



## Projekt

# Určení kvality řečové nahrávky bez reference

*Studijní program:*

*Studijní obor:*

*Autor práce:*

*Vedoucí práce:*

B0613A140005 – Informační technologie

B0613A140005IS – Inteligentní systémy

**Viktoriiia Sergeeva**

Ing. Jiří Málek Ph.D.

Liberec 2023

Tento list nahradte  
originálem zadání.

## Prohlášení

Prohlašuji, že svůj projekt jsem vypracovala samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mého projektu a konzultantem.

Jsem si vědoma toho, že na můj projekt se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

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

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

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

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

10. 3. 2023

Viktoriiia Sergeeva

## Určení kvality řečové nahrávky bez reference

### Abstrakt

V tomto projektu představujeme nový přístup ke zjišťování poměru signálu k šumu (SNR) daného signálu bez použití referencí. Naše metoda využívá neuronovou síť, která je natrénována a vyhodnocena na množině zarušených řečových signálů. Soubor dat je vytvořen uměle, z každého nezkresleného řečového signálu je vytvořeno několik zarušených variant s různou úrovní šumu. Prostřednictvím experimentů a vyhodnocení jsme prokázali, že naše síť je schopna detekovat SNR v reálných scénářích obsahujících vysokou přesnost a má velký potenciál pro praktické aplikace v oblastech, jako jsou telekomunikace a zpracování signálů.

**Klíčová slova:** SNR, neuronové sítě, Signal To Noise Ratio, lokální SNR, kvalita řečové nahrávky

## Determining the quality of a speech recording without reference

### Abstract

In this project, we present a novel approach to detecting the signal-to-noise ratio (SNR) of a given signal without the use of references. Our method employs a neural network, which has been trained and evaluated on a dataset of corrupted speech signals. The data set is artificially created, with several distorted variants of each uncorrupted speech signal generated with varying levels of noise. Through experimentation and evaluation, we demonstrate that our network is able to detect SNR in real-world scenarios with high accuracy and has a strong potential for practical applications in fields such as telecommunications and signal processing.

**Keywords:** SNR, neural networks, Signal To Noise Ratio, local SNR, speech recording quality

## Poděkování

Rádá bych poděkovala svému vedoucímu, Ing. Jiřímu Málkovi, PhD., za jeho trpělivost, důkladné vysvětlení materiálů a veškerou pomoc při tvorbě tohoto bakalářského projektu.

# Obsah

Seznam zkratek . . . . .	7
<b>1 Úvod a motivace</b>	<b>8</b>
<b>2 Formální popis problému</b>	<b>11</b>
2.1 Existující metody pro odhad SNR bez reference . . . . .	12
<b>3 Popis řešení</b>	<b>13</b>
3.1 Co jsou hluboké neuronové sítě? . . . . .	13
3.1.1 Váhy a bias . . . . .	13
3.2 Konvoluční sítě . . . . .	14
3.2.1 Motivace k použití . . . . .	14
3.3 Melův spektrogram . . . . .	15
3.4 Architektura sítě . . . . .	16
3.5 Trénovací a testovací dataset . . . . .	16
<b>4 Experimentální ověření</b>	<b>18</b>
4.1 Ztrátová funkce na trénovacích a testovacích datech . . . . .	18
4.2 Jak byla vypočítaná přesnost? . . . . .	19
4.3 Jak byla vypočítaná RMSE? . . . . .	19
4.4 Analýza výsledků provedení testů . . . . .	20
<b>5 Závěr</b>	<b>21</b>
<b>Použitá literatura</b>	<b>22</b>

