlásek se nezabýváme daným jazykem ani jejím významem. Pokud ovšem zkoumáme řečové zvuky z hlediska stavby konkrétního jazyka, zavádíme tím lingvistickou základní řečovou jednotku foném. Foném je nejmenší jednotka zvukové stránky řeči, která má v daném jazyce rozlišovací funkci [39]. Foném pro nás bude v této práci velmi důležitým kouskem skládačky pro vytvoření funkčního statistického modelu, jak uvidíme v pozdější fázi práce. V kapitole 2 je popsán postup převodu řečové nahrávky na posloupnost fonémů a v kapitole 4 fonémy využijeme pro návrh metriky hodnotící kvalitu nahrávky.

1.3 Řečové technologie

Od popisu tvorby řeči v lidském těle a zavedení pojmu foném se přesuneme k velice úzce spjatému tématu řečových technologií. Řeč je jedním z nejběžnějších způsobů lidské komunikace. V současné době, kdy je nesmírně rozšířené použití počítačů, je snaha o ovládnutí této formy komunikace tak, abychom mohli komunikovat s počítačem za pomocí řeči stejně jako s běžným člověkem. Vývoj řečových technologií je velmi aktuální téma, a čím dál více se setkáváme s nejrůznějšími aplikacemi řečových technologií v každodenním životě. Ať už jde o hlasové ovládání chytrých zařízení, hlasovou identifikaci po telefonu či vyřízení reklamace s hlasovým robotem místo živého operátora, dokáží nám tyto technologie ulehčit život a otvírají nové možnosti využití hlasu. Současný výzkum se zaměřuje například i na zlepšení životních podmínek hendikepovaných občanů. Zde se dostáváme k tématům jako je hlasová syntéza nebo automatický překlad mluveného slova do znakové řeči [38].

Řečové technologie můžeme rozdělit na několik podskupin. První z nich je oblast známá jako *automatické zpracování řeči* [13]. Tato oblast zpracování řeči se zkoumá již více jak padesát let. Jde o systémy, které mají usnadnit jak řečovou komunikaci mezi člověkem a počítačem, tak i mezi lidmi. Mezi populární aplikace pro komunikaci s počítačem jsou hlasové vyhledávání [37], osobní digitální asistenti, hraní her, systémy pro chytré domácnosti nebo např. pro ovládání auta hlasovými pokyny [31]. Mezi lidmi může automatické zpracování řeči pomoci např. jako živý překladač z jednoho jazyka do druhého [9]. Bez řečových technologií je pro konverzaci mezi dvěma lidmi ovládajícími jiné jazyky potřeba lidský překladatel. Můžeme si představit, že pro člověka, který neovládá čínštinu, je velmi obtížné cestovat Čínou na vlastní pěst. Ve zkratce, jsou to řečové technologie zkoumající *obsah* řeči. Tyto technologie využívají hojně i vládní či komerční instituce. Ve vládním sektoru mohou být použity

pro boj proti zločinu a odhalování bezpečnostních hrozeb [1]. V komerčním sektoru zase např. jako analytické nástroje v call-centrech.

Druhou skupinou je *hlasová biometrie*. Na rozdíl od automatického zpracování řeči nezkoumá hlasová biometrie, co daný člověk říká. Dokáže z hlasu získat unikátní atributy a na jejich základě ho zařadit do některé z dostupných kategorií. Hlasovou biometrii můžeme využít pro hlasovou identifikaci člověka, zjištění pohlaví, odhad věku či rozpoznání jazyka, kterým se mluví. Hlasová identifikace se nyní stává čím dál rozšířenější součástí call-center společností jako přídavná vrstva zabezpečení proti podvodníkům. Velmi užitečná je i pro forenzní specialisty, kteří mají hlasovou nahrávku jako důkazní materiál a potřebují prokázat identitu řečníka na nahrávce [5].

Dalším využitím řečových technologií jsou systémy *hlasové syntézy*. Zde se řeší, jak přimět počítač mluvit co nejvíce jako člověk. Využití sahá od automatických telefonních hlášek, přes usnadnění života lidem s hendikepem. Tato oblast se v posledních letech velmi vyvinula a nejlepší systémy hlasové syntézy už rozhodně nezní roboticky. Jedním z prvních využití byly čtecí systémy pro nevidomé. Takový systém by vzal text elektronické knihy a přetvořil ho na hlasovou nahrávku. V každodenním životě využijeme hlasovou syntézu např. v navigacích při hlasových instrukcích či stále častěji pro komunikaci s hlasovým robotem [12][17].

1.4 Použité řečové technologie

V průběhu této diplomové práce jsem postupně použil několik řečových technologií. Šlo o technologie přepisu řeči na text, přepisu řeči na posloupnost fonémů a generátor signálových metrik pro danou nahrávku. Všechny tyto technologie byly poskytnuty brněnskou firmou Phonexia. V krátkosti zde každou z technologií představím.

1.4.1 Phonexia Speech Quality Estimator

Phonexia Speech Quality Estimator (SQE) kvantifikuje akustickou kvalitu audionahrávek. Pomáhá uživateli zjistit, zda je daná nahrávka dostatečně kvalitní pro zpracování dalšími řečovými technologiemi (např. pro přepis nahrávky do textové podoby).

Důvodem k použití SQE jako rozhodovacího nástroje před dalším zpracováním je jeho rychlost a výpočetní nenáročnost. Oproti náročnějším technologiím jako je přepis řeči na text je SQE mnohonásobně rychlejší a dokáže odfiltrovat velké množství nekvalitních nahrávek.

Vstupem SQE je audionahrávka ve formátu **.wav* nebo **.raw*, *a-law* nebo *μ-law* kódování, *PCM* (pulsní kódová modulace).

Výstupem SQE pro danou audionahrávku je seznam metrik signálové kvality a jejich hodnot (význam metrik je vysvětlen v kapitole 3). Z těchto metrik je poté spočítáno výsledné *skóre* nahrávky, tj. procentuální vyjádření kvality nahrávky z intervalu $< 0, 100 >$, podle kterého se námi zvoleným prahem rozhodujeme, zda je pro nás daná nahrávka dostatečně kvalitní. V současné podobě je výpočet skóre realizován váhovaným průměrem metrik s ručně nastavenými vahami. Výstup SQE je zobrazen na Obr. 6.2.

1.4.2 Phonexia Phoneme Recognizer

Phonexia Phoneme Recognizer (PHR) je nástroj pro přepis audionahrávky do fonetického zápisu. Vstupem je opět audionahrávka ve formátu **.wav* nebo **.raw*, *a-law* nebo *μ-law* kódování, *PCM* (pulsní kódová modulace). Výstupem je posloupnost fonémů detekovaných v nahrávce. Detekované fonémy jsou závislé na jazyku, protože každý jazyk může mít odlišnou sadu fonémů. Dokonce i pro jeden jazyk mohou existovat různé sady fonémů. Více informací v kapitole 4. Jak vypadá výstup PHR můžeme vidět na Obr. 6.3.

1.4.3 Phonexia Speech to Text

Textová podoba řeči má velkou řadu výhod, zabírá méně místa, dá se v ní rychleji hledat informace a otevírá nové možnosti zpracování (např. metody přirozeného zpracování jazyka - NLP). Text můžeme snadno číst i upravovat.

Phonexia Speech to Text (STT) pracuje s akustickým a jazykovým modelem. Akustický model popisuje výslovnosti v daném jazyce a jazykový model zahrnuje statistiky o tom jak jsou daná slova v jazyce používaná dohromady.

Vstupem pro STT je řečová nahrávka ve formátu **.wav* nebo **.raw*, *a-law* nebo *μ-law* kódování, *PCM* (pulsní kódová modulace), popřípadě stream ve formátech *RTP* či *HTTP*.

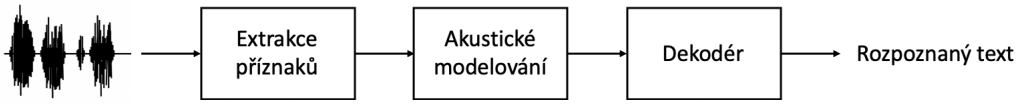
Výstupem STT je textový soubor obsahující přepis řečové nahrávky. Příklad takového výstupu lze vidět na Obr. 6.4.

2 Fonetický přepis

V kapitole 1.2 jsme zavedli pojem foném a zlehka jsme zmínili, že pro nás fonémy budou v této práci velmi důležitým nástrojem. Nyní se podíváme na proces, jakým z řečové nahrávky vznikne posloupnost fonémů. Popíšeme jednotlivé součásti fonetického rozpoznávače a uvedeme moderní přístupy v těchto součástech používané.

2.1 Fonetický rozpoznávač

Fonetický rozpoznávač má za úkol převést audionahrávku obsahující řeč na posloupnost fonémů, které je možné dále zpracovávat. Obecné schéma na Obr. 2.1 popisuje proces přetvoření řečového signálu na text.



Obr. 2.1: Blokové schéma převodu řeči na text

Vstupem prvního bloku je řečový signál, má za úkol extrakci příznaků, odstranění nepotřebných informací a kompresi důležitých informací. V druhém bloku probíhá přiřazení částí signálů k uchovaným vzorům řečových jednotek (v našem případě k fonémům). Dekodér poté nalezne nejlepší pořadí detekovaných fonémů za využití vlastností konkrétního jazyka [34].

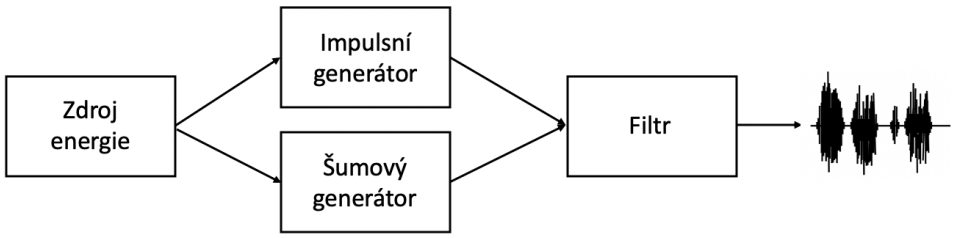
2.2 Model řeči

Prvním krokem k fonetickému přepisu řečové nahrávky je tedy extrakce příznaků. Cílem je vyjádřit řečový signál v takové podobě, aby zachoval co nejvíce užitečné informace a zároveň odstranil neužitečné složky. V současné době je nejrozšířenějším způsobem vyjádření řečového signálu pomocí tzv. *koeficientů z banky Mel filtrů*. Pro jeho pochopení je potřeba začít signálovým modelem řeči.

V kapitole 1.1 je popsán proces vzniku řeči u člověka. Důležitou roli hrají plíce, které slouží jako zdroj energie. Dále prochází vzduch hrtanem, kde rozkmitá hlasivky. Hlasivky vydají zvuk, který je upraven průchodem artikulačním traktem tak, jak jej poté můžeme slyšet. Hlasivky mají v procesu tvoření řeči dva různé stavy. Prvním z nich je kmitání, kdy hlasivky periodicky kmitají frekvencí základního tónu (pro muže se pohybuje kolem 90-120 Hz, pro ženy 150-300 Hz a pro děti 350-400 Hz).

Tento stav se projevuje jako znělá hláska. Druhým stavem je klid, kdy mohou vznikat neznělé hlásky [19].

Analogií s využitím signálové terminologie můžeme považovat plíce jako zdroj energie, hlasivky jako buzení signálu (impulsní generátor generující impulsy o frekvenci základního tónu hlasu) a artikulační trakt jako signálový filtr upravující finální podobu řeči (ilustrace na Obr. 2.2).



Obr. 2.2: Blokové schéma modelu tvorby řeči

Výsledný řečový signál tedy můžeme vyjádřit jako

$$s(t) = g(t) * h(t), \tag{2.1}$$

kde $s(t)$ je řečový signál proměnný v čase, $g(t)$ je impuls (buzení hlasivek), $h(t)$ je impulsní charakteristika filtru (artikulačního traktu) a $*$ je operátorem konvoluce.

Ve frekvenční oblasti vypadá vztah následovně:

$$S(f) = G(f)H(f), \tag{2.2}$$

kde $S(f)$ je spektrum řečového signálu proměnného v čase, $G(f)$ je spektrum impulsu a $H(f)$ je přenosová charakteristika filtru.

2.3 Cepstrum

Pro další zpracování řeči je výhodné buzení a filtraci oddělit. Abychom každý řečový signál dokázali rozdělit na tyto dvě složky, je potřeba udělat jeho dekonvoluci. Je to proces, při kterém se snažíme oddělit vliv buzení a artikulačního traktu. Protože jak v časové, tak i ve frekvenční oblasti je toto oddělení netriviální, zavedeme pojem *cepstrum* [24]. Cepstrum nám pomůže nahradit součin z 2.2 součtem, ze kterého pak oddělení bude možné.

Cepstrum definujeme jako

$$c(n) = \mathscr{F}^{-1}\{ln|\mathscr{F}[s(n)]|^2\}, \tag{2.3}$$

kde $c(n)$ jsou cepstrální koeficienty, $\mathscr{F}$ značí Fourierovu diskrétní transformaci a $s(n)$ je původní řečový signál [24].

Dále můžeme upravovat:

$$\begin{aligned} c(n) &= \mathscr{F}^{-1}\{ln|\mathscr{F}[s(n)]|^2\} = \\ &= \mathscr{F}^{-1}\{ln|S(f)|^2\} = \\ &= \mathscr{F}^{-1}\{ln|G(f)|^2|H(f)|^2\} = \\ &= \mathscr{F}^{-1}\{ln|G(f)|^2 + ln|H(f)|^2\} \end{aligned} \tag{2.4}$$

Zpětná Fourierova transformace je lineární, můžeme tedy psát:

$$\begin{aligned} c(n) &= \mathscr{F}^{-1}\{ln|G(f)|^2\} + \mathscr{F}^{-1}\{ln|H(f)|^2\} = \\ &= c_g(n) + c_h(n) \end{aligned} \tag{2.5}$$

Pro cepstrum $c(n)$ obsahující M koeficientů můžeme najít m takové, že $c(1)$ až $c(m-1)$ odpovídají $c_h(n)$ a $c(m+1)$ až $c(M)$ odpovídají $c_g(n)$.

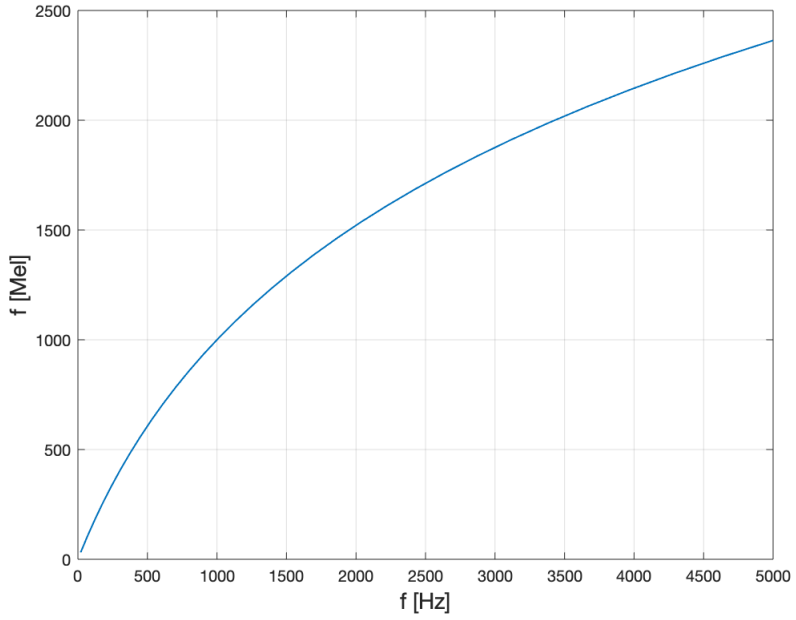
2.4 Mel banka filtrů

Lidské ucho vnímá zvuk nelineárně (nízké frekvence rozlišujeme lépe než vysoké frekvence) [26]. Jelikož výpočet cepstra tuto skutečnost vůbec neuvažuje, zavedeme tzv. *koeficienty Mel banky filtrů*. Jde o výstupy energií sady trojúhelníkových filtrů nelineárně rozmístěných tak, aby odpovídaly fyziologickému vnímání zvuku. Těmito koeficienty nahradíme $\mathscr{F}[s(n)]|^2$ v 2.3 [29].

Pro vytvoření takovýchto nelineárních filtrů se využívá převod Hertzů na Mely podle vzorce:

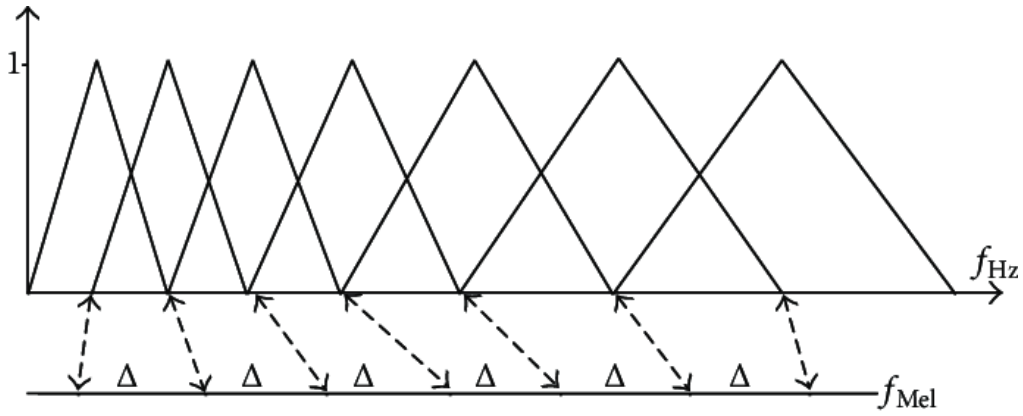
$$F_{Mel} = 2959\ log(1 + \frac{F_{Hz}}{700}), \tag{2.6}$$

kde F_{Mel} je frekvence v Melech a F_{Hz} je frekvence v Hz [33].



Obr. 2.3: Převodní funkce frekvence v Hz na frekvence v Mel

Lineární rozmístění filtrů na Mel ose vyústí v nelineární rozmístění na Hz ose. Ilustraci Mel banky filtrů vidíme na Obr. 2.4.



Obr. 2.4: Ilustrace Mel banky filtrů, převzato z [29]

Výpočet energie daného filtru provedeme násobením umocněné Diskrétní Fourierovy transformace signálu s odpovídajícím trojúhelníkovým oknem a sečtením všech hodnot. Cepstrum vypočtené pomocí energií Mel banky filtrů označujeme jako *Mel-frekvenční cepstrální koeficienty (MFCC)*.

Druhým krokem fonetického přepisu je akustické modelování. Moderní metody akustického modelování využívají neuronových sítí. Vstupem neuronové sítě jsou Mel-bank koeficienty, není tedy nutný celý výpočet Mel-frekvenčních cepstrálních

koeficientů. Výstupem neuronových sítí jsou pravděpodobnosti výskytu fonémů pro daný (standardně 25 ms dlouhý) úsek.

Dekodér nakonec převádí matici pravděpodobností na fonémový řetězec. Pro určení nejpravděpodobnější sekvence fonémů může dekodér využít statistiky přechodů mezi jednotlivými fonémy.

3 Metriky signálové kvality audia

Při zpracování audionahrávek je velmi žádoucí zpracovávat jen ty z nich, které jsou pro zpracování vhodné. V praxi můžeme v prostředí komerčního i vládního sektoru očekávat zpracování obrovského množství audionahrávek každý den. Mluvíme zde o objemech audia od desítek po desítky tisíc hodin. Konkrétně ve vládním sektoru je naprosto běžné, že zpracovávané nahrávky obsahují nejrůznější ruchy okolí, šum popř. dlouhé úseky ticha. Odstraněním nežádoucích nahrávek nepřinášejících žádnou hodnotu můžeme ušetřit čas i peníze.

Máme tedy u hodnocení kvality nahrávek dva základní požadavky. Prvním z nich je co nejpřesnější určení vhodnosti pro další zpracování. Procedury jako identifikace řečníka z řeči, identifikace jazyka či přepis mluvené řeči na text jsou výpočetně náročné. Každá nahrávka, která je takhle zpracována a poskytne buď špatné nebo žádné výsledky, znamená ztracený čas. To může ve větším měřítku znamenat např. pozdní odhalení kriminálníků nebo neodvrácení veřejné hrozby. Druhý požadavek víceméně vyplývá z prvního, je jím rychlost ohodnocení. Aby bylo hodnocení užitečné, musí být výrazně rychlejší než všechny výše zmíněné technologie.

Pro hodnocení kvality nahrávek používáme námi zvolené parametry audia, tzv. metriky. V této kapitole se zaměříme na metriky založené na signálové kvalitě. Zde pracujeme se zpracováním řečového signálu v časové i ve frekvenční oblasti. Tento přístup ovšem přináší jistá omezení. Nejsou ojedinělé případy, kdy je nahrávka ohodnocena jako kvalitní, avšak při následném strojovém přepisu mluvené řeči na text na první pohled vidíme, že v nahrávce buď vůbec žádná řeč není, nebo je nesrozumitelná, a tedy technologiemi nezpracovatelná. Tento problém se pokusíme vyřešit v kapitole 4, kde představíme metody hodnocení obsahové kvality audionahrávek. Zde se budeme zabývat fonetickým přepisem audia a vlivem statistického rozložení fonémů v nahrávce na výslednou kvalitu přepisu řeči do textu.

Postupně vysvětlíme všechny signálové parametry a metriky, které v práci používáme. Všechny tyto metriky vychází z Speech Quality Estimatoru (SQE) popsaného v kapitole 1.4.1. U každé metriky také uvádíme její název ve výstupním souboru SQE pro snadnou orientaci v praktické části této práce.

3.1 Vzorkovací frekvence

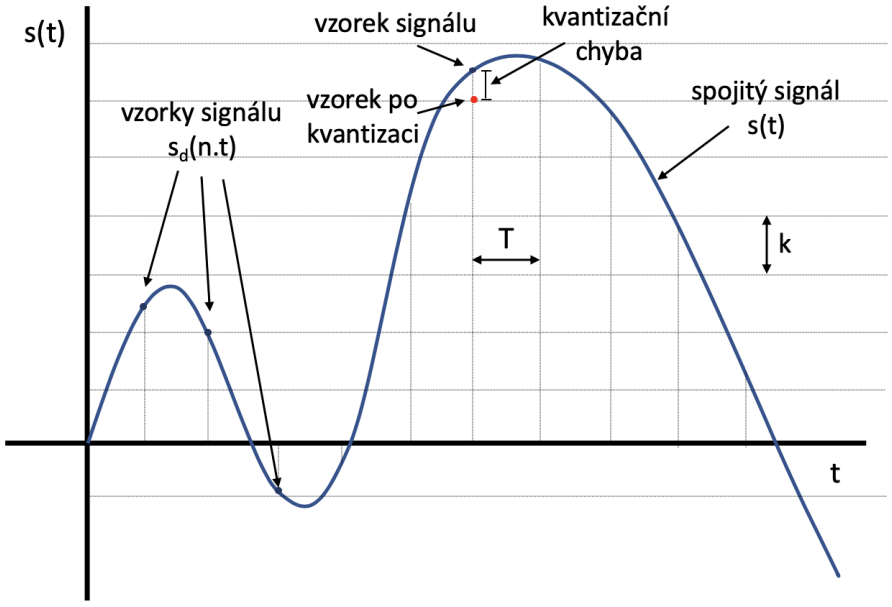
Pro počítačové zpracování řeči je potřeba převést spojitý řečový signál $s(t)$ na diskrétní posloupnost vzorků $s_d = s(nT)$, kde T je perioda vzorkování a $n = 0, ..., \infty$ (viz Obr. 3.1). Vzorkovací frekvenci následně vypočítáme jako $f_{vz} = 1/T$. Tato frekvence má omezení vyplývající z Nyquistova vzorkovacího teorému [28].

V SQE najdeme vzorkovací frekvenci pod názvem *waveform_sample_freq* a je vyčtena z metadat audio souboru. Jako dostatečná hodnota je brána $f_{vz} = 8$ kHz.

3.2 Počet bitů

Po navzorkování řečového signálu je ještě nutné převést stávající hodnoty signálu na hodnoty z konečného souboru hodnot tzv. kvantizaci. Jde o aproximaci analogové hodnoty vzorku řečového signálu jednou z konečného počtu hodnot. Při tomto procesu dochází ke ztrátě informace nazývané kvantizační zkreslení (viz Obr. 3.1) [28].

Je nutné znát počet hodnot (kvantizačních úrovní) a kvantizační krok. Počet úrovní se zpravidla volí ve tvaru 2^B, kde B je počet bitů [28].



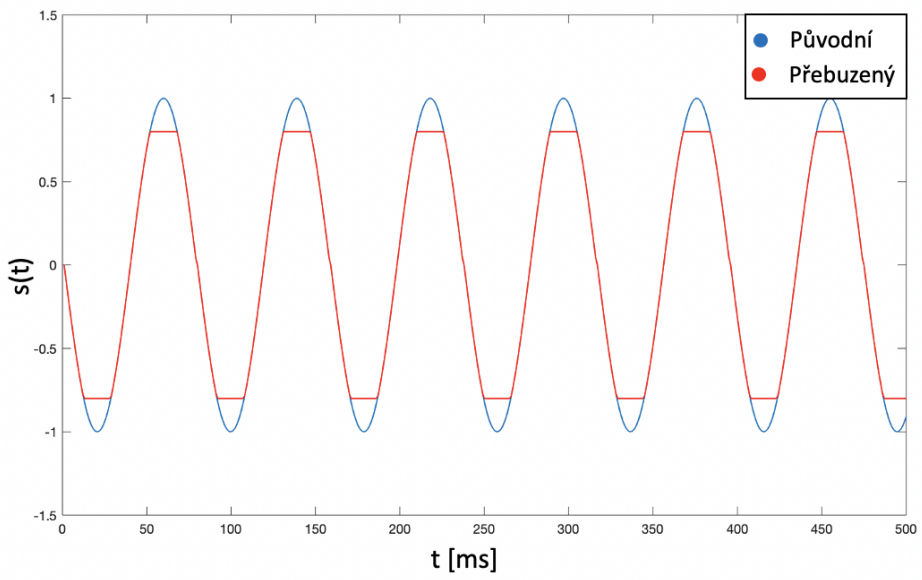
Obr. 3.1: Ilustrace procesu vzorkování a kvantizace. T označuje vzorkovací periodu, k označuje kvantizační krok.

V SQE se počet bitů zjišťuje přibližně postupným procházením hodnot a zapamatováním unikátních hodnot. Podle počtu unikátních hodnot y_{un} lze zjistit, kolik alespoň bitů dané kvantování obsahuje výpočtem $N_{bitů} = \log_2 y_{un}$. Jako dostačující

hodnota se bere $B \geq 10$ při lineárním kvantování a $B \geq 8$ při logaritmickém kvantování ($a-law$ nebo $\mu-law$). Hodnota je označena jako *waveform_n_bits* a související hodnotu počtu kvantizačních úrovní najdeme pod názvem *waveform_n_levels*.

3.3 Množství přebuzeného signálu

Přebuzení signálu je zkreslení signálu způsobující ztrátu hodnot vyšších než je maximální možná hodnota. Maximální možná hodnota je dána dynamickým rozsahem signálu, jinými slovy počtem kvantizačních rovin. Například při 16-bitovém znaménkovém kvantování je maximální kladná hodnota 32 767 a všechny vyšší hodnoty původního signálu budou po kvantování rovny právě této hodnotě (viz Obr. 3.2) [42].



Obr. 3.2: Přebuzení signálu. Modře je znázorněn původní signál a červeně přebuzený signál.

Množství přebuzeného signálu (angl. *signal clipped length*) nám tedy říká, jestli je používaný dynamický rozsah dostačující, nebo o významnou část signálu přicházíme. Na Obr. 3.2 můžeme pozorovat osekání špiček signálu. Přebuzení navíc přidává parazitní frekvence prakticky v celém spektru.

Pro výpočet délky přebuzeného signálu se v SQE používá následující přístup. Signál se rozdělí na úseky dělky 25 ms a pokud je v daných 25 ms alespoň jedna hodnota signálu vyšší než daný práh (standardně nastavený na 90% maximální možné hodnoty), je tento úsek označen jako přebuzený. Výsledná hodnota je tedy dána součtem všech přebuzených úseků.

Přebuzení signálu může být způsobeno např. chybným nastavením předzesilovače. Dříve se přebuzení mohlo vyskytnout u pevných linek. Nynější mobily většinou využívají dynamické přizpůsobení rozsahu.

Tato metrika je v SQE označená jako *waveform_clipped_length*. Velikost prahu je zde označena jako *waveform_clipping_threshold*.

3.4 Střední hodnota signálu

Střední hodnota signálu slouží jako odhad pozice nulové izolinie signálu. Tu chceme mít ideálně v hodnotě 0. Použitím špatného kodeku, popř. špatným zadáním parametrů kódování audia se ovšem nulová izolinie signálu může posunout. Tento problém se dá řešit např. filtrem typu horní propust, který odstraní nižší frekvence a tím i stejnosměrnou složku signálu.

Průměr signálu spočítáme jako:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \tag{3.1}$$

kde $\bar{x}$ označuje průměr hodnot signálu, x_i i-tý vzorek signálu a n počet vzorků.

V SQE najdeme tuto metriku pod názvem *waveform_mean*.

3.5 Směrodatná odchylka

Směrodatná odchylka nám říká, jak moc se průměrně liší hodnoty signálu od jeho průměru. Vypočítáme ji jako:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{3.2}$$

kde σ označuje směrodatnou odchylku, $\bar{x}$ průměr hodnot signálu, x_i i-tý vzorek signálu a n počet vzorků.

Podíváme-li se na histogram řečového signálu, směrodatná odchylka nám udává šířku rozložení. Tohoto faktu se dá využít, jelikož histogram čisté řeči se dá aproximovat Gamma rozložením, tj. rozložením s malou šířkou a výraznou strmostí. Naopak histogram šumu se vyznačuje spíše Gaussovským rozložením [2].

V SQE je tato hodnota označena jako *waveform_standard_deviation*.

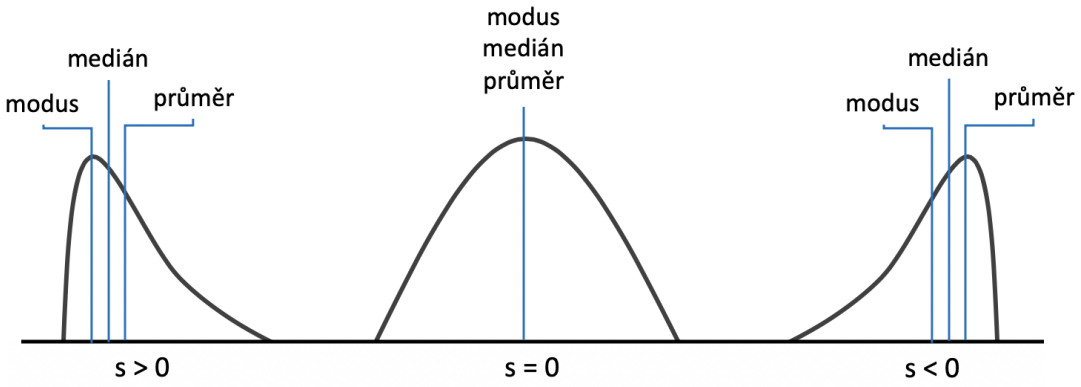
3.6 Koeficient šikmosti

Z rozložení hodnot vychází i další metrika koeficient šikmosti (angl. *skewness*). Směrodatná odchylka nám sice dává informaci o šířce rozložení, ovšem nevíme nic o jeho tvaru. Koeficient šikmosti vyjadřuje míru asymetrie rozložení (Obr. 3.3) [3]. Vypočítáme ho jako:

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}, \tag{3.3}$$

kde s označuje hodnotu koeficientu šikmosti , $\bar{x}$ průměr hodnot signálu, x_i i-tý vzorek signálu a n počet vzorků.

Pro dokonale symetrické rozložení je hodnota skewness rovna 0. Čistá řeč by měla mít symetrické rozložení.



Obr. 3.3: Příklad nesymetrie rozložení. Koeficient šikmosti je označen jako s. Vlevo je znázornění pro kladnou hodnotu koeficientu šikmosti, uprostřed pro hodnotu 0 a vpravo pro zápornou hodnotu s.

Kladná hodnota udává vyosení doleva, záporná hodnota vyosení doprava. V SQE má metrika název *waveform_skewness*.

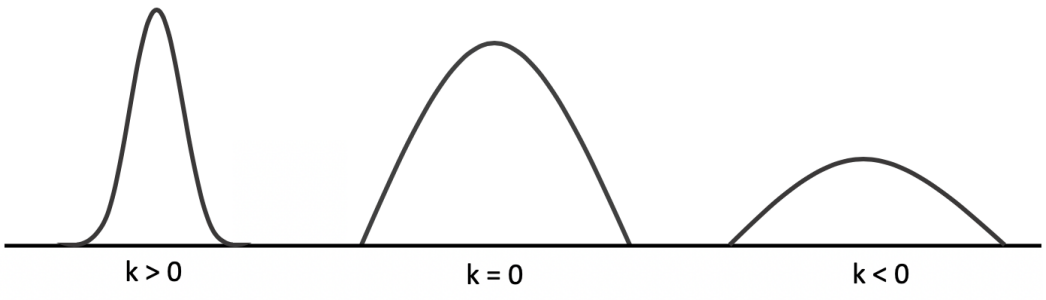
3.7 Koeficient špičatosti

Další informaci o tvaru rozložení můžeme získat metrikou koeficient špičatosti (angl. *kurtosis*). Ten nám popisuje strmost histogramu [3]. Vypočítáme ho jako:

$$k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}, \tag{3.4}$$

kde k označuje hodnotu koeficientu špičatosti , $\bar{x}$ průměr hodnot signálu, x_i i-tý vzorek signálu a n počet vzorků.

Normální rozložení má koeficient špičatosti roven 0. Kladná hodnota značí, že většina hodnot signálu se příliš neliší od střední hodnoty. Záporná hodnota naopak značí rovnoměrnější rozdělení (Obr. 3.4) [11].



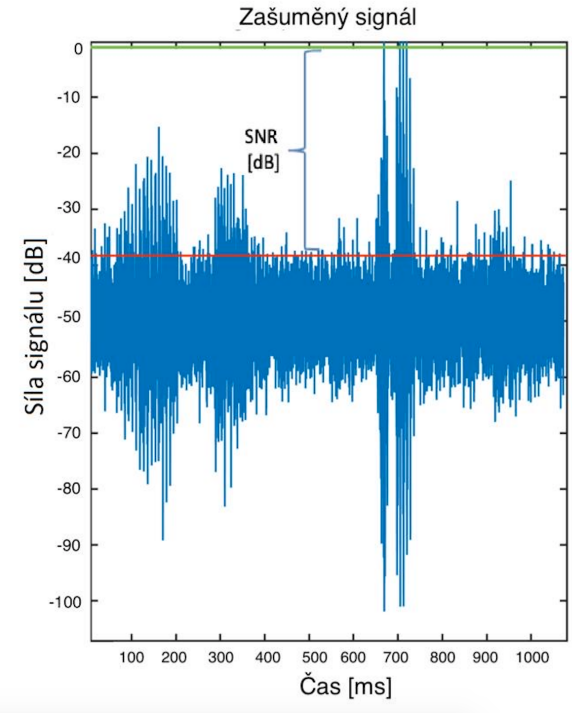
Obr. 3.4: Příklad rozložení pro kladnou (vpravo), nulovou (uprostřed) a zápornou (vlevo) hodnotu koeficientu špičatosti. Koeficient špičatosti je označen jako k.

V SQE má metrika název *waveform_kurtosis*.

3.8 Poměr signálu a šumu

Poměr signálu a šumu (SNR) říká, jaký je poměr energie užitečného signálu (v tomto případě řeči) a energie šumu (Obr. 3.5). Hodnota je udávaná v dB.

Hodnota $SNR = 0$ znamená, že je v daném signálu stejné množství řeči i šumu. Signál je považován za kvalitní při $SNR > 15$ dB.



Obr. 3.5: Ilustrace významu SNR - zeleně je označena úroveň zpracovaného signálu a červeně úroveň šumu.

V SQE je SNR počítáno dvěma různými způsoby:

- zkoumáním podobnosti rozložení vzorků signálu s Gamma/Gaussovským rozložením a analytickým převodem na SNR
- použitím detektoru řečové aktivity a explicitním výpočtem energie řeči a energie šumu

3.8.1 SNR na základě tvaru rozložení vzorků signálu

Nevýhodou tohoto způsobu je např. nezvládání rozeznání technických signálů (zvuky nejrůznějších přístrojů apod.). Technický signál totiž může mít také Gamma rozložení, SQE jej tedy chybně vyhodnotí jako kvalitní signál vhodný pro následující přepis [14]. Více informací lze najít v [39].

Tuto hodnotu najdeme v SQE pod názvem *waveform_snr*.

3.8.2 SNR na základě detekce řečové aktivity a explicitního výpočtu

Tento způsob vyžaduje použití externí Voice Activity Detection (VAD) technologie. Ta je schopna detekovat úseky řeči a úseky ticha. Následně se SNR vypočítá ze vztahu $SNR = 10\log(\frac{R}{T})$, kde R je energie řeči a T energie ticha.

Nevýhodou tohoto přístupu je nutná přítomnost úseků řeči i úseků ticha v nahrávce.

Tuto hodnotu najdeme v SQE pod názvem *wfilter_snr*.

3.9 Minimální a maximální absolutní hodnota

Máme-li všechny vzorky konečně dlouhého řečového signálu o délce N vzorků $(x_1, \ldots, x_N)$, pak

$$\hat{x}_{min} = min(|x_1|, \ldots, |x_N|) \tag{3.5}$$

a

$$\hat{x}_{max} = max(|x_1|, \ldots, |x_N|), \tag{3.6}$$

kde $\hat{x}_{min}$ je minimální absolutní hodnota signálu a $\hat{x}_{max}$ je maximální absolutní hodnota signálu.

Pro zpracování řečového signálu je ideální, aby hodnoty signálu pokrývaly celý dynamický rozsah. Pokud je maximální absolutní hodnota nízká, dochází k velkým numerickým chybám (viz. 3.2). Může být také indikátorem, že je řeč příliš tichá. Podle minimální absolutní hodnoty lze odhadnout úroveň šumu v signálu.

Metriky najdeme v SQE pod názvy *waveform_min_abs_value* a *waveform_max_abs_value*.

3.10 Minimální a maximální hodnota

Obdobně můžeme definovat i minimální a maximální hodnotu audia,

$$x_{min} = min(x_1, \ldots, x_N) \tag{3.7}$$

a

$$x_{max} = max(x_1, \ldots, x_N), \tag{3.8}$$

kde x_{min} je minimální hodnota signálu a x_{max} je maximální hodnota signálu.

Porovnáním těchto dvou hodnot můžeme odhadnout například posunutí nulové izolinie. Taktéž získáváme informaci o pokrytí dynamického rozsahu a indikaci tiché řeči.

V SQE jdou metriky uvedeny pod názvy *waveform_min_value* a *waveform_max_value*.

3.11 Délka ticha

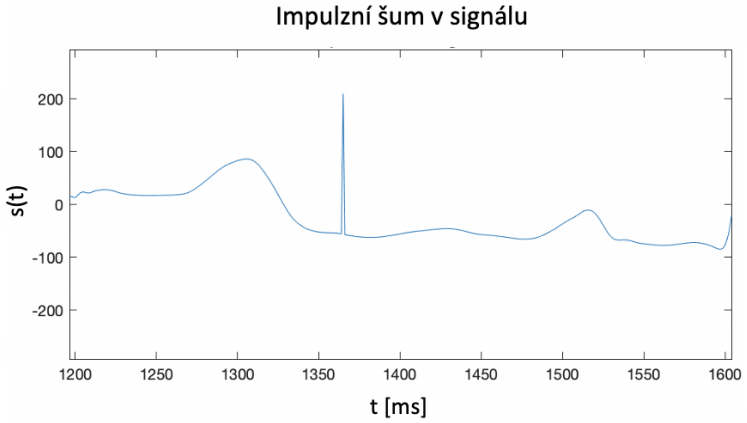
Pro přibližnou detekci úseků ticha je využíván energetický detektor řečové aktivity. Řečový signál je převeden na energetickou funkci počítanou s 25 ms oknem (každých 25 ms je vypočtena jedna hodnota energie). Podle ručně nastaveného prahu se zkoumá, zda je hodnota energie okna nižší než práh nebo ne. Pokud ano, daný úsek je označen jako ticho. Výsledná délka je tedy součtem všech takto označených úseků v sekundách.

Tato metrika je velmi užitečná, např. v nahrávkách z dlouhodobě odposlouchávajících zařízení se mohou objevovat až hodinové úseky ticha. Díky této metrice je možné je odstranit a zpracovávat mnohem kratší nahrávky, což šetří výpočetní sílu a tím pádem i čas.

V SQE nalezneme délku ticha jako *wfilter_silence_length*. S touto metrikou souvisí i metrika poměru ticha, který udává poměr mezi délkou ticha a délkou celého signálu. Je označena jako *wfilter_silence_ratio*.

3.12 Délka úseků obsahující impulsní šum

Cílem této metriky je odhalit přítomnost impulsního rušení k jeho následnému odstranění (např. mediánovým filtrem). Také nám dává informaci o tom, jak velká část signálu je tímto druhem šumu postižena. Řečový signál je procházen 25 ms oknem a v případě ojedinělé velmi vysoké či velmi nízké hodnoty je úsek označen jako úsek obsahující impulsní šum. Celková délka je pak součtem všech těchto úseků v sekundách. Ukázka rušení impulsního typu na Obr. 3.6.



Obr. 3.6: Příklad impulsního rušení signálu

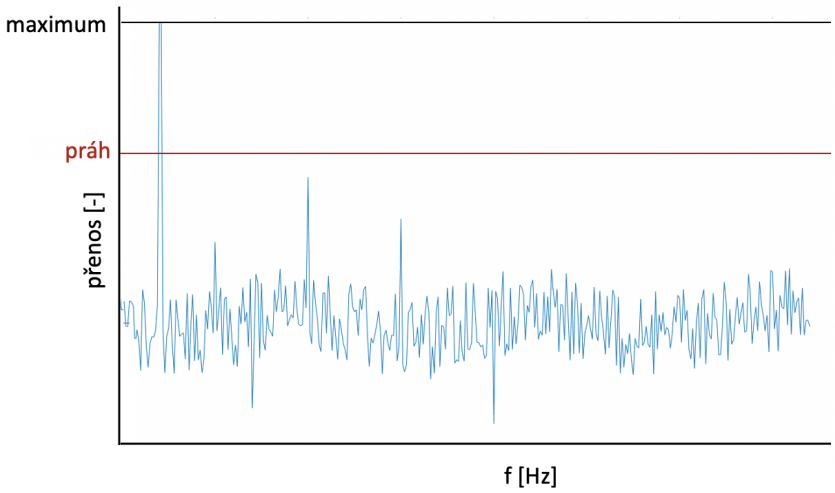
Podobně jako v předchozím odstavci najdeme v SQE celkovou délku pod parametrem *wfilter_intermittent_noise_length* a poměr této délky k celkové délce signálu

jako *wfilter_intermittent_noise_ratio*.

3.13 Délka technického signálu

Technický signál neboli zvuky vydávané přístroji, je jedním z běžných nežádoucích artefaktů vyskytujících se v řečových signálech. I přesto, že SQE využívá více různých přístupů k jeho detekci, právě zde často selhává.

Jedním z přístupů SQE k jeho detekci je hledání plochého spektra. Systémy s rozprostřeným spektrem mají přibližně stejnou amplitudu přes celé využívané spektrum řečového signálu (vyskytuje se např. u modemů). Dalším přístupem je hledání tónů, což jsou zvuky složené z jedné či více frekvencí, ve spektru tudíž výrazné. Tato metoda je založená na hledání ostrých maxim ve spektru a porovnání amplitudy nalezené frekvence s amplitudou okolních frekvencí, přičemž práh je závislý na maximální hodnotě ve spektru (viz Obr. 3.7).



Obr. 3.7: Příklad detekce tónu prahováním spektra

Třetím přístupem je hledání v čase neměnných spekter ve spektrogramu (každé spektrum patří jednomu 25 ms dlouhému úseku). Hlas je neměnný několik desítek ms, u delších úseků jde o umělý signál.

Pokud se některá z těchto vlastností spekter projeví, je úsek označen jako technický signál. Součet všech takto označených úseků je právě délka technického signálu.

V SQE je tato metrika označena jako *wfilter_technical_signal_length*. Stejně jako v předchozích případech je v SQE i metrika *wfilter_technical_signal_ratio* vyjadřující poměr délky technického signálu a celkové délky signálu.

3.14 Délka signálu určeného k odfiltrování

Součet všech nežádoucích úseků, tj. úseků technického signálu, impulsního šumu a ticha je označen jako délka signálu určeného k odfiltrování (v SQE *wfilter_filtered_length*), poměr délky nežádoucích úseků a celkové délky signálu jako *wfilter_filtered_ratio*.

3.15 Délka nahrávky

Poslední metrikou spíše informačního charakteru je délka nahrávky v sekundách, v SQE označená jako *waveform_length*. Ta nám může pomoci např. odstranit příliš krátké nahrávky, u kterých se nedá čekat rozumný informační přínos.

4 Návrh metriky obsahové bohatosti audia

Metriky signálové kvality popsané v kapitole 3 mohou selhávat u jistých typů (např. technických) signálů a označit je jako vhodné pro následný přepis do textu. Cílem je vyvarovat se nesmyslných výsledků. V této kapitole tedy navrhnu metriky obsahové bohatosti audia. Obsahovou bohatostí se rozumí informační přínos dané řečové nahrávky. Pro převod řečové nahrávky na fonetický přepis využíváme nástroj Phonexia Phoneme Recognizer (viz. kapitola 1.4.2).