## Seznam zkratek

<b>SNR</b>	Poměr signálu k šumu (Signal To Noise Ratio)
<b>PESQ</b>	Percepční vyhodnocení srozumitelnosti řeči (Perceptual Evaluation Of Speech Quality)
<b>STOI</b>	Krátkodobá objektivní míra srozumitelnosti (Short-Time Objective Intelligibility)
<b>POLQA</b>	Percepční objektivní analýza kvality poslechu (Perceptual Objective Listening Quality Analysis)
<b>SIR</b>	Poměr signálu k rušení (Signal To Interference Ratio)
<b>T60</b>	Doba dozvuku (Reverberation time)
<b>LAN</b>	Místní síť (Local Area Network)
<b>RF</b>	Rádiová frekvence (Radio Frequency)
<b>WADA</b>	Analýza rozložení amplitudy tvaru vlny (Waveform amplitude distribution analysis)
<b>VAD</b>	Detektor hlasové aktivity (Voice activity detector)
<b>RMSE</b>	Efektivní hodnota kvadratické chyby (Root Mean Squared Error)
<b>MSE</b>	Střední kvadratická chyba (Mean squared error)

# 1 Úvod a motivace

Analýza řečových signálů je důležitou oblastí strojového učení. Signály obsahují cenné informace, které jsou však často zkresleny šumem. Toto rozdělení není stálé: například, při vytváření hlasového ovládání je důležitý pouze hlas mluvčího a všechny ostatní zvuky a hlasy jsou rušivé. Protipříkladem může být situace, kdy při analýze zvukového signálu není důležitý pouze hlas mluvčího, ale je nutné zachytit i jiné zvuky a hlasy v okolí. Například při rozpoznávání zvuků v prostředí jako je ulice, je důležité zachytit nejen hlas člověka, ale i zvuky vozidel, sirén apod., které mohou mít význam. Určení míry kvality signálu je klíčové pro úspěšné zpracování signálu. Čím kvalitnější je signál, tím jednodušší je práce s ním. Často musíme řešit úlohu převodu řeči do textu: pokud není řeč v takovém signálu zkreslená šumem, převod do textu proběhne téměř bezchybně, ale v zašuměném prostředí se přesnost přepisu sníží.

Většina užitečných signálů jsou náchylné k rušení šumem, reverberací a hlukem v pozadí. Šum je zvuk nechtěných okolních zdrojů, které se mísí s užitečným signálem, ale není součástí jeho užitečné informace. Reverberace je zpětný odraz zvuku, který se projevuje jako postupné odeznívání zvuku po skončení přehrávání signálu.

Informace o srozumitelnosti řeči mohou být užitečné v automobilovém průmyslu pro snížení hluku a zlepšení zvukového výkonu, v lékařském průmyslu pro přizpůsobení sluchadel a kochleárních implantátů a v hudebním průmyslu pro odstranění nežádoucího hluku a zlepšení kvality skladby, jakož i pro automatické nastavení úrovně hlasitosti.

Metrik pro měření kvality řečového signálu je celá řada: percepční ověření srozumitelnosti řeči (PESQ, [1, 2]), percepční objektivní analýza kvality poslechu (POLQA, [3]), poměr signálu k rušení (SIR), doba dozvuku (T60, [4]) a poměr signálu k šumu (SNR).

Všechny uvedené metriky potřebují referenční nezkreslený signál, a proto jejich výpočet je možný jen v laboratorních podmínkách. V praxi máme pouze zkreslený signál a je třeba metriky nějak odhadnout bez reference. Například neuronovou sítí.

## 1. PESQ


Je to široce používaná metrika kvality řeči, která simuluje lidské sluchové vnímání. Porovnává původní a poškozený řečový signál a vypočítá číslo v rozsahu od -0,5 do 4,5, které odráží jejich rozdíl. Vyšší hodnoty znamenají lepší kvalitu řečového signálu a naopak, nižší hodnoty znamenají horší kvalitu. Používá se pro analýzu časové deformace, proměnného zpoždění, překódování a šumu v nahrávce.