Nyní tedy můžeme začít zkoumat zastoupení fonémů v audiu, jejich rozložení a další parametry obsahové bohatosti. Výše jsme zmiňovali unikátnost fonémů pro daný jazyk. V této práci pracuji s datovou sadou nahrávek z českého call-centra. Prvním krokem je tedy znalost všech fonémů vyskytujících se v češtině (seznam fonémů byl poskytnut firmou Phonexia):

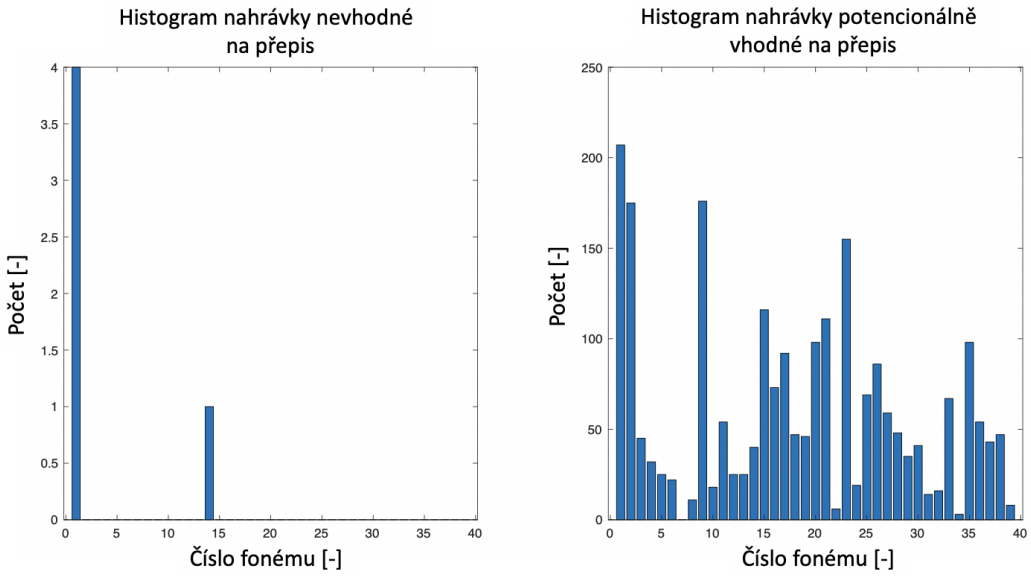
```
sil, @, C, D, N, R, S ,T, Z, a, a:, b ,c ,d ,e, e:, f, g,
h, i, i:, j, k, l, m, n, o, o:, p, r, s, t, u, u:, unk,
v, x, z
```

Jak vypadá takový fonetický přepis řečové nahrávky můžeme vidět na Obr. 6.3 (*sil* je fonémová značka pro ticho (angl. *silence*), v jednom řádku je vždy posloupnost fonémů mezi dvěma značkami ticha, fonémová značka *unk* (angl. *unknown*) značí neznámou). Často tedy jeden řádek představuje např. slovo či větu.

Dále se v této kapitole budeme dívat na histogramy fonémů a pravděpodobnostní rozložení fonémů v nahrávkách. S jejich využitím následně navrhneme metriky hodnotící obsahovou bohatost jakožto metriku obsahové kvality.

4.1 Histogram fonémů

Prvním ukazatelem obsahové bohatosti může být histogram fonémů dané audionahrávky. Ten nám říká, kolikrát se určitý foném v nahrávce vyskytl. Na Obr. 4.1 vlevo vidíme histogram audionahrávky, u které nemůžeme očekávat smysluplný přepis do textu. V jeho fonémovém přepisu se vyskytuje čtyřikrát značka pro ticho a jen jeden jiný foném. Vpravo je poté fonémový přepis audionahrávky, u které je větší pravděpodobnost smysluplného přepisu do textové podoby vzhledem k přítomnosti velké většiny fonémů. To samo o sobě nemusí nutně vypovídat o dobrém přepisu, ale ve vztahu k délce nahrávky lze odvodit minimálně přítomnost rozumného množství fonémů k přepisu.



Obr. 4.1: Příklad histogramů výskytu fonému v nahrávkách - vlevo je znázorněna nahrávka 015.*wav*, která je nevhodná na přepis a vpravo je nahrávka 018.*wav*, která je potencionálně vhodná na přepis.

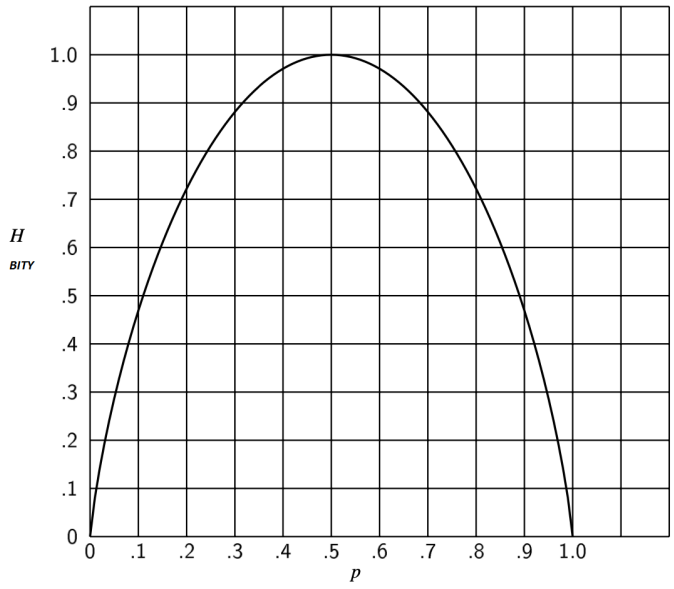
4.2 Entropie

Entropie (zde míněno informační entropie) vychází z teorie informace. Jde o způsob vyjádření, kolik informace nám dané zastoupení fonémů v audionahrávce poskytuje. Nechť máme set n možných událostí (v našem případě výskyt daného fonému) a každá událost nastane s pravděpodobností p_i, kde $i \in \{1, \ldots, n\}$. Tyto pravděpodobnosti výskytu jsou nám známé [32]. Entropii pak vypočítáme jako

$$H = -\sum_{i=1}^{n} p_i \log p_i, \tag{4.1}$$

kde H je entropie, n je počet fonémů a p_i je pravděpodobnost výskytu i-tého fonému.

Entropie je udávána v *bitech*. Příklad entropie pro soubor obsahující dva fonémy s pravděpodobnostmi výskytu p_1 a $p_2 = 1 - p_1$ je zobrazen na Obr. 4.2 [32].



Obr. 4.2: Entropie pro dva fonémy s pravděpodobnostmi výskytu p_1 a $1 - p_1$ [32]

Entropie je rovna 0 tehdy a jen tehdy, když jsou všechny p_i kromě jedné rovny nule, jinými slovy pokud by existovala možnost výskytu jen jednoho fonému v souboru, entropie by byla nulová, protože by takto sestavená nahrávka nepřinášela žádnou informaci [32].

Entropie je maximální (a rovna $\log n$), pokud je pravděpodobnost výskytu fonémů stejná, tedy pokud pro všechny $p_i = \frac{1}{n}$. V tomto případě máme největší nejistotu náhodného rozdělení, tudíž daná posloupnost fonémů přináší nejvíce informace [32]. Pro výpočet entropie dané audionahrávky potřebujeme znát pravděpodobnosti výskytů všech fonémů daného jazyka. Existují výzkumy o obrovských objemech dat, které tyto pravděpodobnosti vyčíslují [41], jejich databáze fonémů se ovšem trochu liší od mé, potřebuji si tedy pravděpodobnosti vypočítat z dostupných dat.

Výpočet provedu z datové sady českého call-centra. Jde o 680 nahrávek v češtině a délky nahrávek se pohybují od 100 sekund po 16 minut. Jde tedy o dostatečně velkou datovou sadu pro referenční pravděpodobnostní rozložení fonémů. Pravděpodobnosti výskytu jednotlivých fonémů tedy vypočítáme jako

$$p_i = \frac{\sum_{j=1}^{680} n_{ij}}{n_{celkem}}, \tag{4.2}$$

kde p_i je pravděpodobnost výskytu i-tého fonému, n_{ij} počet výskytů i-tého fonému v j-té nahrávce a n_{celkem} je součet výskytů všech fonémů ve všech nahrávkách.

4.3 Křížová entropie

Entropie je vhodná pro porovnání různých jazyků, slouží jako indikátor toho, kolik každý foném nese průměrně v konkrétním jazyku informace. My ovšem potřebujeme vyjádřit, jak velkou informaci máme v jedné konkrétní nahrávce. K tomu můžeme využít *křížovou entropii*. Křížová entropie měří relativní entropii mezi dvěma pravděpodobnostními rozloženími (za předpokladu, že se v nich vyskytují stejné události, tedy u nás stejné fonémy). Pro každou nahrávku tedy můžeme tedy vypočítat relativní entropii mezi pravděpodobnostním rozložením fonémů v dané nahrávce a celkovým pravděpodobnostním rozložením fonémů ve všech nahrávkách [25].

Pravděpodobnostní rozložení fonémů v dané nahrávce vypočítáme jako:

$$q_i = \frac{n_i}{n_{celkem}}, \tag{4.3}$$

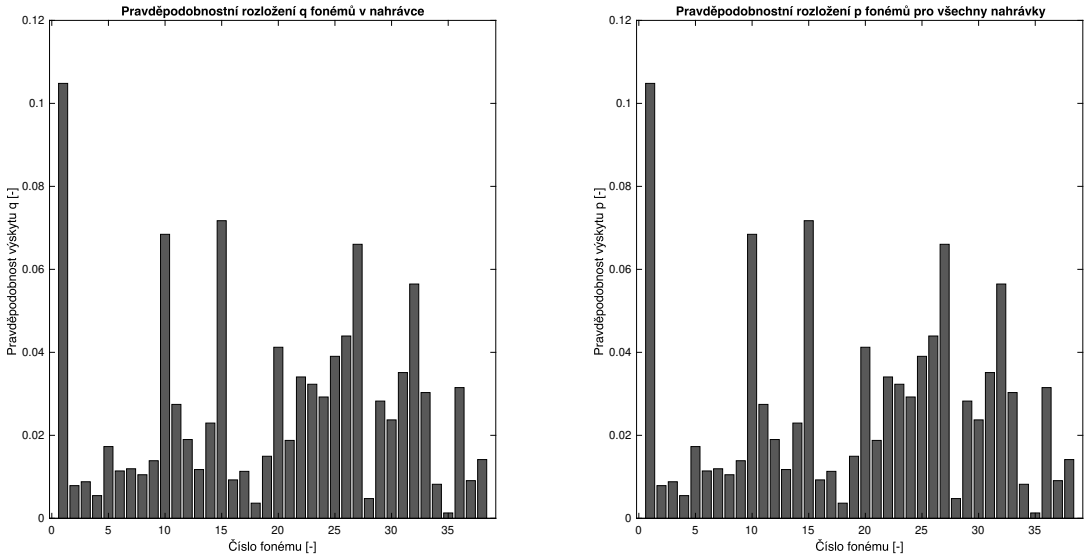
kde q_i je pravděpodobnost výskytu i-tého fonému v nahrávce, n_i počet výskytů i-tého fonému v nahrávce a n_{celkem} je součet výskytů všech fonémů v nahrávce.

Křížovou entropii pak získáme:

$$H(p, q) = -\sum_{i=1}^{n} p_i \log q_i, \tag{4.4}$$

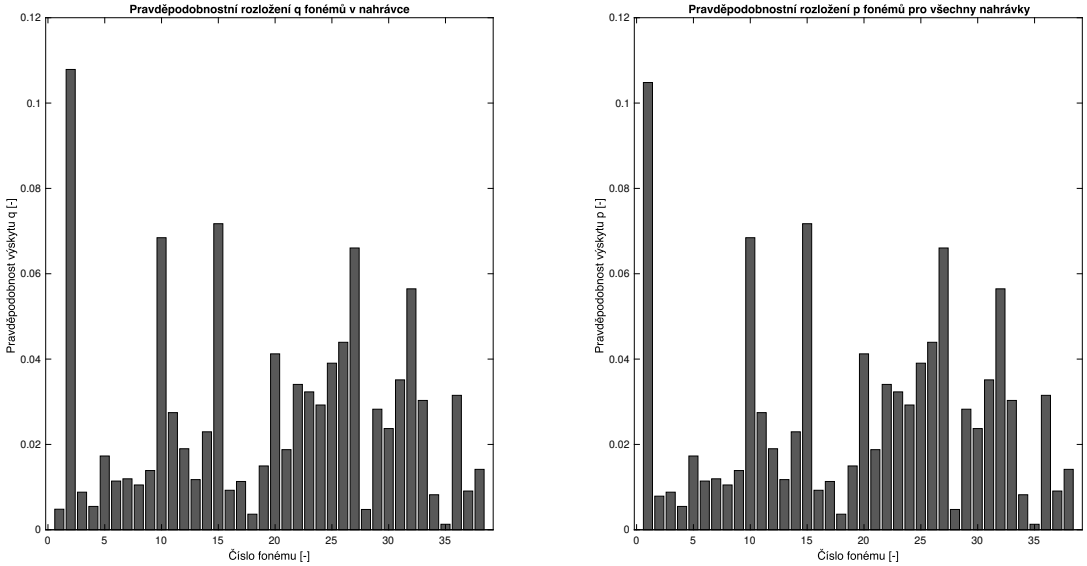
kde $H(p, q)$ je křížová entropie pravděpodobnostních rozložení p a q, n je počet fonémů, p_i je pravděpodobnost výskytu i-tého fonému celkového pravděpodobnostního rozložení a q_i je pravděpodobnost výskytu i-tého fonému pravděpodobnostního rozložení nahrávky [25].

Křížová entropie je nejmenší, je-li pravděpodobnostní rozložení fonémů v nahrávce shodné s celkovým pravděpodobnostním rozložením:



Obr. 4.3: Shodná pravděpodobnostní rozložení

Pro toto rozložení je $H(p, q) = 1,44$. Stačí jakákoliv drobná změna v q (všimněte si změny výšek prvního a druhého sloupce v rozložení q):



Obr. 4.4: Rozdílná pravděpodobnostní rozložení

a křížová entropie se zvýší, pro toto rozložení vychází $H(p, q) = 1,58$. Pokud tedy bude nahrávka obsahovat málo fonémů nebo bude její pravděpodobnostní rozdělení úplně odlišné od celkového pravděpodobnostního rozdělení, křížová entropie bude vysoká.

Pro použití na reálné sadě dat je nutné ošetřit zvláštní situace, které mohou nastat. První takovou situací je, pokud $q_i = 0$, jinými slovy pokud i-tý foném nebude

v dané nahrávce vůbec přítomen.

Pak $log(q_i) = -\infty$ a tím pádem $H = \infty$. Takovou situaci vyřešíme adaptací pravděpodobností q_i podle pravděpodobností p_i vztahem

$$q'_i = \alpha p_i + (1 - \alpha) q_i, \tag{4.5}$$

kde q'_i je upravená pravděpodobnost výskytu i-tého fonému, p_i je pravděpodobnost výskytu i-tého fonému celkového pravděpodobnostního rozložení, q_i je pravděpodobnost výskytu i-tého fonému pravděpodobnostního rozložení nahrávky a α je koeficient upravení.

Volbou parametru α určujeme náklonnost adaptované hodnoty k p_i a q_i.

Druhou situací, kterou je potřeba ošetřit, je situace kdy v nahrávce není přítomen žádný foném. Pak $q_i = \frac{n_i}{0}$, což je neplatný výraz. Situaci vyřešíme stejně jako v předchozím případě upravením pravděpodobností q podle referenčního pravděpodobnostního rozložení p, podobným vztahem

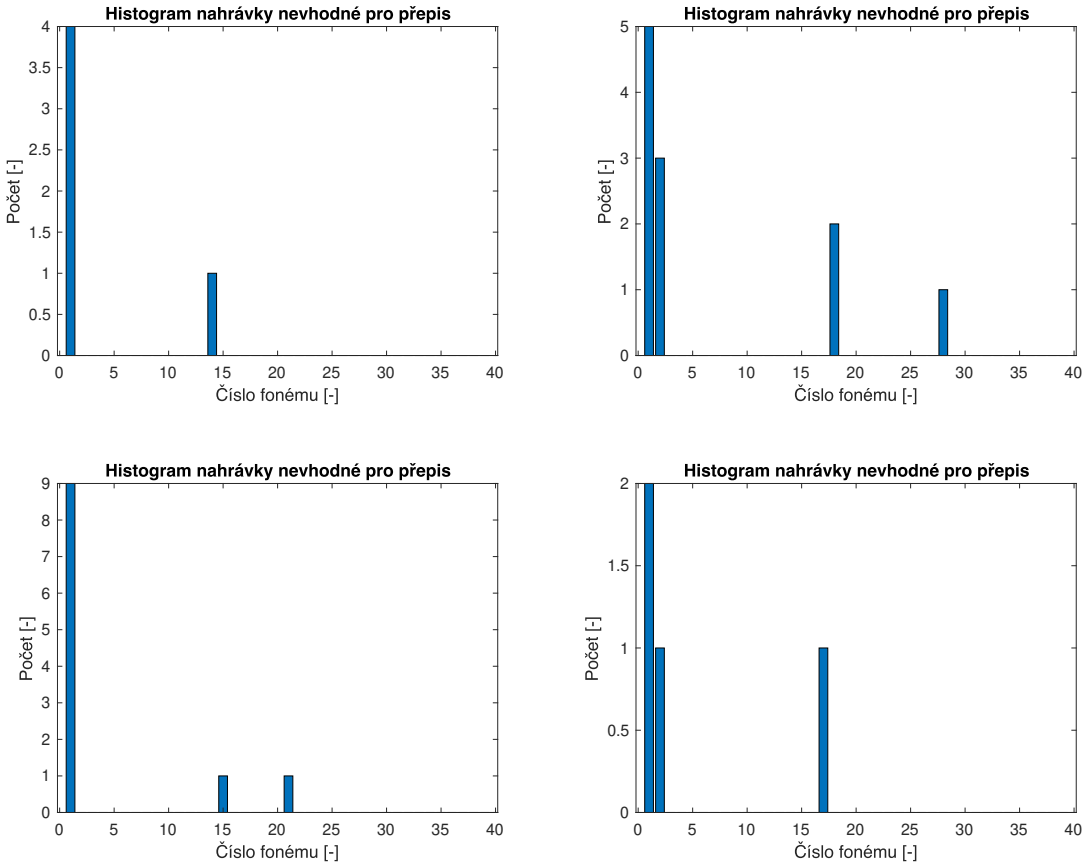
$$q'_i = \beta p_i + (1 - \beta) q_i = \beta p_i, \tag{4.6}$$

kde q' je upravené pravděpodobnostní rozložení dané nahrávky, p je celkové pravděpodobnostní rozložení, q je pravděpodobnostního rozložení nahrávky a β je koeficient upravení.

Třetí situací je případ, kdy $p_i = 0$. To znamená, že v celé datové sadě není ani jednou přítomen i-tý foném. Tato situace nastává např. při použití datových sad obsahující hluky (není vůbec přítomna řeč). Jelikož $p_i = 0$, pak nutně i $q_i = 0$ a dle 4.5 i $q'_i = \alpha 0 + (1 - \alpha)0 = 0$. Při takové situaci jednoduše odstraníme p_i i q_i z p a q a dál počítáme klasickým způsobem s upravenými rozloženími.

4.4 Počet unikátních fonémů

V kapitole 4.1 jsme se podívali na histogram jednotlivých fonémů nahrávek jakožto ukazatel vhodnosti pro přepis. V tento moment se na něj lze dívat jako na subjektivní ukazatel. Jednou z možností jak z něj kvantifikovat informace může být např. počet unikátních fonémů. Říká nám, kolik různých fonémů se v nahrávce objevuje. Na Obr. 4.5 je ukázka čtyř různých nahrávek s malým počtem unikátních fonémů.



Obr. 4.5: Histogramy fonémů nahrávek nevhodných pro přepis - vlevo nahoře vidíme výskyt dvou unikátních fonémů, vpravo nahoře výskyt čtyř unikátních fonémů, vlevo dole tří unikátních fonémů a vpravo dole taky tří unikátních fonémů.

Žádná z těchto nahrávek neobsahuje řeč, tyto nahrávky jsou výrazně nevhodné pro přepis. V české datové sadě je celkem 39 fonémů a drtivá většina nahrávek má počet unikátních fonémů větší než 30. V této datové sadě můžeme bezpečně říci, že pokud má nahrávka méně než 5 unikátních fonémů, je nevhodná pro přepis. Problém může nastat, pokud budeme pracovat s krátkými nahrávkami obsahujícími řeč. Konkrétně třeba s jednoslovnými pokyny typu „ne“. U zpracování databází nahrávek se ovšem dá předpokládat, že velmi krátké nahrávky s pokyny přítomny nebudou.

5 Statistický model

V kapitole 3 jsme představili metriky signálové kvality a v kapitole 4 jsme zavedli metriky obsahové bohatosti. Stěžejní náplní praktické části této práce je tvorba statistického modelu, který na základě těchto metrik dokáže předpovědět přesnost strojového přepisu dané nahrávky. Pro vytvoření tohoto modelu použijeme regresní metody strojového učení. Na začátek představíme datové sady, které budeme pro tvorbu modelu používat a následně zavedeme metriky hodnocení přesnosti strojového přepisu, které nám budou sloužit jako referenční hodnoty statistického modelu. Velmi důležitou částí tvorby statistického modelu je příprava vstupních dat, která se skládá z výběru vhodných metrik a normalizace. Nakonec představíme jednotlivé regresní metody a vysvětlíme si jak rozdělit datovou sadu na trénovací a testovací část pomocí k-křížové validace.

5.1 Použité datové sady

V této diplomové práci využívám dvě různé datové sady nahrávek.

První z nich je sada pořízená od českého kontaktního centra. Jde o 680 nahrávek. Původně šlo o 340 nahrávek formou dialogu, každý dialog se poté rozdělil na dvě nahrávky. Nahrávky byly ručně segmentované pro separaci řečníků a každá nahrávka díky tomu obsahuje jen jednoho řečníka. V nahrávkách jsou běžné konverzace a volná řeč (tedy nejsou to nahrávky čtení psaného textu). Délka nahrávek se pohybuje zhruba od 100 sekund po 16 minut. Vzorkovací frekvence nahrávek je 8 kHz. Tento datový set jsme zvolili vzhledem k faktu, že jsou k němu dostupné ručně dělané přepisy řeči do textové podoby, což je nutný předpoklad k ohodnocení přesnosti přepisu metrikou Word Error Rate (více v kapitole 5.2.1).

Druhou datovou sadou je sbírka 251 nahrávek zvuků neobsahující žádnou řeč. Nahrávky jsou velice krátké, zhruba 2 až 25 sekund. Tyto nahrávky jsou typickým příkladem nahrávek nevhodných pro přepis. Jejich přítomnost ve statistickém modelu nám umožní rozlišit mezi nahrávkou plnou řeči a nahrávkou bez řeči.

Tyto datové sady dále rozdělíme na trénovací a testovací sadu (více v kapitole 5.6).

Obě datové sady byly poskytnuty firmou Phonexia a mají pouze evaluační licenci, není tedy možné tato data dále poskytnout.

5.2 Vyhodnocení přesnosti přepisu řeči na text

V předchozích kapitolách byly probrány jak signálové metriky, tak metriky obsahové bohatosti jako nástroje pro určení kvality nahrávky vzhledem k následnému přepisu řeči v nahrávce na text. V diplomové práci budou využity metody strojového učení pro zjištění důležitosti jednotlivých metrik a k nastavení jejich vah pro predikci přesnosti automatického přepisu.

Aby tohle bylo možné, musíme připravit dvě sady dat, trénovací a testovací. V trénovací sadě je nutné mít referenční metriku, které bude hodnotit přesnost automatického přepisu nahrávek. Tou bude tzv. Word Error Rate (WER). Pro každou nahrávku tedy budeme mít sadu vygenerovaných metrik (signálových i obsahových) a budeme znát přesnost strojového přepisu dané nahrávky do textu díky WER. Na těchto datech natrénujeme statistický model pro predikci WER.

V testovací sadě poté využijeme vygenerované metriky, statistický model a budeme predikovat přesnost přepisu nahrávek z testovací sady v závislosti na daných metrikách. Tyto predikované hodnoty poté porovnáme s vypočítaným WER, jako s referenční hodnotou přesnosti.

5.2.1 Word Error Rate

Word Error Rate je standardní metrika pro hodnocení přesnosti automatického přepisu řečového signálu na text. Rozlišuje tři základní druhy chyb - substituci, vypuštění a vložení slov [30].

1. Substitucí rozumíme nahrazení slova jiným slovem, např. když původní větu:

   `Dnes je krásný den.`

   přepíše technologie automatického přepisu na:

   `Dnes je krásný len.`

2. Vypuštěním rozumíme vynechání slova z původní věty, např.:

   `Dnes je krásný den.`

   přepíše technologie automatického přepisu na:

   `Dnes krásný den.`

3. Vložením rozumíme vložení slova do původní věty, např.:

   `Dnes je krásný den.`

   přepíše technologie automatického přepisu na:

   `Dnes je krásný no den.`

WER vypočítáme jako

$$WER = \frac{S + D + I}{N} \cdot 100, \tag{5.1}$$

kde WER je hodnota WER v procentech, S je počet substitucí, D počet vypuštění, I počet vložení a N je počet slov v referenční větě [40].

Pro výpočet WER využívám *NIST SClite* skórovací nástroj [23]. Tento program vrací jak počet substitucí, vložení a vypuštění, tak počet slov referenčního přepisu. Tyto hodnoty načítám do prostředí Matlab a používám vzorec 5.1.

Netradiční je možnost dosáhnutí hodnoty $WER > 100\%$. Z tohoto důvodu se někdy využívají jiné metody, např. *correctness* (kapitola 5.2.2).

V datové sadě obsahující zvuky jsem narazil na zajímavý problém. Pokud nahrávka neobsahuje řeč, není v referenčním přepisu ani jedno slovo a tím pádem $N = 0$. Dále $WER = \frac{S+D+I}{0} \cdot 100$, což je neplatný výraz. Pro použití takových nahrávek v trénovací sadě je ovšem nutné tuto metriku přizpůsobit tak, abychom měli referenční hodnotu přesnosti přepisu.

Odvodím tedy metriku WER_{zvuky}, kterou budu moci použít ve statistickém modelu. Při odvození budu vycházet ze skutečnosti, že v české datové sadě se velmi kvalitní přepisy pohybují kolem 20% WER. Naopak velmi špatné přepisy se pohybují kolem 150% WER. Pokud mám nahrávku neobsahující řeč a strojový přepis vyhodnotí prázdný řetězec, považuji to za úspěšný přepis a jeho hodnota by měla odpovídat kvalitním přepisům, tj. hodnotám kolem 20% WER. U takových nahrávek není potřeba uvažovat substituce ani vypuštění, jediné co nás bude zajímat je počet vložení. Počet vložení v této datové sadě nabývá pouze hodnot $0, 1, 2, 3$ a 5. Pro 0 bude tedy $WER_{zvuky} = 20$, přesnost přepisu se pak lineárně zhoršuje až na maximální hodnotu 5 vložení. Tato hodnota by měla odpovídat velmi špatnému přepisu, tj. hodnotě kolem 150% WER. Vytvořím tedy mapovací funkci

$$WER_{zvuky}(I) = 26I + 20, \tag{5.2}$$

kde WER_{zvuky} je hodnota přesnosti přepisu pro nahrávky bez řeči a I je počet vložení.

Velmi související metrikou, se kterou se lze v oblasti vyhodnocení přesnosti strojového přepisu setkat je tzv. Word Accuracy (WAC). Vypočítat ji lze jako:

$$WAC = 1 - WER, \tag{5.3}$$

kde WAC značí Word Accuracy a WER značí Word Error Rate.

5.2.2 Correctness

Další rozšířenou metrikou přesnosti strojového přepisu je *correctness*. Correctness se na rozdíl od WER nebo WAC vůbec nedívá na vložená slova, zajímá se jen o substituce nebo vypuštěná slova. Tento přístup může být praktičtější, protože hodnota correctness leží vždy mezi 0 - 100%. Lehce se tedy převádí na rozsah od 0 do 1. Correctness vypočítáme podle vzorce:

$$correctness = \frac{1 - (S + D)}{N} \cdot 100, \tag{5.4}$$

kde *correctness* je hodnota v procentech, S je počet substitucí, D počet vypuštění a N je počet slov v referenční větě.

5.3 Příprava dat pro statistický model

K predikci přesnosti strojového přepisu využiji metod strojového učení. Strojové učení se používá například pro:

- klasifikaci dat - zařazení objektů do předem definovaných tříd
- regresi - predikci hodnot spojité proměnné na základě spojitých či kategorických vstupů
- shlukování - vytváření skupin objektů na základě jejich podobnosti či rozdílnosti, na rozdíl od klasifikace nejsou předem definované třídy ani jejich počet

V našem případě je potřeba na základě jednotlivých metrik audionahrávek predikovat přesnost strojového přepisu, využijeme tedy regresních metod.

5.3.1 Výběr vhodných metrik

Ne všechny parametry nahrávek, které SQE (kapitola 1.4.1) poskytne, jsou vhodné pro náš statistický model. První skupinou takových parametrů jsou neměnné parametry pro všechny nahrávky, mezi ty patří:

- vzorkovací frekvence
- práh pro určení přebuzení signálu
- počet kvantovacích bitů (a tedy zároveň i počet kvantovacích hladin)
- délka úseků obsahujících impulsní šum

Vzorkovací frekvence, délka úseků obsahující impulsní šum a počet kvantovacích bitů (resp. hladin) se obecně lišit mohou, v této datové sadě jsou však konstantní.

Druhou skupinou jsou duplicitní parametry vyjádřené jednou jako absolutní hodnota a podruhé jako relativní hodnota:

- délka ticha
- délka úseků obsahující impulsní šum

- délka technického signálu
- délka signálu určeného k odfiltrování

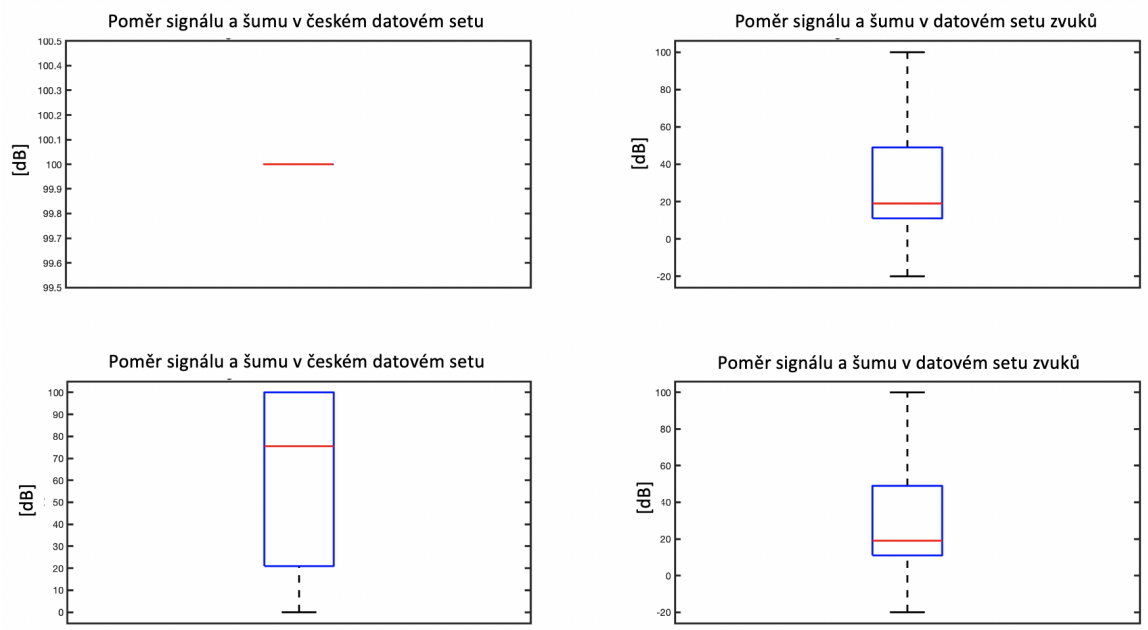
U dalších metrik se podíváme na výskyt jejich hodnot v obou datových sadách a rozhodneme, zda jejich použití může být pro statistický model užitečné či ne.

Pro zobrazení rozložení hodnot daných metrik jsem zvolil tzv. *krabicový graf* (angl. boxplot). Příklad krabicového grafu můžete vidět na Obr. 5.1. U každé metriky zobrazuji vždy dvě rozložení hodnot, vlevo pro českou datovou sadu a vpravo pro datovou sadu obsahující zvuky. V každém grafu je červeně vyznačena hodnota mediánu metriky. Dolní a horní okraje modrého obdélníku vyznačují 50% výskytu všech hodnot. Hranice mimo obdélník zaznačují maximální a minimální hodnoty metrik (kromě hodnot, které jsou považovány za velmi odlehlé, ty jsou označeny červenými křížky).

Poměr signálu a šumu na základě tvaru rozložení vzorků signálu je v praxi pravděpodobně nejdůležitějším ukazatelem signálové kvality nahrávky vzhledem k strojovému přepisu. Jak ale vidíme na Obr. 5.1 (v krabicovém grafu vlevo nahoře), celá česká datová sada je natolik kvalitně připravená, že hodnoty poměru signálu a šumu jsou maximální, tj. 100 dB. Pro přenositelnost výsledků do reálného prostředí tedy bylo potřeba do nahrávek uměle přidat šum. Postup pro umělé zašumění této sady byl následující:

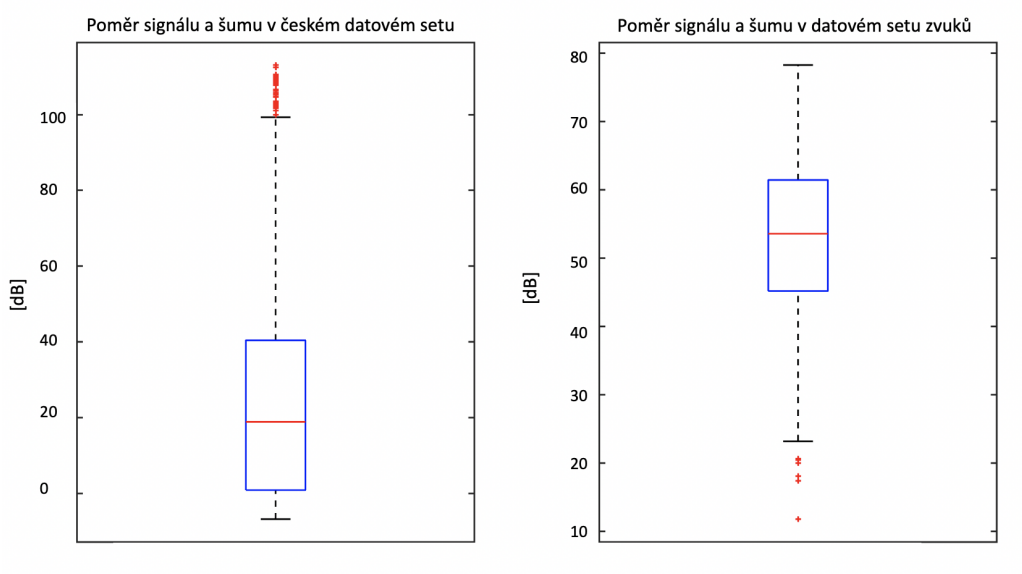
1. rozdělil jsem nahrávky na deset různých skupin (každá skupina obsahovala 68 nahrávek)
2. vytvořil jsem devět úrovní šumů (jedna skupina zůstane nezašuměná) - k tomu jsem použil reálné nahrávky šumu
3. ke každé skupině jsem přiřadil jednu úroveň šumu a následně do všech nahrávek z dané skupiny tuto úroveň šumu přidal

Pro co největší přiblížení realitě jsem vybral reálné šumy z každodenního života. Standardní umělé šumy jako Gaussovský či bílý šum by neodpovídaly reálnému použití, jelikož se v nahrávkách zpracovávaných řečovými technologiemi prakticky nevyskytují. Šlo o šumy obsahující běžné zvuky při pohybu městem, pobytu v restauraci či obchodu. Tyto šumy mi poskytla firma Phonexia pod evaluační licencí. Upravené rozložení poměru signálu a šumu vidíme na Obr. 5.1 (vlevo dole).



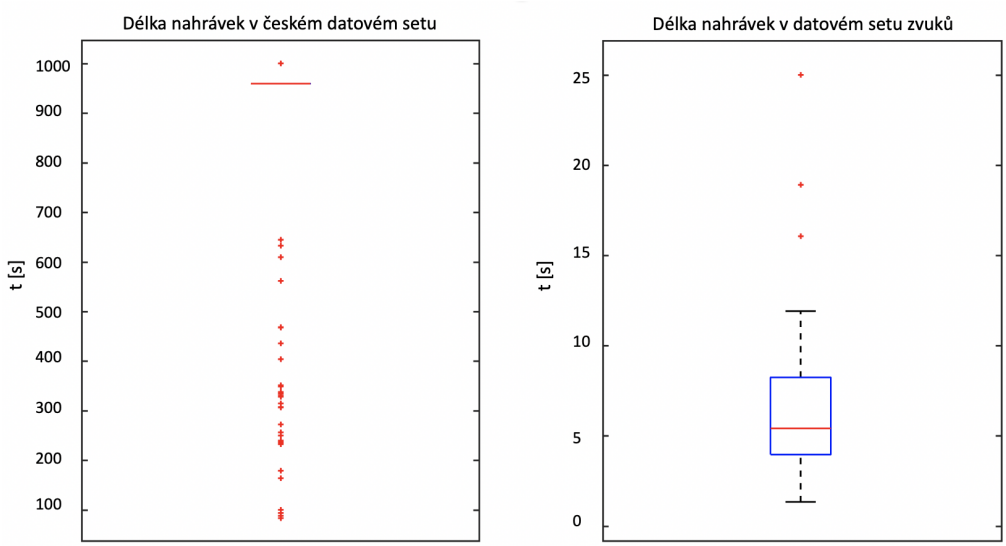
Obr. 5.1: Poměr signálu a šumu na základě tvaru rozložení vzorků signálu pro obě datové sady - vlevo nahoře vidíme rozložení pro původní českou sadu, vlevo dole rozložení pro upravenou českou sadu a na pravé straně je rozložení pro datovou sadu zvuků (pro ilustraci, že byla zašuměna jen česká sada)

Poměr signálu a šumu na základě detekce řečové aktivity a explicitního výpočtu (Obr. 5.2) bude velmi nepřesný zejména pro datovou sadu obsahující zvuky, jelikož pro jeho výpočet vyžaduje přítomnost jak řeči tak ticha a v ani jedné nahrávce z této datové sady se řeč nevyskytuje.



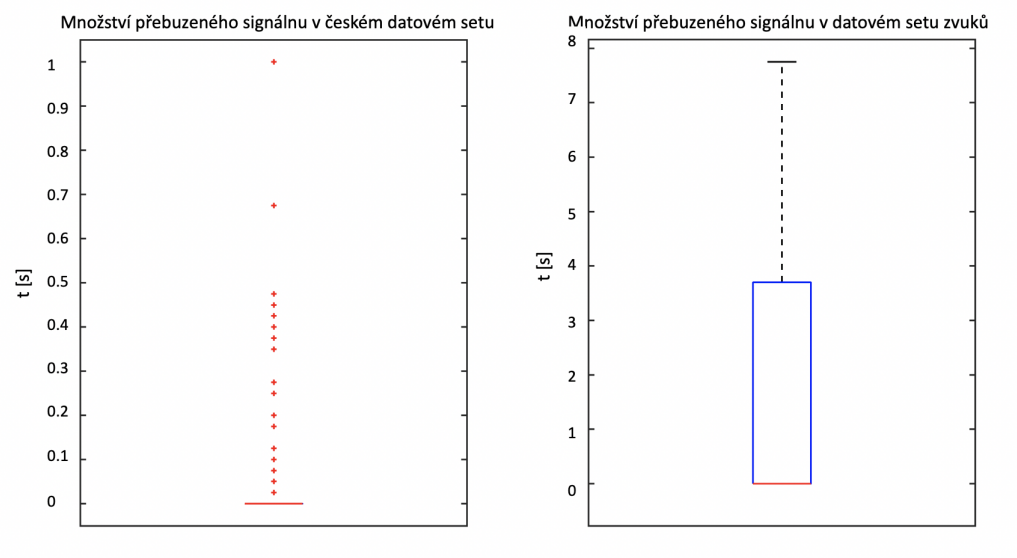
Obr. 5.2: Poměr signálu a šumu na základě detekce řečové aktivity a explicitního výpočtu pro obě datové sady

Délka nahrávky nemá obecně na přesnost přepisu vliv. V našem případě by její použití bylo nežádoucí vzhledem k faktu, že všechny nahrávky z datové sady zvuků jsou velmi krátké a naopak nahrávky z české datové sady jsou dlouhé (Obr. 5.3). Mohlo by se tedy stát, že tato metrika bude mít ve statistickém modelu velkou váhu, nebude ovšem využitelná pro jiné datové sady, jelikož v reálném prostředí se velmi často vyskytují dlouhé nahrávky neobsahující žádnou řeč.



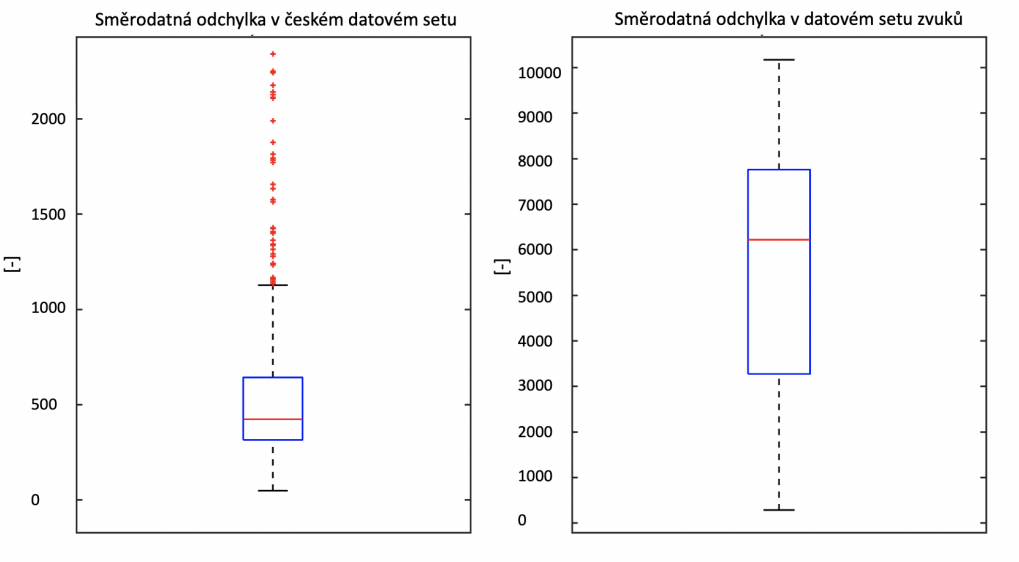
Obr. 5.3: Délky nahrávek pro obě datové sady

Na Obr. 5.4 vidíme, že v české datové sadě nejsou prakticky žádné přebuzené úseky. Naopak v datové sadě zvuků je přebuzení výrazné, zvlášť uvážíme-li velmi krátké délky těchto nahrávek (viz Obr. 5.3).



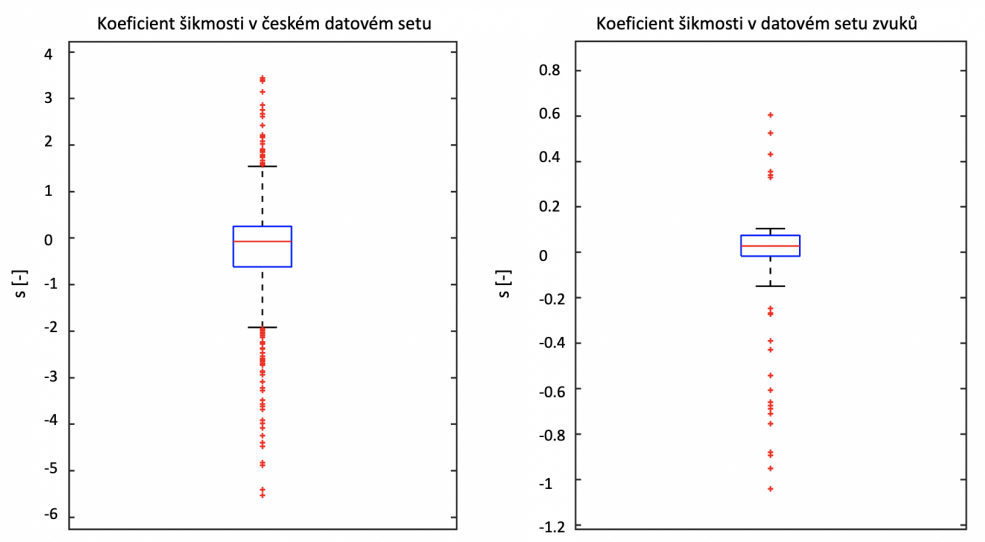
Obr. 5.4: Množství přebuzeného signálu pro obě datové sady

Směrodatné odchylky pro českou datovou sadu nabývají dle Obr. 5.5 výrazně menších hodnot než pro datovou sadu zvuků.