Pro hodnocení kvality signálu metrika zohledňuje následující charakteristiky:


- **Ostrost zvuku** – ovlivňuje jasnost hlasu,
- **Objem hovoru** – ovlivňuje srozumitelnost hlasu,
- **Hluk na pozadí** – ovlivňuje kvalitu hlasu,
- **Poměrné zpoždění** – ovlivňuje plynulost hlasu,
- **Vystřihování** – ovlivňuje kontinuitu hlasu,
- **Zvukový šum** – ovlivňuje čistotu hlasu

## 2. POLQA

Algoritmus POLQA se zaměřuje na hodnocení kvality hlasového signálu v širokopásmovém rozsahu. Tento algoritmus porovnává referenční signál  $X(t)$  se signálem  $Y(t)$ , který prochází různými komponentami komunikačního systému, jako jsou kódování, dekódování, LAN a  komponenty. Výsledkem je předpověď vnímané kvality, kterou by signálu  $Y(t)$  přisoudily osoby při subjektivním poslechovém testu. POLQA rozsahy jsou vyjádřeny na stupnici od 1 do 5, kde 1 představuje nejnižší kvalitu a 5 nejvyšší kvalitu. POLQA se tak vyznačuje vysokou úrovní přesnosti a robustnosti v hodnocení kvality hlasového signálu

## 3. T60

Doba dozvuku, také známá jako T60, je časový údaj, který určuje, jak dlouho posluchač uslyší zvuk, když jeho zdroj přestane být aktivní. T60 závisí na řadě faktorů, jako je objem prostoru, plocha a povaha povrchu, které ovlivňují úroveň pohlcování a odrazu zvuku.

- **Objem:** místnosti s větším objemem mají tendenci mít delší dobu T60, protože větší prostor poskytuje více prostoru pro odrazy zvuku .
- **Plocha povrchu:** při konstantní hlasitosti se T60 snižuje s nárůstem plochy, která je k dispozici pro odrazy, což znamená větší pohlcování zvuku,
- **Povaha plochy:** pohltivost a drsnost povrchu hrají důležitou roli v ovlivnění T60. Měkké, porézní povrchy (jako záclony, koberce nebo čalouněné židle) pohlcují více zvukové energie než tvrdé, neporézní povrchy. Drsný povrch také rozptýlí zvuk do více směrů

T60 je důležitým parametrem pro posouzení kvality akustiky prostoru. Jeho hodnota se používá pro návrh a úpravy prostoru tak, aby byla dosažitelná požadovaná kvalita zvuku uvnitř.

#### 4. SNR

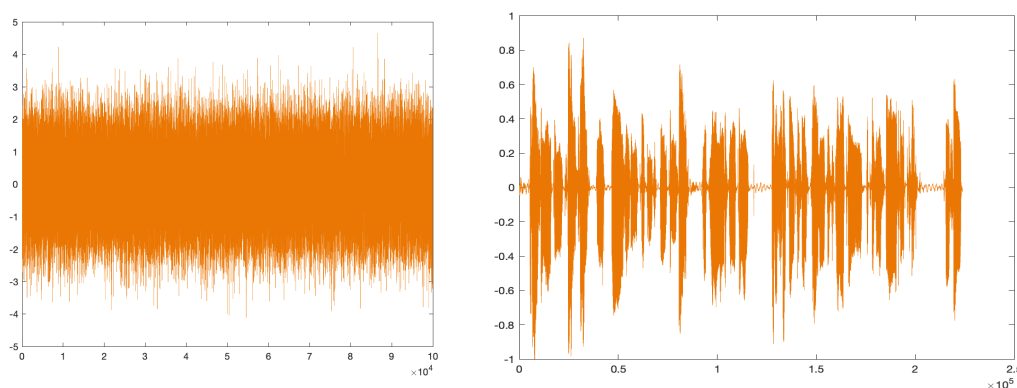
V tomto projektu se zaměříme na SNR. SNR, na rozdíl od perceptuálních mír PESQ a POLQA, je míra objektivní. Rozdíl je v tom, že měří skutečné energie řeči a šumu a nesnaží se je interpretovat tak, jak je vnímá člověk. Jedná se o bezrozměrnou hodnotu, která se rovná poměru výkonu užitečného signálu k výkonu šumu. Často se vyjádří v decibelech. Ačkoli se SNR obvykle udává pro elektrické signály, lze ji použít pro jakoukoli formu signálu, jako jsou hladiny izotopů v ledovém jádře, biochemické signály mezi buňkami nebo signály finančního obchodování. Definice SNR říká že musíme mít referenci signálu, avšak ji lze zajistit jen při testování v laboratoři. Existuje několik metod pro odhad SNR bez reference, na které se zaměříme v kapitole 2.1.