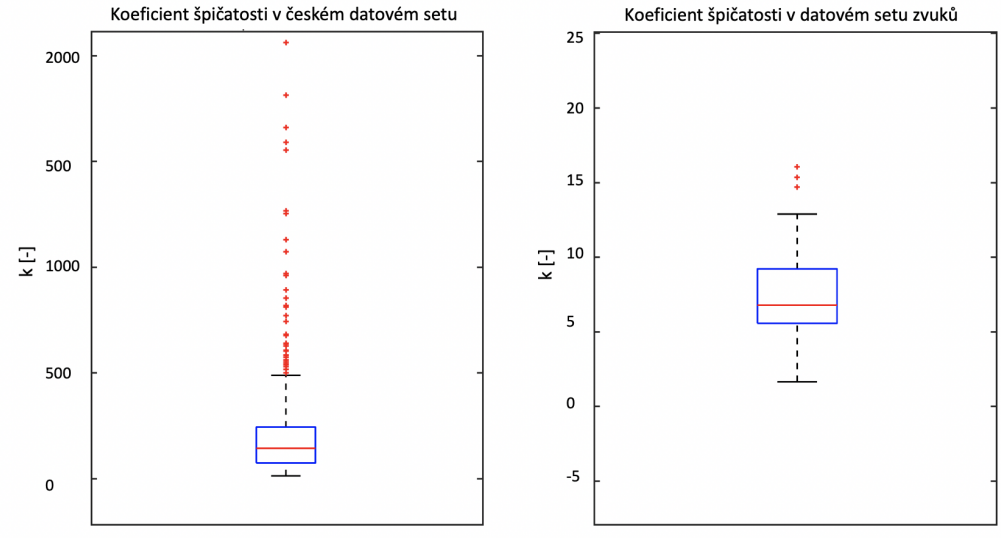
Obr. 5.5: Střední hodnoty pro obě datové sady

Koeficient šikmosti má ve většině případů v obou datových sadách hodnoty blízké nule (Obr. 5.6), neočekávám tedy, že bude ve statistickém modelu hrát významnou roli.



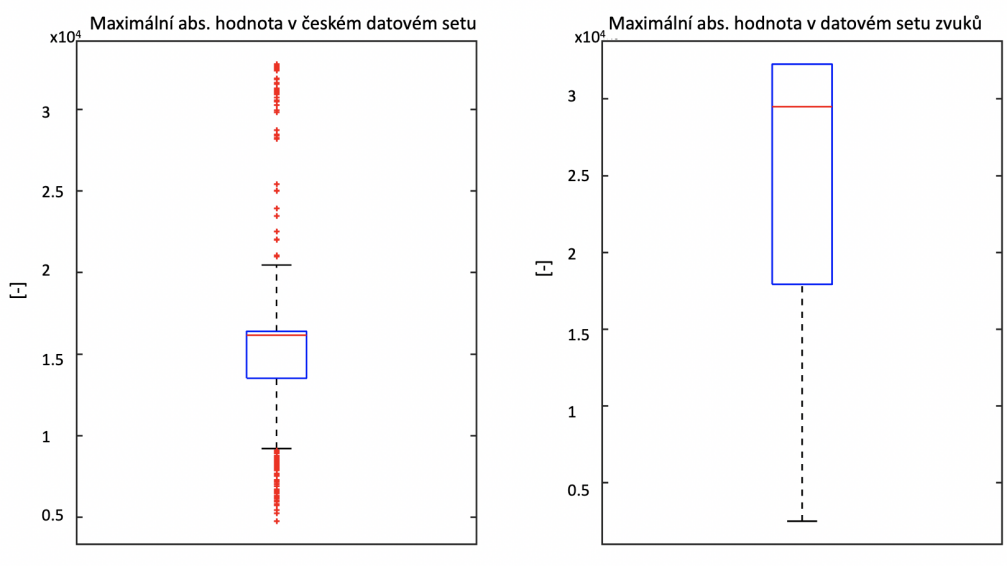
Obr. 5.6: Koeficient šikmosti pro obě datové sady

Koeficient špičatosti má u datové sady obsahující zvuky hodnoty velmi blízké nule, u české datové sady ovšem obsahuje hodnoty v řádu stovek (viz Obr. 5.7).

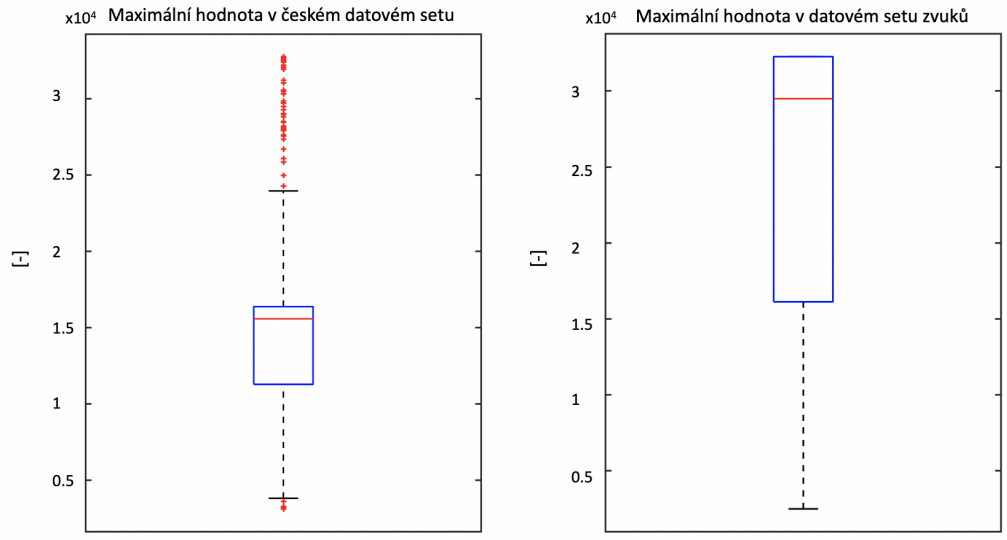


Obr. 5.7: Koeficient špičatosti pro obě datové sady

Maximální absolutní hodnoty (Obr. 5.8) a maximální hodnoty (Obr. 5.9) mají velký rozsah. Vzhledem ke stejnému počtu kvantovacích hladin ve všech nahrávkách bychom ve statistickém modelu mohli hledat souvislost mezi hlasitostí nahrávek a přesností přepisu. Na první pohled jde vidět, že rozložení hodnot těchto dvou metrik je velmi podobné, podíváme se tedy, zda jsou hodnoty navzájem silně korelované, nebo je to jen náhoda.



Obr. 5.8: Maximální absolutní hodnoty pro obě datové sady



Obr. 5.9: Maximální hodnoty pro obě datové sady

Pro výpočet vzájemné korelace využiji *Pearsonův korelační koeficient*. Hodnoty koeficientu se nachází v intervalu $< -1, 1 >$, přičemž kladné hodnoty vyjadřují kladnou lineární korelaci a záporné hodnoty zápornou lineární korelaci. Čím vyšší je absolutní hodnota tohoto koeficientu, tím vyšší korelace mezi dvěma metrikami je. Extrémem jsou hodnoty 1 a -1, které znamenají dokonalou korelaci [4].

Pro lepší představu lze využít tzv. *Evansovu příručku* navrženou pro absolutní hodnotu koeficientu [10]:

- 0,00 - 0,19 . . . velmi slabá
- 0,20 - 0,39 . . . slabá
- 0,40 - 0,59 . . . střední
- 0,60 - 0,79 . . . silná
- 0,80 - 1,00 . . . velmi silná

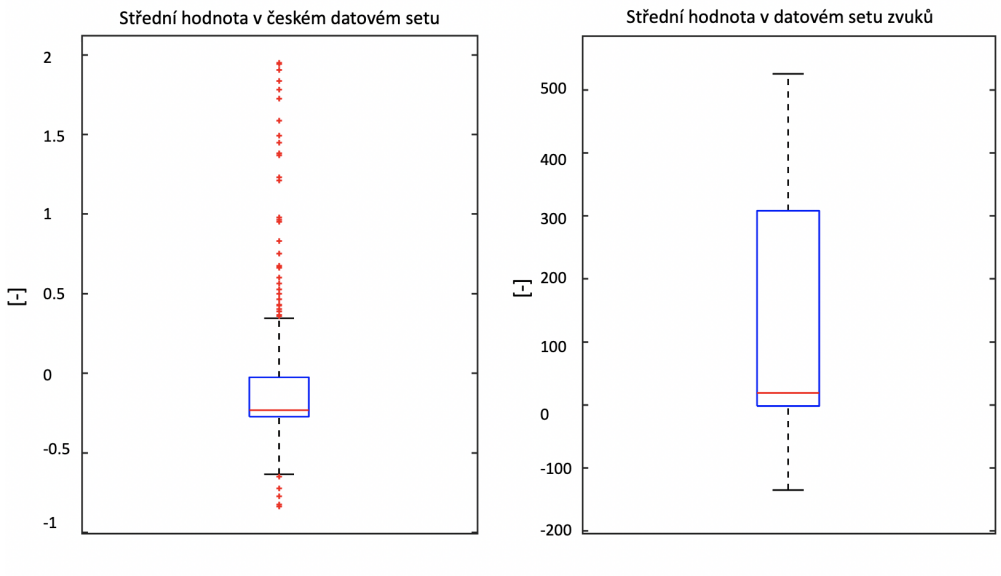
Pearsonův koeficient spočítáme dle vztahu:

$$Pearsonuv\ koeficient = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{5.5}$$

kde $\bar{x}$ je aritmetický průměr prvního vektoru a $\bar{y}$ aritmetický průměr druhého [4].

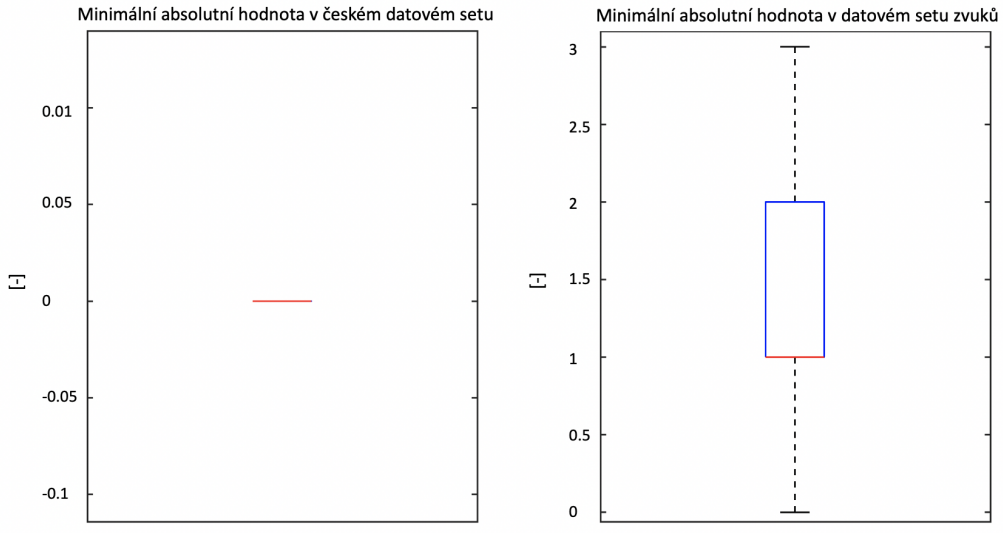
Pearsonův korelační koeficient pro maximální absolutní hodnoty a maximální hodnoty vychází $0,9855$, vidíme tedy že tyto dvě metriky jsou velmi silně korelované a do statistického modelu stačí vybrat jen jednu z nich.

Při pohledu na střední hodnoty si můžeme všimnout, že u české datové sady se tyto hodnoty ve většině případů nacházejí kolem 0, naopak u datové sady zvuků vidíme hodnoty i v řádu stovek (Obr. 5.10).

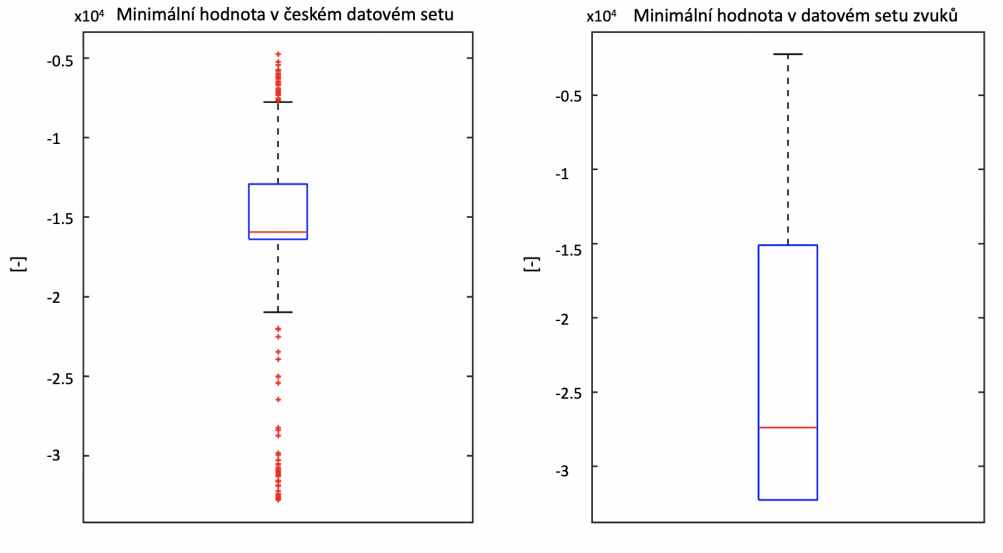


Obr. 5.10: Střední hodnoty pro obě datové sady

Minimální absolutní hodnota je pro českou datovou sadu vždy přibližně rovna nule, pro datovou sadu obsahují zvuky jsou její hodnoty nepatrně vyšší (stále maximálně jednotky) viz Obr. 5.11. Ve statistickém modelu i vzhledem k významu této metriky neočekáváme, že by tato metrika hrála důležitou roli. Minimální hodnoty v obou datových sadách si rozsahem odpovídají a stejně jako u předchozí metriky by neměly na přesnost přepisu mít výrazný vliv.

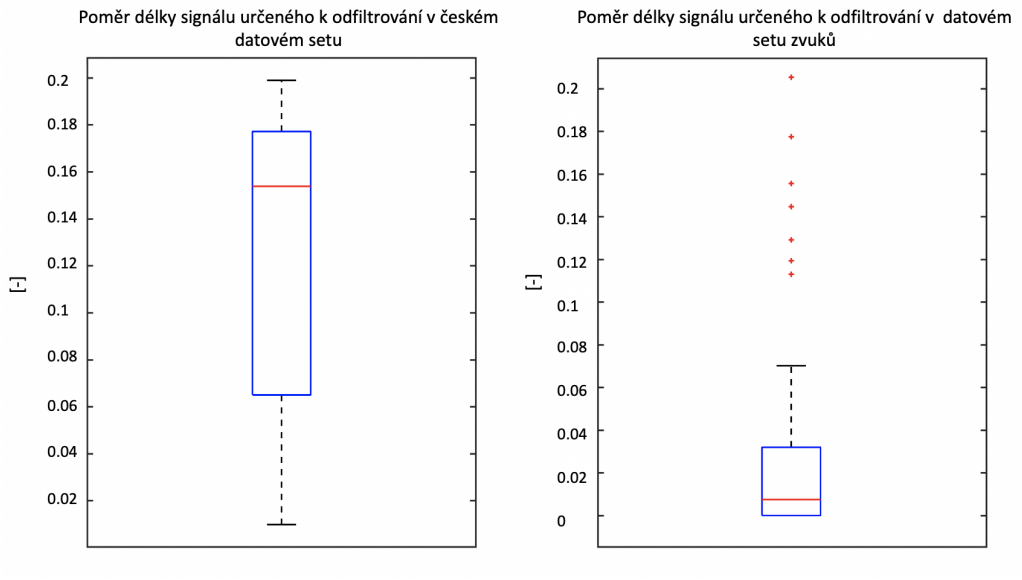


Obr. 5.11: Minimální absolutní hodnoty hodnoty pro obě datové sady

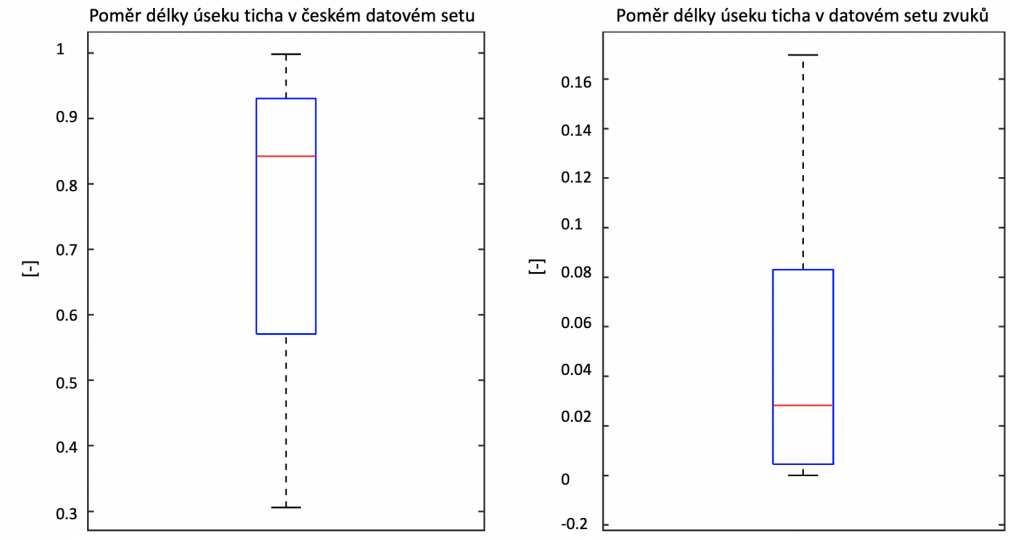


Obr. 5.12: Minimální hodnoty pro obě datové sady

Poměr signálu určeného k filtrování je v české datové sadě velmi vysoký (Obr. 5.13). Je to zejména proto, jakým způsobem byly tyto nahrávky vytvořeny, tj. ručně rozdělený dialog dvou lidí na dva monology. Obsahuje tedy velké množství ticha, což dokazuje i Obr. 5.14. Zároveň jsou na první pohled tyto dvě metriky velmi podobné co se týče rozsahu. Jejich Pearsonův korelační koeficient vychází 0.9367, do statistického modelu vybereme tedy jen jednu z těchto metrik.

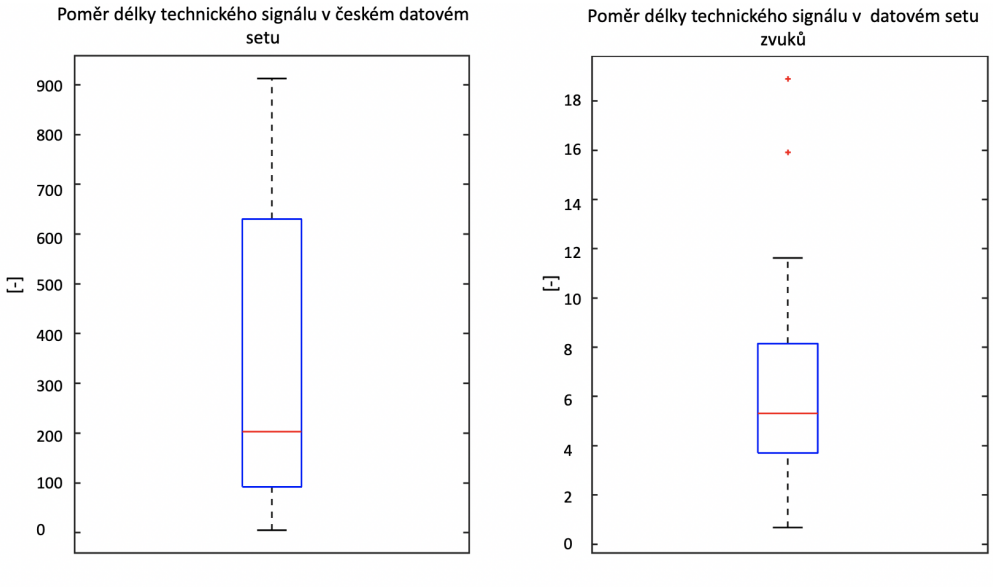


Obr. 5.13: Poměr signálu určeného k filtrování pro obě datové sady



Obr. 5.14: Poměr ticha pro obě datové sady

Poměr technického signálu je pro obě datové sady minimální a jeho hodnoty se pohybují kolem 0. Vzhledem k tomu, že celý datový set zvuků obsahuje technické signály, můžeme pozorovat nesnadnost jejich detekce.



Obr. 5.15: Poměr technického signálu pro obě datové sady

5.3.2 Normalizace dat

Pro zlepšení funkčnosti statistického modelu je žádoucí provést datovou normalizaci. Jde o změnu rozsahu jednotlivých metrik tak, aby jejich řády odpovídaly. Statistický model se pak vytvoří přesněji a je menší šance uvíznutí v lokálním minimu při numerických metodách optimalizace. Pro přepočítání využiji vzorec:

$$x_{i,m} = \frac{x_{i,m} - \overline{x_i}}{max(x_i) - min(x_i)}, \tag{5.6}$$

kde $x_{i,m}$ je i-tá metrika pro m-tou nahrávku, $\overline{x_i}$ je průměr i-té metriky, $max(x_i)$ je maximální možná hodnota i-té metriky a $min(x_i)$ je minimální možná hodnota i-té metriky.

Každá metrika bude díky této normalizaci vyjádřena hodnotami v intervalu $< -1, 1 >$ [34].

5.4 Lineární regrese

První použitou metodou pro vytvoření statistického modelu je lineární regrese. Cílem lineární regrese je predikovat hodnotu spojité proměnné y na základě n-dimenzionálního vektoru $\boldsymbol{x}$ vstupních proměnných [21][43].

Máme-li trénovací datový set obsahující M různých vektorů $\boldsymbol{x^{(m)}}$, kde $m = 1, 2, \ldots, M$ (M značí v našem případě počet nahrávek) a k nim odpovídající hodnoty $\boldsymbol{y^{(m)}}$, chceme umět predikovat hodnotu y na základě nového vektoru $\boldsymbol{x}$.

Takovou predikci hodnoty y označujeme jako hypotézu $h(x_1^{(m)}, \ldots, x_n^{(m)})$ a vypočítáme ji:

$$h(x_1^{(m)}, \ldots, x_n^{(m)}) = \omega_0 + \omega_1 x_1^{(m)} + \ldots + \omega_n x_n^{(m)}, \tag{5.7}$$

kde $h(x_1, \ldots, x_n)$ označuje hodnotu hypotézy pro vstupní proměnné $x_1, \ldots, x_n$ a $\omega_0, \ldots, \omega_n$ jsou parametry hypotézy. [36]

Cílem lineární regrese je tedy nalézt takové parametry $\omega_0, \ldots, \omega_n$, pro které budou predikce $h(x_1, \ldots, x_n)$ v testovacím datovém setu dostatečně odpovídat jejich referenčním hodnotám y.

Dále potřebujeme číselně vyjádřit, jak moc predikce $h(x_1, \ldots, x_n)$ odpovídá referenčním hodnotám y. K tomu nám poslouží tzv. *ztrátová funkce*. V kombinaci s lineární regresí se standardně používá *střední kvadratická chyba (angl. Mean Squared Error)* [6]. Tu spočítáme jako:

$$J(w_1, \ldots, w_n) = \frac{1}{2M} \sum_{i=1}^{M} \left[h(x_1^{(m)}, \ldots, x_n^{(m)}) - y^{(m)} \right]^2, \tag{5.8}$$

kde $J(w_1, \ldots, w_n)$ je hodnota střední kvadratické chyby, M je počet audionahrávek v trénovací sadě, $h(\boldsymbol{x})$ je hodnota predikce pro vektor $\boldsymbol{x}$ a $y^{(i)}$ je referenční hodnota [36].

Cílem je tedy najít parametry $\omega_0, \ldots, \omega_n$ takové, aby hodnota $J(w_1, \ldots, w_n)$ byla minimální.

V rámci zjednodušení zápisu (a vzhledem k jednoduchosti práce s maticemi v prostředí Matlab) si zavedeme maticový zápis lineární regrese [16]. Máme-li vektory $(x_1^{(m)}, ..., x_n^{(m)})$ a jejich váhy $(\omega_0, \ldots, \omega_n)$, pak $\omega_0 + \omega_1 x_1^{(m)} + \ldots + \omega_n x_n^{(m)} = \omega_0 x_0^{(m)} + \omega_1 x_1^{(m)} + \ldots + \omega_n x_n^{(m)}$, kde $x_0^{(m)} = 1$ pro všechny m. Odteď tedy vektorem $\boldsymbol{x^{(m)}}$ budeme označovat vektor $(x_0^{(m)}, ..., x_n^{(m)})$, kde $x_0^{(m)} = 1$ pro všechny m a vektorem $\boldsymbol{\omega}$ označíme vektor $(\omega_0, \ldots, \omega_n)$. Nyní můžeme psát:

$$h(x_0^{(m)}, \ldots, x_n^{(m)}) = h(\boldsymbol{x^{(m)}}) = \boldsymbol{\omega}^T \boldsymbol{x^{(m)}}, \tag{5.9}$$

$$J(w_0, \ldots, w_n) = J(\boldsymbol{\omega}) = \frac{1}{2M} \sum_{i=1}^{M} \left[h(\boldsymbol{x}^{(m)}) - y^{(m)} \right]^2, \tag{5.10}$$

Dalšími úpravami dostaneme:

$$\begin{aligned} J(\boldsymbol{\omega}) &= \frac{1}{2M} (\boldsymbol{X\Omega} - \boldsymbol{y})^T (\boldsymbol{X\Omega} - \boldsymbol{y}) = \\ &= \frac{1}{2M} (\boldsymbol{X\Omega} - \boldsymbol{y})^T (\boldsymbol{X\Omega} - \boldsymbol{y}) \end{aligned} \tag{5.11}$$

$J(\boldsymbol{\omega})$ budeme později pokládat rovno 0 pro nalezení minima, zbavíme se tedy zlomku $\frac{1}{2M}$.

$$\begin{aligned} J(\boldsymbol{\omega}) &= ((\boldsymbol{X\Omega})^T - \boldsymbol{y}^T)(\boldsymbol{X\Omega} - \boldsymbol{y}) = \\ &= (\boldsymbol{X\Omega})^T\boldsymbol{X\Omega} - (\boldsymbol{X\Omega})^T\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{X\Omega}) + \boldsymbol{y}^T\boldsymbol{y} \end{aligned} \tag{5.12}$$

Všimneme si, že $\boldsymbol{X\Omega}$ je vektor. Stejně tak $\boldsymbol{y}$. Pokud je tedy navzájem vynásobíme, nezáleží na jejich pořadí. Můžeme tedy dále zjednodušit:

$$J(\boldsymbol{\omega}) = \boldsymbol{\Omega}^T\boldsymbol{X}^T\boldsymbol{X\Omega} - 2(\boldsymbol{X\Omega})^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} \tag{5.13}$$

Pro nalezení minima funkce $J(\boldsymbol{\Omega})$ ji zderivujeme podle $\boldsymbol{\Omega}$ a položíme rovnu nule:

$$\frac{\partial J}{\partial \boldsymbol{\Omega}} = 2\boldsymbol{X}^T\boldsymbol{X\Omega} - 2\boldsymbol{X}^T\boldsymbol{y} = 0 \tag{5.14}$$

$$\boldsymbol{X}^T\boldsymbol{X\Omega} = \boldsymbol{X}^T\boldsymbol{y} \tag{5.15}$$

Předpokládáme-li, že je matice $\boldsymbol{X}^T\boldsymbol{X}$ regulární, dostaneme

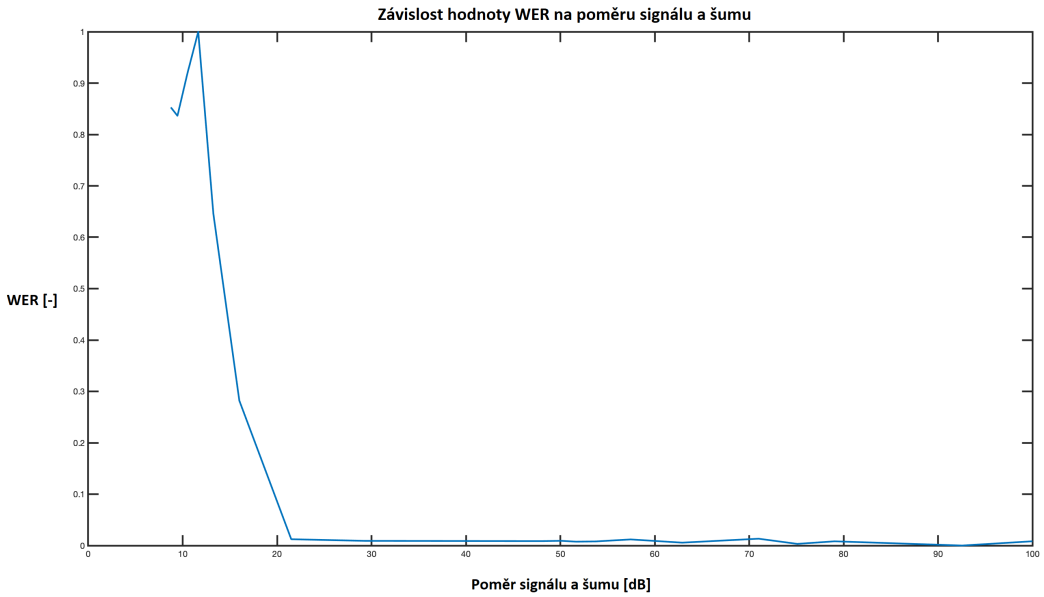
$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{\Omega}, \tag{5.16}$$

jakožto explicitní vzorec pro výpočet $\boldsymbol{\Omega}$.

Výraz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ označujeme jako pseudo-inverzní matici.

5.5 Logistická regrese

U lineární regrese se snažíme předpovědět přesnost strojového přepisu nahrávky pomocí lineární kombinace metrik a jejich vah. Nyní se ale můžeme podívat, jak vypadají závislosti některých metrik na referenční hodnotě přesnosti přepisu (WER). Na Obr. 5.16 je vykreslena závislost WER na poměru signálu a šumu u české datové sady. K vytvoření grafu jsme použili průměrné hodnoty poměru signálu a šumu pro každou z úrovní zašumění (viz. kapitola 5.3.1) a jejich odpovídající celkovou hodnotu WER.



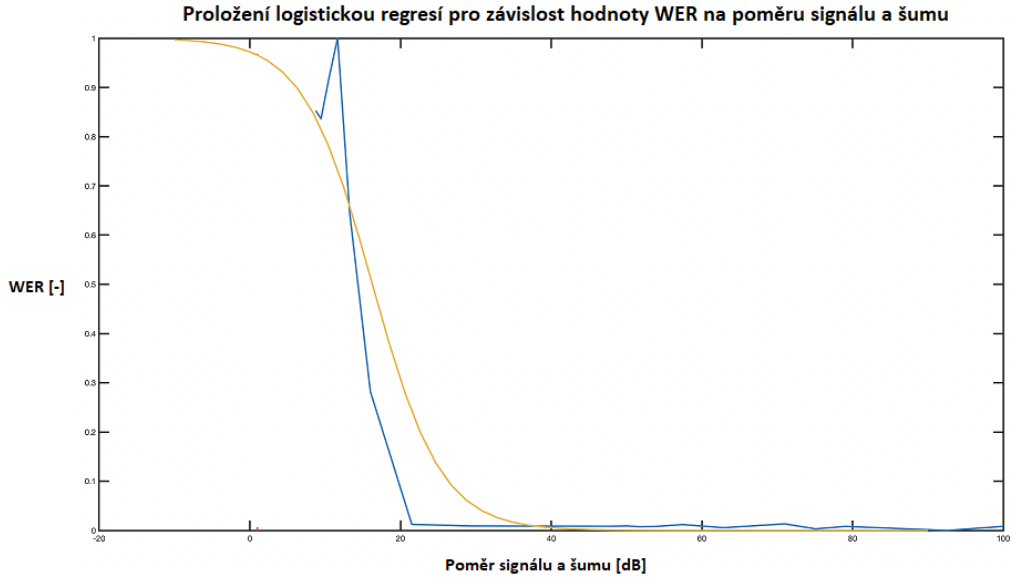
Obr. 5.16: Závislost hodnoty WER na poměru signálu a šumu

Na první pohled jde vidět, že lineární proložení této konkrétní závislosti není nejlepším řešením. Pomoci nám zde může tzv. *logistická regrese*. Narozdíl od lineární regrese se logistická regrese snaží proložit a následně predikovat data pomocí sigmoidní funkce $h(x)$:

$$h(\boldsymbol{x}^{(m)}) = \frac{1}{1 + e^{-\boldsymbol{\omega}^T \boldsymbol{x}^{(m)}}}, \tag{5.17}$$

kde $h(\boldsymbol{x}^{(m)})$ označuje hodnotu hypotézy pro vstupní vektor $\boldsymbol{x}^{(m)}$ a $\boldsymbol{\omega}$ je vektor parametrů hypotézy. [15]

Na Obr. 5.17 vidíme příklad proložení logistickou regresí pro 1-dimenziální vektor $\boldsymbol{x}^{(m)}$.



Obr. 5.17: Proložení logistickou regresí pro závislost z Obr. 5.16

Hodnoty logistické regrese $h(\boldsymbol{x}^{(m)})$ leží vždy v intervalu $< 0, 1 >$. Pro použití tohoto modelu je tedy nutné normalizovat referenční hodnoty $\boldsymbol{y}^{(m)}$ (normalizované hodnoty WER) do stejného intervalu. Pak se lze na tyto normalizované hodnoty $\boldsymbol{y'}^{(m)}$ dívat jako na pravděpodobnostní rozložení. Každou hodnotu lze formulovat jako pravděpodobnost, s jakou se daná řečová nahrávka správně přepíše. Normalizovaná hodnota $\boldsymbol{y'}^{(m)} = 1$ by znamenala velmi špatný přepis a naopak hodnota $\boldsymbol{y'}^{(m)} = 0$ velmi dobrý přepis. Každá predikovaná hodnota $h(\boldsymbol{x}^{(m)})$ pro $m = 1, \ldots, M$ udává predikci $\boldsymbol{y'}^{(m)} = 1$. Pro porovnání dvou pravděpodobnostních rozložení opět využijeme křížovou entropii a budeme ji zde považovat za ztrátovou funkci. Pro každé dvě hodnoty $\boldsymbol{y'}^{(m)}$ a $h(\boldsymbol{x}^{(m)})$ vypočítáme hodnotu křížové entropie jako [20]:

$$H(y'^{(m)}, h(\boldsymbol{x}^{(m)})) = -y'^{(m)} log\ h(\boldsymbol{x}^{(m)}) - (1 - y'^{(m)}) log(1 - h(\boldsymbol{x}^{(m)})) \tag{5.18}$$

Křížová entropie se velmi dobře vypořádá s extrémy rozdílů dvou porovnávaných hodnot. Použijeme-li ji jako ztrátovou funkci, můžeme očekávat, že při její minimalizaci se úspěšně odstraní velmi nepřesné hodnoty predikce (např. pro extrémní predikci 0 při referenční hodnotě 1 je hodnota křížové entropie ∞) [22].

Ztrátovou funkci $J(\boldsymbol{\Omega})$ je třeba upravit pro obsažení všech dvojic referenčních a predikovaných hodnot:

$$J(\Omega) = \frac{1}{M} \sum_{m=1}^{M} H(y'^{(m)}, h(\boldsymbol{x}^{(m)})), \tag{5.19}$$

kde $J(\Omega)$ je hodnota ztrátové funkce, M je počet nahrávek, $\boldsymbol{y'^{(m)}}$ je normalizovaná referenční hodnota a $h(\boldsymbol{x^{(m)}})$ je predikovaná hodnota.

Křížová entropie jakožto ztrátová funkce se liší od střední kvadratické chyby používané u lineární regrese především rozložením chyb. Zatímco u střední kvadratické chyby jsou stejné chyby rozprostřeny přes všechny hodnoty. Křížová entropie místo toho zabraňuje obrovským chybám (např. že se $h(\boldsymbol{x^{(m)}})$ predikuje jako velmi nízká a přitom referenční hodnota bude velmi vysoká), neporadí si ale tak přesně s predikcí v celém rozsahu. Důležitými vlastnostmi křížové entropie jsou také její nezápornost a fakt, že pokud budou všechny predikované hodnoty blízké referenčním, bude se hodnota křížové entropie velmi blížit nule.

Pro minimalizaci ztrátové funkce využijeme *metodu nejstrmějšího klesání gradientu* (dále jako gradientní metoda). V této metodě iterativně měníme jednotlivé parametry $\omega_0, \ldots, \omega_n$ podle gradientu ztrátové funkce $J(w_1, \ldots, w_n)$. Rychlost změny určujeme parametrem α:

$$w_{j,k+1} = w_{j,k} - \alpha \cdot \frac{\partial}{\partial w_j} J(w_1, \ldots, w_n), \tag{5.20}$$

kde $w_{j,k+1}$ je j-tý parametr v $k+1$ iteraci, α je krok metody a $J(w_1, \ldots, w_n)$ je ztrátová funkce.

Pro derivaci ztrátové funkce využijeme řetízkové pravidlo pro derivace složených funkcí [27]:

$$\frac{\partial J}{\partial \omega_j} = \frac{\partial J}{\partial a}\frac{\partial a}{\partial z}\frac{\partial z}{\partial \omega_j}, \tag{5.21}$$

kde

$$\begin{aligned} \frac{\partial J}{\partial a} &= \frac{\partial J}{\partial a}\Big[- y'^{(m)} log\ a - (1 - y'^{(m)}) log(1-a)\Big] = \\ &= y'^{(m)}\frac{1}{a} + \frac{1 - y'^{(m)}}{1-a}, \end{aligned} \tag{5.22}$$

dále

$$\begin{aligned} \frac{\partial a}{\partial z} &= \frac{\partial a}{\partial z}\frac{1}{1+e^z} = \\ &= \frac{\partial a}{\partial z}(1+e^z)^{-1} = \\ &= (1+e^z)^{-2}\frac{\partial a}{\partial z}(1+e^z) = \\ &= (1+e^z)^{-2}\frac{\partial a}{\partial z}e^z = \\ &= (1+e^z)^{-2}e^z\frac{\partial a}{\partial z}z = \\ &= (1+e^z)^{-2}e^z\frac{\partial a}{\partial z}z = \\ &= (1-z)z \end{aligned} \tag{5.23}$$

a

$$\frac{\partial z}{\partial w_j} = x_j^{(m)} \tag{5.24}$$

Po vynásobení dostaneme konečný výraz:

$$\frac{\partial J}{\partial \omega_j} = \frac{1}{M} \sum_{m=1}^{M} x_j (h(\boldsymbol{x}^{(\boldsymbol{m})}) - y'^{(m)}), \tag{5.25}$$

Výpočet nových parametrů tedy provedeme dle vzorce:

$$w_{j,k+1} = w_{j,k} - \alpha \frac{1}{M} \sum_{m=1}^{M} x_j (h(\boldsymbol{x}^{(\boldsymbol{m})}) - y'^{(m)}), \tag{5.26}$$

kde $w_{j,k+1}$ je hodnota j-té metriky v $k+1$ iteraci, $w_{j,k}$ je hodnota j-té metriky v k-té iteraci, M je počet nahrávek, $h(\boldsymbol{x}^{(\boldsymbol{m})})$ hodnota predikce a $y'^{(m)}$ referenční hodnota.

5.6 K-křížová validace

Než přistoupíme k samotným regresím, je potřeba rozdělit datové sady na trénovací a testovací data (používané datové sady jsou popsány v kapitole 5.1). Standardním postupem je zde využití tzv. *k-křížové validace* [35]. Tato metoda má jediný parametr, a to konstantu k, která značí počet skupin, do kterých datový set rozdělíme.

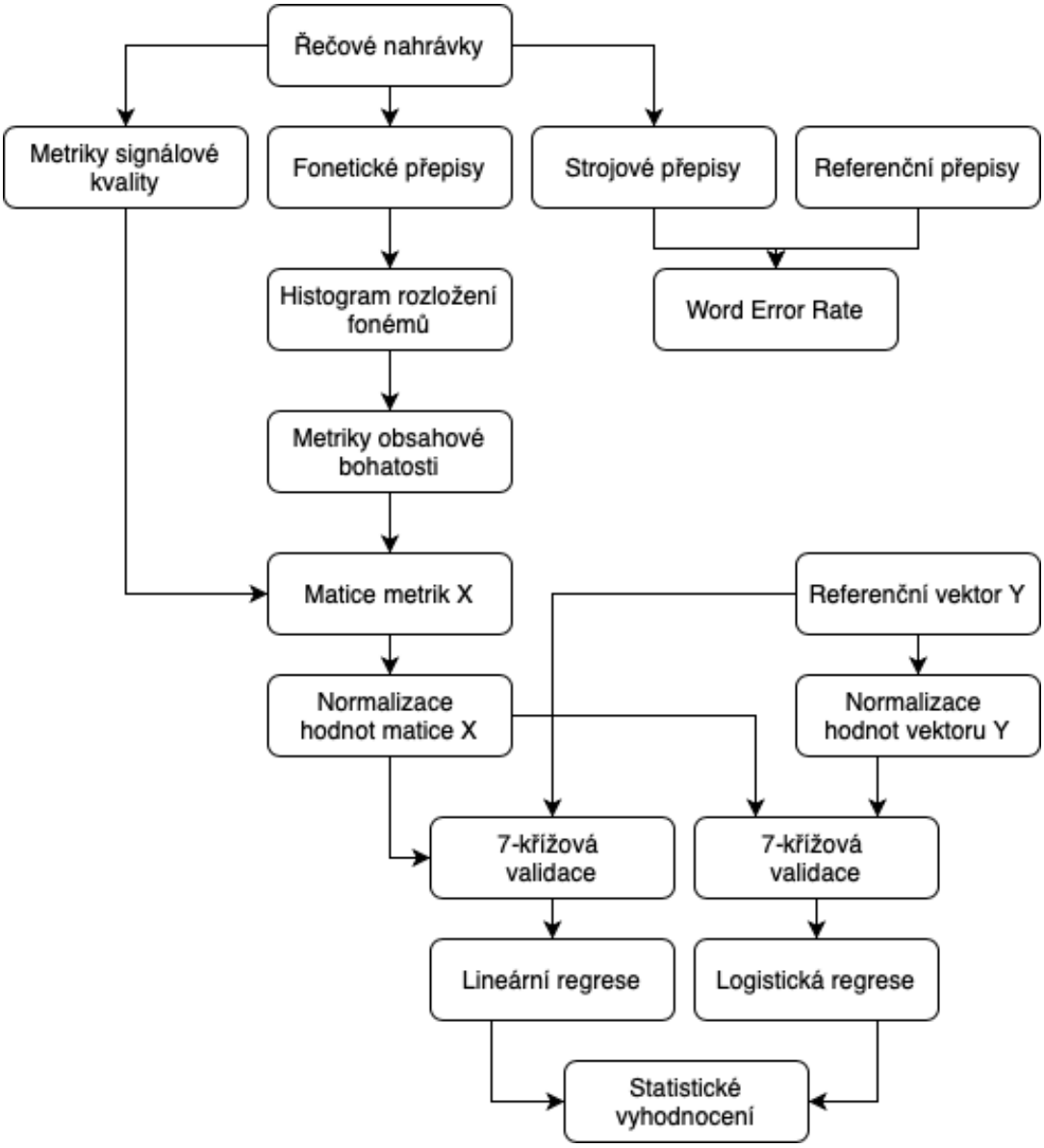
Celý postup vypadá následovně:

- datové sady (českou datovou sadu a sadu zvuků) spojíme do jedné (budeme mít tedy 931 vektorů $\boldsymbol{x}^{(m)}$ a k nim přilehlých hodnot $y^{(m)}$)
- náhodně promícháme pořadí vektorů v datové sadě $\boldsymbol{x}^{(m)}$ (a k nim přilehlé hodnoty $y^{(m)}$)
- rozdělíme datovou sadu na k unikátních skupin se stejným (popř. přibližně stejným) počtem nahrávek
- postupně použijeme každou skupinu jako testovací sadu, přičemž všechny ostatní skupiny budou sloužit jako trénovací sada
- statisticky vyhodnotíme všech k scénářů

Datová sada obsahuje 931 nahrávek, je tedy potřeba zvolit hodnotu k takovou, aby byla trénovací sada reprezentativní a zároveň zbyl rozumný počet nahrávek v testovací sadě. V našem případě jsme zvolili hodnotu $k = 7$, tedy v každé testovací sadě bude 133 nahrávek a 798 jich zbyde pro natrénování modelu.

6 Programová realizace

V práci jsme pro zpracování nahrávek používali řečové technologie popsané v kapitole 1.4. K vytvoření statistického modelu jsme využili prostředí Matlab. Hlavním skriptem diplomové práce je soubor *main.m*. Z něj se pak volají ostatní potřebné funkce. Pomocné skripty jsou umístěny ve složce *Pomocne* a používaná data ve složce *Data*. Na Obr. 6.1 je zobrazeno blokové schéma celé programové realizace.