## 2 Formální popis problému

Pro použití metriky SNR mimo laboratoře se používá několik metod:

1. **WADA** – funguje s Gaussovským šumem
2. **VAD**
3. **End-to-end systém**

Metody (1) a (2) jsou vhodné pouze pro stacionární signály (spektrum takového signálu se nemění), zatímco pro nestacionární signály se využívá metoda (3).



Obrázek 2.1: Příklady stacionárního i nestacionárního signálů

Abychom mohli vypočítat SNR, musíme vědět, jaká část signálu je užitečná a jaká neužitečná.

Globální SNR se vypočítá pro celý signál:

$$SNR = 10 \log_{10} \frac{\sigma_s^2}{\sigma_v^2}, \quad (2.1)$$

kde  $\sigma_s^2$  je průměrný výkon signálu a  $\sigma_v^2$  je průměrný výkon zašumění.

Lokální SNR se vypočítá po částech signálu:

$$SNR_i = 10 * \log_{10} \frac{\sigma_{s,i}^2}{\sigma_{v,i}^2}, \quad (2.2)$$

kde  $\sigma_{s,i}^2$  a  $\sigma_{v,i}^2$  jsou průměrné výkony signálu a zašumění v i-tem rámci.

Jak globální, tak i lokální SNR podle vzorků musí mít reference signálu. Reference signálu je předem známá informace o tom, která část signálu je užitečná a která je neužitečná. Tvoříme metodu, která bude odhadovat SNR bez reference. Taková metoda je na rozdíl od odhadu s referencí prakticky použitelná.

## 2.1 Existující metody pro odhad SNR bez reference

Takovými metodami jsou například WADA a metoda založená na detekci řečové aktivity.

- Algoritmus WADA [5] předpokládá, že šum v testovaném signálu je stacionární, že amplitudové rozložení čisté řeči lze aproximovat Gama rozložením s tvarovacím parametrem 0,4 a že aditivní šumový signál má Gaussovske rozložení. Na základě tohoto předpokladu se SNR odhaduje zkoumáním amplitudového rozložení šumem zkreslené řeči,
- VAD [6, 7] je metoda, která zjišťuje přítomnost lidské řeči v audio signálu. Tato technika určí intervaly, odpovídající řečové aktivitě. Detektor nalezne okamžiky, kde není řeč aktivní a v těchto okamžicích odhadne výkon šumu. Nejprve spočítá energii každého vzorku signálu, poté se stanoví referenční hodnota výkonu šumu jako průměrná energie v prvních vzorcích, které neobsahují řeč. Tato metoda se používá většinou pro stacionární šum. Při práci s dynamickým šumem, je nutné tuto hodnotu adaptivně měnit, což nezajišťuje dostatečně kvalitní výsledky. Vzorek se označuje za řečový, pokud jeho energie překročí stanovený práh. VAD se uplatňuje v mnoha aplikacích, jako jsou hlasově ovládané zařízení jako chytré telefony, hlasové asistenty a další
- End-to-end systém je metoda realizovaná pouze neuronovou sítí. V této metodě neprobíhá žádné další zpracování. My se snažíme o vytvoření metody použitelné i pro nestacionární šum a zvolili jsme end-to-end systém jako řešení našeho problému.

## 3 Popis řešení

### 3.1 Co jsou hluboké neuronové sítě?

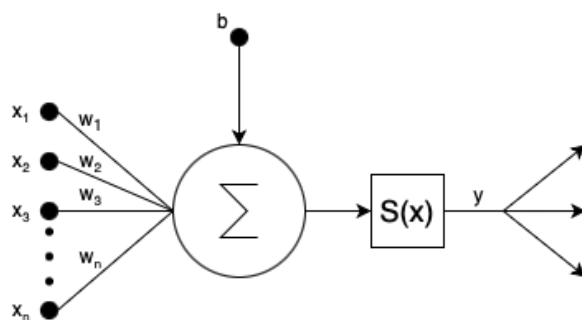
Neuronové sítě jsou mocným nástrojem pro modelování složitých vzorců v datech. Jsou inspirovány strukturou a funkcí lidského mozku a skládají se ze vzájemně propojených uzlů neboli umělých neuronů, které zpracovávají informace prostřednictvím vážených spojení.

Na vysoké úrovni neuronové sítě přijímají vstupní data, šíří je sítí umělých neuronů a vytvářejí výstup. Váhy spojení mezi neurony se upravují v procesu učení, což síti umožňuje přizpůsobit se vstupním datům a provádět předpovědi nebo rozhodnutí na základě informací, na kterých byla vyškolená.

#### 3.1.1 Váhy a bias

Váhy a bias jsou parametry, které se neuronová síť učí během trénování, aby mohla provádět předpovědi nebo rozhodnutí na základě vstupních dat. Řídí vliv každého vstupního prvku a slouží jako základní predikce.

Váhy jsou skalární hodnoty, které se používají k řízení vlivu jednotlivých vstupních prvků na výstup.



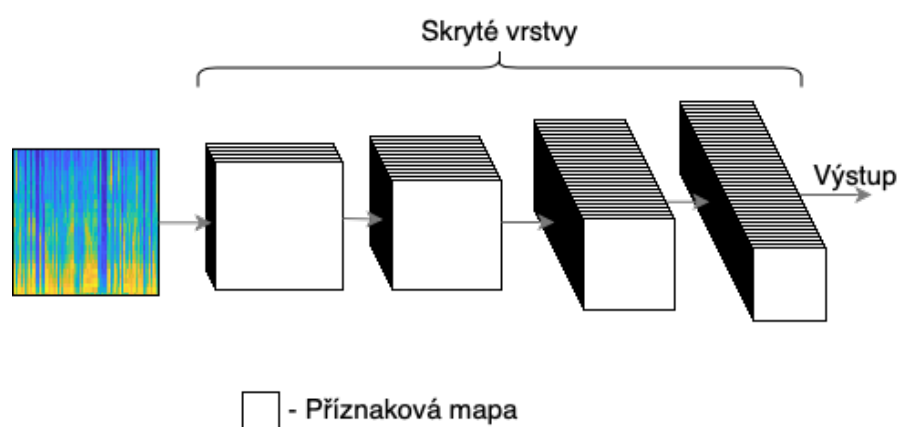
Obrázek 3.1: Schéma neuronu

Každé spojení mezi dvěma neurony v síti má svou vlastní váhu a tyto váhy se používají k výpočtu váženého součtu vstupů do každého neuronu, který je poté převeden přes aktivační funkci a vytváří výstup neuronu. Bias obsahuje skalární hodnoty, které se přičítají k váženému součtu vstupů do každého neuronu. Slouží jako "základní linie" předpovědi, i když jsou vstupy nulové.

V bakalářském projektu se používá hluboká neuronová síť k předpovídání úrovně SNR.

## 3.2 Konvoluční síť

Konvoluční neuronové sítě (CNN) se používají k rozpoznávání vzorů v obrazech, videích a zvukových vlnách. Využívají konvoluční vrstvy k detekci specifických rysů a další vrstvy, jako je sdružování a plně propojené vrstvy, ke zmenšování vzorků a transformaci výstupů. Nalezené rysy se pak používají pro analýzu vstupních obrázků a rozhodnutí, jestli oni mají specifické rysy a co oni znamenají. Trénování zahrnuje aktualizaci vah s cílem minimalizovat rozdíl mezi předpovídanými a skutečnými výstupy pomocí zpětného šíření.



Obrázek 3.2: Obyčejné schéma CNN

### 3.2.1 Motivace k použití

Jednou z klíčových vlastností CNN je jejich schopnost extrahovat prostorové informace z dat. V tomto projektu máme nestacionární šum - je to prostorově proměnlivá informace.