Obr. 6.1: Zjednodušené blokové schéma programové realizace

Pro výpočet metrik signálové kvality jsme použili Phonexia Speech Quality Estimator (viz kapitola 1.4.1). Je to aplikace spustitelná z příkazového řádku v prostředí Windows či Linux. Vstupem je složka obsahující audionahrávky a výstupem je složka, ve které je pro každou nahrávku jeden **.txt* soubor. Každý soubor obsahuje signálové metriky dané nahrávky (Obr. 6.2):

```
waveform_clipped_length, value=2.050000, min=0.000000e+00, max=4.312000e+02, str='', score=0.000000, weight=0.000000
waveform_clipping_threshold, value=31128.650391, min=0.000000e+00, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_kurtosis, value=7.224835, min=0.000000e+00, max=3.402823e+38, str='', score=0.000000, weight=0.000000
waveform_length, value=431.200012, min=0.000000e+00, max=3.402823e+38, str='', score=0.000000, weight=0.000000
waveform_max_abs_value, value=32768.000000, min=0.000000e+00, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_max_value, value=32767.000000, min=-3.276700e+04, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_mean, value=-129.922775, min=-3.276700e+04, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_min_abs_value, value=0.000000, min=0.000000e+00, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_min_value, value=-32768.000000, min=-3.276700e+04, max=3.276700e+04, str='', score=0.000000, weight=0.000000
waveform_n_bits, value=10.000000, min=0.000000e+00, max=1.000000e+01, str='', score=100.000000, weight=0.500000
waveform_n_levels, value=1024.000000, min=0.000000e+00, max=1.024000e+03, str='', score=0.000000, weight=0.000000
waveform_sample_freq, value=8000.000000, min=0.000000e+00, max=3.402823e+38, str='', score=0.000000, weight=0.000000
waveform_skewness, value=0.170575, min=-3.402823e+38, max=3.402823e+38, str='', score=0.000000, weight=0.000000
waveform_snr, value=2.000000, min=-3.402823e+38, max=3.402823e+38, str='', score=10.000000, weight=0.500000
waveform_standard_deviation, value=3793.715088, min=0.000000e+00, max=3.402823e+38, str='', score=0.000000, weight=0.000000
wfilter_filtered_length, value=26.504999, min=0.000000e+00, max=4.311750e+02, str='', score=0.000000, weight=0.000000
wfilter_filtered_ratio, value=0.061472, min=0.000000e+00, max=1.000000e+00, str='', score=0.000000, weight=0.000000
wfilter_intermittent_noise_length, value=0.000000, min=0.000000e+00, max=4.311750e+02, str='', score=0.000000, weight=0.000000
wfilter_intermittent_noise_ratio, value=0.000000, min=0.000000e+00, max=1.000000e+00, str='', score=0.000000, weight=0.000000
wfilter_silence_length, value=0.000000, min=0.000000e+00, max=4.311750e+02, str='', score=0.000000, weight=0.000000
wfilter_silence_ratio, value=0.000000, min=0.000000e+00, max=1.000000e+00, str='', score=0.000000, weight=0.000000
wfilter_snr, value=67.685547, min=-3.402823e+38, max=3.402823e+38, str='', score=0.000000, weight=0.000000
wfilter_speech_signal_length, value=404.669983, min=0.000000e+00, max=4.311750e+02, str='', score=0.000000, weight=0.000000
wfilter_technical_signal_length, value=121.245003, min=0.000000e+00, max=4.311750e+02, str='', score=0.000000, weight=0.000000
wfilter_technical_signal_ratio, value=0.281197, min=0.000000e+00, max=1.000000e+00, str='', score=0.000000, weight=0.000000
```

Obr. 6.2: Výstup SQE - textový soubor s metrikami a jejich specifikacemi

V tomto souboru nás zajímají první čtyři sloupce. První sloupec je název dané metriky, druhý sloupec je hodnota metriky pro daný soubor, třetí a čtvrtý sloupec obsahují minimální a maximální možné hodnoty této metriky. Hodnoty metrik načítáme do Matlabu funkcí *vytvor_tabulku_parametru.m*. Funkce má na vstupu cestu ke složce s **.txt* soubory, na výstupu buňkové pole obsahující všechny signálové metriky pro všechny parametry.

K vytvoření fonetických přepisů jsme použili Phonexia Phoneme Recognizer (viz kapitola 1.4.2). Stejně jako u SQE jde o aplikaci spustitelnou z příkazového řádku v prostředí Windows či Linux. Vstupem je složka obsahující audionahrávky a výstupem je složka, ve které je pro každou nahrávku jeden **.txt* soubor. Každý soubor obsahuje fonetický přepis dané nahrávky (Obr. 6.3):

Obr. 6.3: Textový soubor - ukázka výstupu z PHR

Fonémy načítáme do Matlabu funkcí *foneticke_prepisy.m*. Funkce má na vstupu cestu ke složce s **.txt* soubory, na výstupu buňkové pole fonetických přepisů pro každou nahrávku.

Dalším krokem po získání fonetických přepisů je tvorba histogramů fonémů pro jednotlivé nahrávky. Pro tento účel používáme funkci *vytvor_histogramy.m*. Na vstupu jsou jednak buňkové pole fonetických přepisů, ale i jeden referenční **.txt* soubor obsahující všechny fonémy daného jazyka. Na výstupu je buňkové pole obsahující všechny histogramy pro všechny nahrávky.

Hned následujícím logickým krokem je výpočet metriky počtu unikátních fonémů jako první z metrik obsahové bohatosti. Tento výpočet provádíme pomocí funkce *vypocti_unikatni_fonemy.m*, která má na vstupu buňkové pole všech histogramů a referenční **.txt* soubor obsahující všechny fonémy daného jazyka. Na výstupu dává buňkové pole obsahující pro každý souboru počet unikátních fonémů.

Křížovou entropii počítáme funkcí *vypocti_entropii.m*. Na vstupu načítáme buňkové pole histogramů a na výstupu dává buňkové pole entropií pro všechny nahrávky.

K vyhodnocení WER je prvním krokem přepsání nahrávek automatickým strojovým přepisem. K tomuto účelu jsme použili Phonexia Speech To Text (kapitola 1.4.3). Znovu jde o aplikaci spustitelnou z příkazového řádku v prostředí Windows či Linux. Vstupem je složka obsahující audionahrávky a výstupem je složka, ve které je pro každou nahrávku jeden **.trn* soubor. Každý soubor obsahuje textový přepis dané nahrávky:

```
0 1100000 <silence/> 0.000000 0.000000 0
1100000 4700000 <segment> 0.000000 1.000000 0
4700000 10700000 prosím 0.000000 1.000000 0
10700000 10700000 </segment> 0.000000 1.000000 0
10700000 69000000 <silence/> 0.000000 0.000000 0
69000000 72900000 <segment> 0.000000 1.000000 0
72900000 78300000 ano 0.000000 1.000000 0
78300000 78300000 </segment> 0.000000 1.000000 0
78300000 114300000 <silence/> 0.000000 0.000000 0
114300000 118200000 <segment> 0.000000 1.000000 0
118200000 123300000 jó -0.146103 0.864068 0
123300000 125400000 <silence/> 0.000000 1.000000 0
125400000 126900000 na -0.065006 0.937062 0
126900000 132000000 krátký 0.000000 1.000000 0
132000000 138000000 určitě 0.000000 1.000000 0
138000000 138000000 </segment> 0.000000 1.000000 0
138000000 364000000 <silence/> 0.000000 0.000000 0
364000000 367900000 <segment> 0.000000 1.000000 0
367900000 369900000 m -0.395872 0.673093 0
369900000 372700000 tak -0.049670 0.951543 0
372700000 374400000 u -0.204275 0.815239 0
374400000 374800000 <silence/> 0.000000 1.000000 0
```

Obr. 6.4: Ukázka výstupu strojového přepisu nahrávky na text

Dále se postup liší v závislosti na datové sadě. Pro českou datovou sadu bylo potřeba upravit strojový přepis na formát **.ctm*. Tento soubor obsahuje všechny strojové přepisy pro všechny nahrávky a má formát:

```
<F> <C> <BT> <DUR> word,
```

kde v prvním sloupci *<F> (z angl. file)* značí název souboru, ve druhém sloupci *<C> (z angl. channel)* značí akustický kanál souboru, přičemž v našem případě pracujeme s mono (jednokanálovými) nahrávkami, akustický kanál se tedy nemění, ve třetím sloupci *<BT> (z angl. begin time)* značí začátek časového úseku v sekundách, ve čtvrtém sloupci sloupci *<DUR> (z angl. duration)* značí délku trvání slova v sekundách a *word* je slovo v daném časovém úseku.

Současně bylo třeba upravit referenční přepisy do formátu **.stm*. Tento soubor obsahuje veškeré referenční přepisy pro všechny nahrávky a jeho formát vypadá následovně:

```
<F> <C> <S> <BT> <ET> transcript,
```

kde v prvním sloupci *<F> (z angl. file)* značí název souboru, ve druhém sloupci *<C> (z angl. channel)* značí akustický kanál souboru, ve třetím sloupci *<S> (z angl. speaker)* značí číslo řečníka, přičemž v našem případě je v každé nahrávce jen jeden řečník, číslo se tedy nemění, ve čtvrtém sloupci *<BT> (z angl. begin time)* značí začátek časového úseku v sekundách, v pátém sloupci sloupci *<ET> (z angl. end time)* značí konec časového úseku v sekundách a *transcript* je referenční přepis daného časového úseku.

Tyto soubory je možné porovnat skórovacím nástrojem *NIST SClite*. Na vstupu jsou **.ctm* a **.stm* soubory, na výstupu dostaneme jeden **.pra* soubor. Na Obr. 6.5 vidíme část tohoto souboru:

```
id: (1-026)
File: cs0000.l
Channel: a
Scores: (#C #S #D #I) 4 0 1 0
REF:  NO no no no no
HYP:  ** no no no no
Eval: D

id: (1-027)
File: cs0000.l
Channel: a
Scores: (#C #S #D #I) 1 0 0 0
REF:  jo
HYP:  jo
Eval:
```

Obr. 6.5: Ukázka části výstupu skórovacího nástroje

Nalezneme zde porovnání referenčního a strojově přepsaného strojového řetězce. U každého textového úseku jsou pro nás důležité parametry název nahrávky a počty správně přepsaných slov C, substitucí S, vypuštění D a vložení I. Tento soubor načítáme funkcí *vytvor_wer.m*, která má na vstupu zmíněný **.pra* soubor a na výstupu buňkové pole obsahující hodnotu WER pro každou nahrávku. Kvůli ochraně soukromých dat tento soubor nelze zveřejnit, nahradím tedy tuto část nahráním už spočítaných WER příkazem *load('Data/WER_CZ.mat')*.

Pro datovou sadu obsahující zvuky je postup jiný, jelikož nelze jeho WER standardní cestou spočítat (viz kapitola 5.2.1). K výpočtu upravené hodnoty WER používáme funkci *noises_wer.m*, která má na vstupu složku obsahující všechny **.trn* soubory (Obr. 6.4) a na výstupu dává buňkové pole ve stejném formátu jako u české datové sady. Uvnitř funkce *noises_wer.m* používáme příkaz programovacího jazyka *bash*, který nám spočítá počet slov daného souboru. Pro uživatele Windows je tedy potřeba nahrát tabulku WER pro zvuky náhradním příkazem *load('Data/WER_NOISES.mat')*.

Výstupem funkce *vytvor_tabulku_parametru.m* jsme získali buňkové pole všech metrik pro všechny parametry. Pro další použití je ovšem buňkové pole nepraktické vzhledem k jednoduché práci s maticemi v prostředí Matlab. Pro převedení do maticového tvaru využíváme funkci *vytvor_matici_parametru.m*, která má na vstupu buňkové pole všech signálových metrik a na výstupu dává matici $\boldsymbol{x}$ signálových metrik. Tato matice odpovídá formátu matice $\boldsymbol{x}$ z kapitoly 5.4. Maticový zápis je pro nás velmi výhodný, díky němu je možné jednoduše získat informace o konkrétních metrikách jako jejich minimální hodnotu, maximální hodnotu, průměr a další.

Stejně tak usnadňuje zobrazení metrik v grafech a umožní použití regresních funkcí pro vytvoření statistického modelu.

Než je možné použít regresní funkce, musíme normalizovat hodnoty signálových metrik do intervalu $< -1, 1 >$. Pro normalizaci používáme funkci *normalizuj_matici.m*, které na vstupu bere matici signálových metrik, buňkové pole maxim signálových metrik a buňkové pole minim signálových metrik. Výstupem je dle očekávání normalizovaná matice signálových metrik.

V kapitole 5.3 jsme vybrali metriky, které nejsou pro náš statistický model vhodné. Tyto metriky odstraníme funkcí *odstran_nepotrebne.m*, která má na vstupu normalizovanou matici signálových metrik a číselný vektor, ve kterém každé číslo odpovídá číslu řádku metriky v buňkovém poli signálových metrik, kterou chceme odstranit. Na výstupu dává normalizovanou matici signálových metrik vybraných pro statistický model.

Poslední úpravou matice $\boldsymbol{x}$ je přidání metrik obsahové bohatosti. Přidání provádíme funkcemi *pridej_unikatni_fonemy.m* a *pridej_entropii*. Obě funkce mají na vstupu normovanou matici $\boldsymbol{x}$ vybraných metrik a buňkové pole požadované metriky k přidání. Na výstupu tu stejnou matici s přidanou metrikou.

Funkci *vytvor_vektor_y.m* používáme k vytvoření referenčního vektoru $\boldsymbol{y}$, jak je uvedeno v kapitole 5.4. Na vstupu má buňkové pole WER a na výstupu dle očekávání vektor $\boldsymbol{y}$ obsahující referenční hodnoty WER. Pro lineární regresi je tohle už finální vektor $\boldsymbol{y}$. Pro logistickou regresi, která nabývá hodnot od 0 do 1 je potřeba tento vektor ještě normalizovat.

K normalizaci vektoru $\boldsymbol{y}$ slouží funkce *normalizuj_y*. Na vstupu přijde původní vektor $\boldsymbol{y}$ a výstupem je jeho normalizovaná verze (hodnoty mezi 0 a 1).

Pro statistické vyhodnocení je třeba nejprve provést 7-křížovou validaci (více v kapitole 5.6). Prvním krokem 7-křížové validace je sloučení obou datových sad a náhodné promíchání pořadí vektorů $\boldsymbol{x^{(m)}}$. K tomu slouží funkce *sluc_a_promichej.m*. Vstupy funkce jsou v tomto pořadí:

- normovaná matice $\boldsymbol{x}$ první datové sady
- normovaná matice $\boldsymbol{x}$ druhé datové sady
- vektor $\boldsymbol{y}$ první datové sady
- vektor $\boldsymbol{y}$ druhé datové sady

Výstupem je matice obsahující náhodně promíchané vektory $\boldsymbol{x^{(m)}}$ a k nim v posledním sloupci přidané hodnoty $\boldsymbol{y^{(m)}}$.

Druhým krokem je rozdělení promíchané matice na 7 datových sad. K tomu používáme funkci *rozdel_na_7_skupin.m*. Jejím vstupem je právě náhodně přeskládaná matice a výstupem je buňkové pole obsahující obsahující 7 různých rozdělení trénovacích a testovacích dat.

Algoritmus pro lineární regresi (více v kapitole 5.4) spouštíme funkcí *krizova_validace_linearni_regrese.m*. Tato funkce má vstupní parametr buňkové pole rozdělení testovacích a trénovacích dat. Výstupem je buňkové pole které obsahuje tyto hodnoty pro každé ze sedmi rozdělení trénovacích a testovacích dat:

- vektor $\boldsymbol{\Omega}$
- hodnotu ztrátové funkce
- referenční hodnoty WER
- predikce hodnot WER vycházející z tohoto statistického modelu

Výpočet lineární regrese je prováděn dle vzorce 5.16. Pro výpočet pseudo-inverzní matice používáme funkci *pinv* [18].

Algoritmus pro logistickou regresi (více v kapitole 5.5) je spoušten funkcí *krizova_validace_logisticka_regrese.m*. Tato funkce má na vstupu parametry:

- buňkové pole rozdělení testovacích a trénovacích dat (s normalizovanými referenčními hodnotami)
- vektor všech hodnot gradientní metody, pro které se daný algoritmus má spustit
- počet iterací gradientní metody

Výstupem je stejně jako u předchozí metody buňkové pole které obsahuje tyto hodnoty pro každé ze sedmi rozdělení trénovacích a testovacích dat:

- vektor $\boldsymbol{\Omega}$
- hodnotu ztrátové funkce
- referenční hodnoty WER
- predikce hodnot WER vycházející z tohoto statistického modelu

K vytvoření statistického modelu využíváme zmíněné gradientní metody jako nástroje pro minimalizaci ztrátové funkce (viz. rovnice 5.20).

7 Vyhodnocení statistického modelu

Pro vytvoření statistického modelu jsme tedy v kapitole 5.3.1 vybrali tyto metriky:

- délku přebuzeného signálu
- koeficient špičatosti
- maximální absolutní hodnota signálu
- střední hodnota signálu
- minimální absolutní hodnota signálu
- minimální hodnota signálu
- koeficient šikmosti
- poměr signálu a šumu na základě detekce řečové aktivity a explicitního výpočtu
- směrodatná odchylka
- poměr signálu určeného k odfiltrování
- poměr signálu a šumu na základě tvaru rozložení vzorků signálu
- poměr řečového signálu
- poměr technického signálu
- počet unikátních fonémů
- křížová entropie

Jejich význam je popsán s kapitole 3. Je důležité zachovat toto pořadí metrik, jelikož odpovídá pořadí v matici $\boldsymbol{X}$ a výsledný vektor parametrů $\boldsymbol{\Omega}$ bude postupně udávat váhy také přesně v tomhle pořadí. Prvním cílem statistického modelu je co nejlépe predikovat přesnost strojového přepisu na základě těchto metrik. Druhým cílem práce je vyhodnotit vliv dílčích metod na přesnost strojového přepisu.

Pro vytvoření statistického modelu byly použity lineární regrese a logistická regrese. U lineární regrese lze provést explicitní výpočet výsledného vektoru parametrů $\boldsymbol{\Omega}$. Naproti tomu u logistické regrese využíváme iterativní gradientní metody optimalizující ztrátovou funkci. Před použitím každé regresní metody jsme provedli 7-křížovou validaci (viz. kapitola 5.6), dostaneme tedy u každé metody 7 různých výsledků.

U lineární regrese jsme na základě přímého výpočtu (viz. rovnice 5.16) spočítali výsledný vektor parametrů $\boldsymbol{\Omega}$. Výsledky jsou zaznačeny (a pro přehlednost tabulky zaokrouhleny na celá čísla) v tabulce 7.1:

Tab. 7.1: Tabulka parametrů ω pro lineární regresi

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
ω_0	52	50	53	52	53	52	52
ω_1	-43	-40	-39	-40	-42	-42	-39
ω_2	17	16	23	30	17	20	13
ω_3	14	10	8	12	10	14	14
ω_4	-101	-131	-93	-99	-96	-100	-92
ω_5	-25	-21	-20	-19	-27	-25	-23
ω_6	10	7	2	9	5	7	10
ω_7	13	13	12	24	14	21	15
ω_8	-16	-15	-16	-15	-18	-17	-16
ω_9	15	19	5	12	11	15	8
ω_{10}	-16	-17	-16	-15	-15	-13	-15
ω_{11}	-8	-10	-7	-7	-7	-9	-6
ω_{12}	3	1	4	4	4	5	4
ω_{13}	-3	3	7	6	5	8	8
ω_{14}	-24	-22	-26	-25	-27	-24	-24
ω_{15}	-6	-6	-4	-5	-6	-5	-6

Průměrné hodnoty jednotlivých metrik jsou:

$$\Omega = (64, -47, 19, 42, -112, -31, 37, 12, -16, 18, -9, -47, -6)$$

Hodnota Ω_0 odpovídá dle rovnice 5.9 první hodnotě vektoru $\boldsymbol{x}_1^{(m)} = 0$, neodpovídá tedy žádné z metrik. Nejvyšší hodnotu parametru má metrika střední hodnota signálu. Podíváme-li se na rozložení jejich hodnot (viz. Obr. 5.3.1), vidíme že u datové sady zvuků má hodnoty v řádu stovek. Naopak hodnoty pro českou datovou sadu jsou blízké nule a v jednotkách. Sada zvuků má vyšší průměr WER, dá se tedy čekat, že klasifikátor předpokládá, že s vyšší střední hodnotou se zvyšuje i WER.

Velký vliv mají i metriky délka přebuzeného signálu, maximální absolutní hodnota a počet unikátních fonémů. Lze si všimnout, že počet unikátních fonémů má záporné znaménko a maximální absolutní hodnota kladné znaménko. Odpovídá to

skutečnosti, že při nižším počtu unikátních fonémů očekáváme horší strojový přepis. Kladné znaménko u maximální absolutní hodnoty zase odpovídá krabicovému grafu 5.8, kde pro zvuky jsou maximální absolutní hodnoty vyšší než pro českou sadu. Počet unikátních fonémů se jako metrika obsahové bohatosti ukazuje, že její zavedení má smysl a podílí se významnou částí na správném fungování statistického modelu.

Nejmenší vliv v tomto modelu mají metriky poměr signálu určeného k odfiltrování, poměr signálu a šumu na základě detekce řečové aktivity a explicitního výpočtu a poměr technického signálu. Na Obr. 5.13 vidíme, že rozložení hodnot je v řádech desetin až setin. Vzhledem k tomu, že celá datová sada zvuků obsahuje jen zvuky bez řeči, můžeme konstatovat, že tato metrika není schopna detekovat neřečové úseky spolehlivě. Poměr signálu a šumu na základě detekce řečové aktivity a explicitního výpočtu není dle očekávání výrazný (více v kapitole 5.3.1), jelikož v datové sadě zvuků vůbec žádná řeč není.

U lineární regrese jsem jako ztrátovou funkci využíval střední kvadratickou odchylku. Hodnotu ztrátové funkce pro všech 7 případů uvádím v tabulce 7.2:

Tab. 7.2: Ztrátová funkce

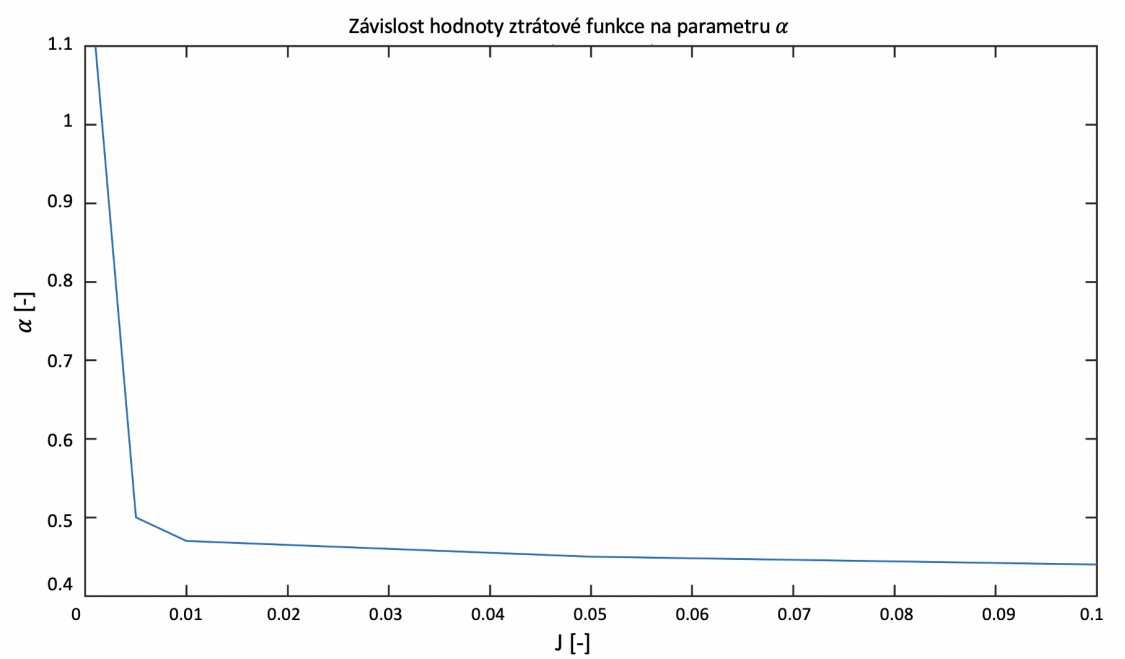
	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
$J(\mathbf{\Omega})$	95	238	234	196	133	198	160

Průměrná hodnota střední kvadratické chyby pro lineární regresi vychází

$$J(\mathbf{\Omega}) = 179.1975.$$

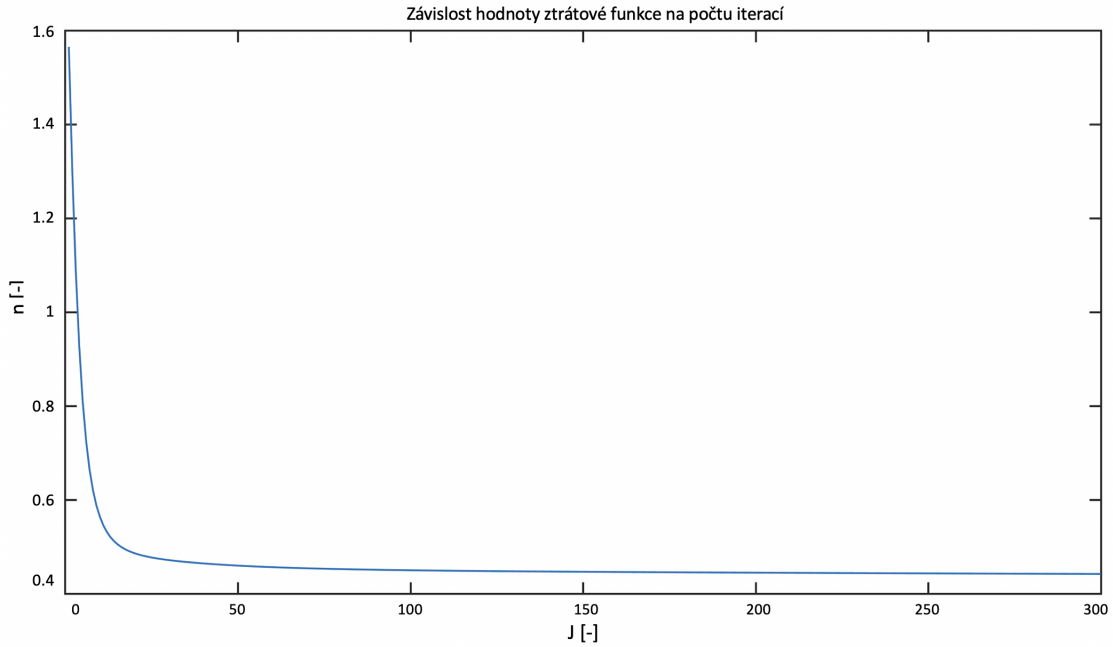
Na jednu dvojici hodnoty predikce $h(\mathbf{\Omega}^{(m)})$ a referenční hodnoty $y^{(m)}$ tedy vychází průměrná odchylka zhruba $13,4$. Vzhledem k povaze přesnosti WER je tohle číslo velmi dobré. WER totiž není závislé jen na kvalitě audionahrávky, velkou roli hraje například i jazyková vrstva. Pokud se v nahrávce vyskytují slova, která nejsou ve slovníku technologie automatického přepisu, není možné je správně přepsat. Predikovat hodnotu WER je tedy obecně složitá záležitost.

Druhým statistickým modelem je logistická regrese. U logistické regrese používám k minimalizaci ztrátové funkce gradientní metodu (viz rovnice 5.20). Prvním krokem tedy bude nalezení optimálního parametru α. Na Obr. 7.1 můžeme vidět závislosti průměrné hodnoty ztrátové funkce (ze 7 rozdělení dat) výsledků na parametru α.



Obr. 7.1: Závislost hodnoty ztrátové funkce na parametru α

Nejlepší hodnoty ztrátové funkce dosáhla pro hodnotu parametru $\alpha = 0,1$. Pro vyšší hodnoty α už roste hodnota ztrátové funkce nade všechny meze. Na Obr. 7.2 zobrazuji závislost hodnoty ztrátové funkce na počtu iterací gradientní metody.



Obr. 7.2: Závislost hodnoty ztrátové funkce na počtu iterací

Z grafu je patrné, že po 150. iteraci se už hodnota ztrátové funkce prakticky nemění. Vektory parametrů $\mathbf{\Omega}$ po 150. iteraci gradientní metody jsou zaznačeny v tabulce 7.3:

Tab. 7.3: Tabulka parametrů ω pro logistickou regresi

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
ω_0	-1.20	-1,18	-1,22	-1,17	-1,16	-1,21	-1,17
ω_1	-1,25	-1,19	-1,20	-1,28	-1,31	-1,21	-1,30
ω_2	1,06	1,13	1,06	1,03	1,00	1,02	1,02
ω_3	0,99	0,99	0,95	1,01	1,01	0,95	1,02
ω_4	0,04	0,03	-0,02	0,15	-0,00	0,02	-0,03
ω_5	-0,15	-0,24	-0,09	-0,20	-0,12	-0,06	-0,29
ω_6	1,28	1,31	1,24	1,32	1,33	1,26	1,25
ω_7	0,86	0,92	0,89	0,86	0,86	0,91	0,81
ω_8	-0,64	-0,62	-0,61	-0,62	-0,61	-0,59	-0,67
ω_9	0,24	0,21	0,27	0,15	0,28	0,25	0,25
ω_{10}	-0,13	-0,08	-0,07	-0,18	-0,11	-0,06	-0,12
ω_{11}	0,21	0,28	0,27	-0,17	0,17	0,27	0,17
ω_{12}	0,76	0,85	0,86	0,72	0,76	0,89	0,74
ω_{13}	0,78	0,78	0,80	0,77	0,86	0,79	0,74
ω_{14}	-0,86	-0,93	-0,82	-0,90	-0,93	-0,87	-0,92
ω_{15}	-0,11	-0,09	-0,08	-0,10	-0,10	-0,09	-0,10

Průměrné hodnoty jednotlivých metrik jsou:

$$\Omega = (-1.19, -1.25, 1.05, 0.99, 0.03, -1.17, 1.28, 0.87, -0.62, 0.23, -0.11, 0.22, 0.80, 0.79, -0.89, -0.09)$$

Zde mají největší vliv metriky délka přebuzeného signálu, minimální hodnota signálu a minimální absolutní hodnota signálu. Zajímavé je, že v kapitole 5.3.1 jsme nepředpokládali, že tyto metriky budou hrát významnou roli. Nejmenší vliv mají naopak střední hodnota signálu, poměr signálu určeného k odfiltrování a křížová entropie.

Ztrátovou funkcí pro logistickou regresi je křížová entropie (více v kapitole 5.5). Hodnotu ztrátové funkce pro všech 7 případů uvádím v tabulce 7.4:

Tab. 7.4: Ztrátová funkce

	1.	**2.**	**3.**	**4.**	**5.**	**6.**	**7.**
$J(\boldsymbol{\Omega})$	0,43	0,43	0,43	0,43	0,43	0,43	0,43

Průměrná hodnota je $J(\boldsymbol{\Omega}) = 0,43$. Pro možnost porovnání výsledků lineární a logistické regrese převedeme výsledné hodnoty hypotézy i normalizované referenční hodnoty do původního rozsahu hodnot. Hodnota střední kvadratické chyby pro toto rozložení vychází 246. U lineární hodnoty vyšla hodnota střední kvadratické chyby 179. Toto porovnání ovšem neznamená, že je model vytvořený za pomocí logistické regrese horší. Vzhledem k tomu, že u něj byla jako ztrátová funkce použita křížová entropie, očekáváme menší pravděpodobnost identifikace velmi špatné hodnoty jako velmi dobré, než u lineární regrese, kde byla jako ztrátová funkce použita střední kvadratická chyba. Při porovnání výsledků je maximální odchylka u lineární regrese $h(\boldsymbol{X}^{(m)}) - y^{(m)}$ rovna 76. Přitom u logistické regrese je maximální odchylka 47. Můžeme tedy vidět, že logistická regrese má horší průměrnou odchylku na jakékoliv porovnání predikované a referenční hodnoty, je u ní ale menší pravděpodobnost velmi velké chyby. Pro výběr vhodnějšího modelu je tedy třeba zvážit, která z těchto dvou možností nám vyhovuje více.

Pro vytvoření statistického modelu jsou tedy vhodné obě použité metody. Pro výběr mezi nimi je potřeba vyhodnotit, zda je pro nás důležitější mít lepší průměrnou odchylku na jednu dvojici predikce a referenční hodnoty, což zajišťuje model vytvořený na základě lineární regrese, nebo zamezit extrémním nepřesnostem za cenu větší průměrné odchylky s použitím logistické regrese. Vzhledem k povaze hodnoty Word Error Rate a její závislosti nejen na kvalitě řečové nahrávky, ale velmi výrazně i na používaných výrazech v těchto nahrávkách, jsou výsledky dosažené za pomoci těchto dvou modelů dostatečné a splňují svůj účel. I lehce nepřesný odhad přesnosti strojového přepisu může vyústit ve výraznou úsporu času (a v komerčním prostředí zákonitě i peněz) díky odstranění nahrávek s nízkou pravděpodobností kvalitního přepisu. Jako nejdůležitější metriky pro lineární regresi se ukázaly délka přebuzeného signálu, maximální absolutní hodnota a počet unikátních fonémů. U logistické regrese hrály velkou roli metriky délka přebuzeného signálu, minimální hodnota a minimální absolutní hodnota signálu.

Závěr

Tuto diplomovou práci je možné rozdělit na teoretickou a praktickou část.

Teoretická část se v první kapitole zabývá řečí a řečovými technologiemi. Je zde popsán proces vzniku řeči a současné možnosti využití řečových technologií. Také zavádím velmi důležitý pojem foném a vysvětluji jeho důležitost pro další části práce. V další kapitole je popsán současně nejrozšířenější způsob fonetického přepisu a vysvětlena role fonému v přepisu řeči do textu. Následně jsou představeny metriky signálové kvality, kde se zabýváme časovou, frekvenční i časově-frekvenční analýzou řečového signálu. Je vysvětlen jejich význam a použití. Ve čtvrté kapitole přidávám pojem obsahové bohatosti jako jiný pohled na hodnocení kvality audionahrávek. Zde zkoumám zejména fonetické rozložení nahrávek. Poslední teoretickou kapitolu věnuji regresním modelům, které následně v praktické části použiji k vytvoření statistického modelu pro predikci přesnosti strojového přepisu na základě hodnocení kvality řečové nahrávky.

Praktická část práce se nejprve zabývá návrhem metrik obsahové bohatosti, jakožto metrik zkoumajících fonetické rozložení dané řečové nahrávky. V šesté kapitole je poté popsána programová realizace. Zde ukazuji použití řečových technologií, standardní metody vyhodnocení přesnosti řečových technologií, použité datové sady a vysvětluji proces vytvoření statistického modelu. Poslední kapitolu věnuji samotnému statistickému vyhodnocení tohoto modelu. Je zde zhodnocena přesnost jednotlivých statistických modelů a vysvětleny možnosti jejich použití. Zároveň je vyhodnocena důležitost dílčích metrik pro dané statistické modely.

Literatura

[1] AITKEN, Colin G.G. a Franco TARONI. *Statistics and the Evaluation of Evidence for Forensic Scientists* [online]. Chichester, UK: John Wiley & Sons, 2004 [cit. 2020-05-10]. DOI: 10.1002/0470011238. ISBN 9780470011232.

[2] AKKALKOTKAR, Ameya, Kevin Scott BROWN a Boris PODOBNIK. An algorithm for separation of mixed sparse and Gaussian sources. *PLOS ONE* [online]. 2017, **12**(4) [cit. 2020-04-28]. DOI: 10.1371/journal.pone.0175775. ISSN 1932-6203. Dostupné z: https://dx.plos.org/10.1371/journal.pone.0175775

[3] ALTMAN, Douglas G. *Practical Statistics for Medical Research*. Boca Raton: Chapman & Hall/CRC, 1999. ISBN 0412276305.

[4] BENESTY, Jacob, Jingdong CHEN, Yiteng HUANG a Israel COHEN. Pearson Correlation Coefficient. *Noise Reduction in Speech Processing* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, 2009-3-5, s. 1-4 [cit. 2020-04-05]. Springer Topics in Signal Processing. DOI: 10.1007/978-3-642-00296-0_5. ISBN 978-3-642-00295-3. Dostupné z: http://link.springer.com/10.1007/978-3-642-00296-0_5

[5] BERANEK, Brett. Voice biometrics: success stories, success factors and what's next. *Biometric Technology Today* [online]. 2013, **2013**(7), 9-11 [cit. 2020-05-13]. DOI: 10.1016/S0969-4765(13)70128-0. ISSN 09694765. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0969476513701280

[6] BINIELI, Moshe. Machine learning: an introduction to mean squared error and regression lines [online]. 2018 [cit. 2020-05-10]. Dostupné z: https://www.freecodecamp.org/news/machine-learning-mean-squared-error-regression-line-c7dde9a26b93/

[7] ČIHÁK, Radomír. *Anatomie*. 2., upr. a dopl. vyd. Ilustroval Milan MED. Praha: Grada, 2001. ISBN 80-7169-970-5.

[8] DRŠATA, Jakub, CHROBOK, Viktor, ed. *Foniatrie - hlas*. Havlíčkův Brod: Tobiáš, 2011. Medicína hlavy a krku. ISBN 978-80-7311-116-8.

[9] DUARTE, Tiago, Rafael PRIKLADNICKI, Fabio CALEFATO a Filippo LANUBILE. Speech Recognition for Voice-Based Machine Translation. *IEEE Software* [online]. 2014, **31**(1), 26-31 [cit. 2020-05-13]. DOI: 10.1109/MS.2014.14. ISSN 0740-7459. Dostupné z: http://ieeexplore.ieee.org/document/6750466/

[10] EVANS, James D., 1996. Straightforward statistics for the behavioral sciences [online]. Pacific Grove: Brooks/Cole Pub. Co [cit. 2017-05-17]. ISBN 0534231004 9780534231002. Dostupné z: http://www.worldcat.org/title/straightforward-statistics-for-the-behavioralsciences/oclc/32465263

[11] GROENEVELD, Richard A. a Glen MEEDEN. Measuring Skewness and Kurtosis. *The Statistician* [online]. 1984, **33**(4) [cit. 2020-05-13]. DOI: 10.2307/2987742. ISSN 00390526. Dostupné z: https://www.jstor.org/stable/2987742?origin=crossref

[12] GUDIVADA, Venkat N. *Cognitive computing: theory and applications*. Amsterdam: Elsevier/North-Holland, North Holland is an imprint of Elsevier, [2016]. Handbook of statistics (Amsterdam, Netherlands), v. 35. ISBN 9780444637444.

[13] KARPAGAVALLI, S. a E. CHANDRA. A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing: Image Processing and Pattern Recognition*. 2016, **9**(4), 393-404. DOI: 10.14257/ijsip.2016.9.4.34.

[14] KIM, Chanwoo a Richard STERN. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. *Proc. Interspeech*. 2008, (1), 2598-2601.

[15] KLEINBAUM, David G. a Mitchel KLEIN. *Logistic Regression* [online]. New York, NY: Springer New York, 2010 [cit. 2020-04-28]. Statistics for Biology and Health. DOI: 10.1007/978-1-4419-1742-3. ISBN 978-1-4419-1741-6.

[16] Lecture 13: Simple Linear Regression in Matrix Format [online]. 2015 [cit. 2020-04-28]. Dostupné z: https://www.stat.cmu.edu/ cshalizi/mreg/15/lectures/13/lecture-13.pdf

[17] LEE, Chin-Hui. Speech Recognition and Production by Machines. *International Encyclopedia of the Social & Behavioral Sciences* [online]. Elsevier, 2015, 2015, s. 259-263 [cit. 2020-05-17]. DOI: 10.1016/B978-0-08-097086-8.52023-6. ISBN 9780080970875. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/B9780080970868520236

[18] MathWorks. *pinv* [online]. [cit. 2020-04-27]. Dostupné z: https://www.mathworks.com/help/matlab/ref/pinv.html

[19] MORNSTEIN, Vojtěch. *Lékařská fyzika a biofyzika*. Brno: Masarykova univerzita, 2018. ISBN 978-80-210-8984-6.

[20] NASR, George E., E. A. BADR a C JOUN. *Cross Entropy Error Function in Neural Networks: Forecasting Gasoline Demand.* [online]. AAAI Press, 2002, s. 381-384 [cit. 2020-04-29]. ISBN 1-57735-141-X. Dostupné z: http://www.aaai.org/Library/FLAIRS/2002/flairs02-075.php

[21] NG, Andrew. CS229 Lecture notes [online]. [cit. 2020-04-27]. Dostupné z: http://cs229.stanford.edu/notes/cs229-notes1.pdf

[22] NIELSEN, Michael A. *Neural networks and deep learning.* Determination Press, 2015.

[23] *NIST SClite Scoring Package* [online]. [cit. 2020-04-05]. Dostupné z: http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

[24] OPPENHEIM, A.V. a R.W. SCHAFER. Dsp history - From frequency to quefrency: a history of the cepstrum. *IEEE Signal Processing Magazine* [online]. 2004, **21**(5), 95-106 [cit. 2020-04-18]. DOI: 10.1109/MSP.2004.1328092. ISSN 1053-5888. Dostupné z: http://ieeexplore.ieee.org/document/1328092/

[25] PAPAIOANNOU, Iason, Sebastian GEYER a Daniel STRAUB. Improved cross entropy-based importance sampling with a flexible mixture model. *Reliability Engineering & System Safety* [online]. 2019, **191** [cit. 2020-04-27]. DOI: 10.1016/j.ress.2019.106564. ISSN 09518320. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0951832019301528

[26] PASCAL, Jérôme, Antoine BOURGEADE, Michel LAGIER a Claude LEGROS. Linear and nonlinear model of the human middle ear. *The Journal of the Acoustical Society of America* [online]. 1998, **104**(3), 1509-1516 [cit. 2020-04-28]. DOI: 10.1121/1.424363. ISSN 0001-4966. Dostupné z: http://asa.scitation.org/doi/10.1121/1.424363

[27] CHRIS, Piech. Logistic Regression [online]. 2016 [cit. 2020-05-23]. Dostupné z: https://web.stanford.edu/class/archive

[28] PSUTKA, Josef. *Mluvíme s počítačem česky.* Praha: Academia, 2006. Česká matice technická (Academia). ISBN 80-200-1309-1.

[29] SALHI, Lotfi a Adnane CHERIF. Robustness of Auditory Teager Energy Cepstrum Coefficients for Classification of Pathological and Normal Voices in Noisy Environments. *The Scientific World Journal* [online]. 2013, **2013**, 1-8 [cit. 2020-04-28]. DOI: 10.1155/2013/435729. ISSN 1537-744X. Dostupné z: http://www.hindawi.com/journals/tswj/2013/435729/

[30] SELJAN, S.; DUNDER, I. *Combined Automatic Speech Recognition and Machine Translation in Business Correspondence Domain for English-Croatian*, World Academy of Science, Engineering and Technology International Journal of Industrial and Systems Engineering, 2014

[31] SELTZER, Michael, Yun-Cheng JU, Ivan TASHEV, Ye-Yi WANG a Dong YU. In-Car Media Search. *IEEE Signal Processing Magazine* [online]. 2011, **28**(4), 50-60 [cit. 2020-04-27]. DOI: 10.1109/MSP.2011.941065. ISSN 1053-5888. Dostupné z: http://ieeexplore.ieee.org/document/5888651/

[32] SHANNON, Claude Elwood a Warren WEAVER. *The mathematical theory of communication*. Chicago: University of Illionois Press, 1998. ISBN 0252725484.

[33] SIGURDSSON, Sigurdur, Kaare Brandt PETERSEN a Tue LEHN-SCHIØLER. Mel Frequency Cepstral Coefficients: An Evaluation of Robustness of MP3 Encoded Music. *ISMIR*. 2006, , 286-280.

[34] SCHWARZ, P. *Phoneme recognition based on long temporal context*, Doctoral thesis, Brno, Brno University of Technology, Faculty of Information Technology, 2008

[35] WONG, Tzu-Tsung. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* [online]. 2015, **48**(9), 2839-2846 [cit. 2020-05-05]. DOI: 10.1016/j.patcog.2015.03.009. ISSN 00313203. Dostupné z: https://linkinghub.elsevier.com/retrieve/pii/S0031320315000989

[36] XIN, Yan a Gang Su XIAO. *Linear Regression Analysis: Theory and Computing*. World Scientific Publishing Co. Pte., 2009. ISBN 981-283-410-9.

[37] YE-YI WANG, DONG YU, YUN-CHENG JU a A. ACERO. An introduction to voice search. *IEEE Signal Processing Magazine* [online]. 2008, **25**(3), 28-38 [cit. 2020-04-26]. DOI: 10.1109/MSP.2008.918411. ISSN 1053-5888. Dostupné z: http://ieeexplore.ieee.org/document/4490199/

[38] YU, Dong a Deng LI. Automatic Speech Recognition: A Deep Learning Approach. London: Springer, 2015. ISBN 978-1-4471-5779-3.

[39] ZEDEK, Martin a Petr SYSEL. *Fonetická transkripce českého jazyka*. 2014.

[40] ZECHNER, Klaus a Alex WEIBEL. Minimizing Word Error Rate in Textual Summaries of Spoken Language [online]. Pittsburgh,USA, 2002 [cit. 2020-04-18]. Dostupné z:

https://www.researchgate.net/publication/2539978_Minimizing_Word_Error_Rate_in_Textual_Summaries_of_Spoken_Language

[41] ZIOLKO, A, Jakub GAŁKA, Suresh MANANDHAR a Richard WILSON. *Triphone Statistics for Polish Language*. 2007, , 63-73. DOI: 10.1007/978-3-642-04235-5_6.

[42] ZOTTOLA, Tino. *Vacuum Tube and Guitar and Bass Amplifier Servicing*. The Bold Strummer, 1995. ISBN 0-933224-97-4.

[43] ZOU, Kelly H., Kemal TUNCALI a Stuart G. SILVERMAN. Correlation and Simple Linear Regression. Radiological Society of North America. 2003(Vol. 227, 3). DOI: 10.1148/radiol.2273011499. ISSN 1527-1323.