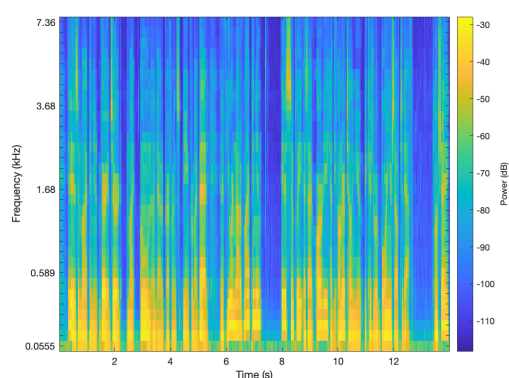
Konvoluční neuronové sítě ji dobře a rychle následující problémy analýzy obrázků:

1. Příliš mnoho výpočtů
2. Zachází s lokálními pixely stejně jako s pixely vzdálenými od sebe
3. Citlivé na umístění objektu v obraze

### 3.3 Melův spektrogram

Melův spektrogram je vizuální znázornění spektrální hustoty zvuku nebo řečového signálu. Zobrazuje spektrální energii zvuku na frekvenční ose Melovy stupnice, přičemž na ose x je čas a na ose y frekvence. Melova stupnice napodobuje způsob, jakým funguje lidské vnímání zvuku. Vyjadřuje frekvenci zvuku v podobě výšky tónu a má být intuitivnější a pro člověka snáze pochopitelná ve srovnání s lineárními frekvenčními stupnicemi.

V Melovém spektrogramu představují tmavší barvy vyšší úroveň energie (tj. větší amplitudu) na určité frekvenci a v určitém čase, zatímco světlejší barvy představují nižší úroveň energie. Tato reprezentace umožňuje vizualizovat vzorce ve zvuku a identifikovat důležité rysy, jako je výška tónu a rytmus.

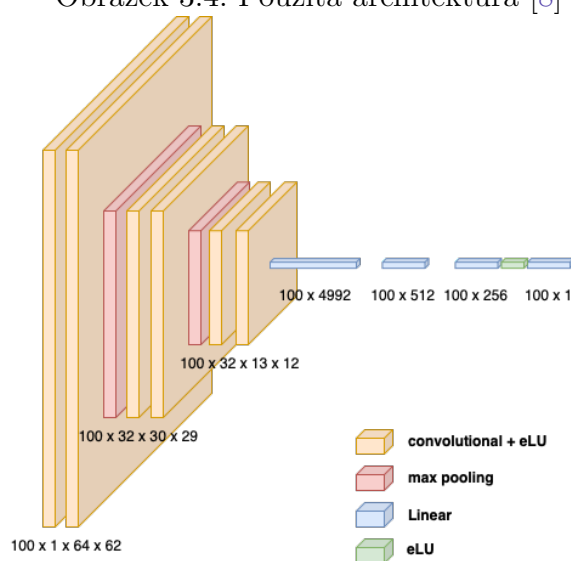


Obrázek 3.3: Příklad Melova spektrogramu

### 3.4 Architektura sítě

Na základě toho, že dataset obsahoval nestacionární signál, zvolili jsme vytvoření End-to-end systému. Byla použita hluboká neuronová síť [5] se třemi skrytými vrstvami (dvě skryté a jedna výstupní). Schéma vytvořené architektury je představeno na obrázku (3.4). Architektura se skládá ze tří skrytých vrstev a zvolená velikost jádra je  $5 \times 5$ . Po prvních dvou vrstvách následují max-pooling vrstvy. Velikosti jejich jader jsou stejné –  $2 \times 2$ . Přes celou síť je použita aktivační funkce eLU. Hyperparametry sítě jsou zvoleny na základě výsledků předběžných experimentů. Byla zvolena střední kvadratická chyba (MSE). Je to střední hodnota druhých mocnin rozdílu mezi skutečnou a predikovanou hodnotou. Vstupem do sítě je obrázek (Melův spektrogram) o velikosti  $64 \times 62$ . Výstupem je predikovaná hodnota SNR.

Obrázek 3.4: Použitá architektura [8]



### 3.5 Trénovací a testovací dataset

Audio signály byli zpracovány do Melovského spektrogramu po segmentech s počtem vzorků 512, délkou skoku 256 a počtem melů 64.

K vytvoření sady dat byla použita veřejně dostupná online knihovna LibriSpeech [9] se 100 hodinami čisté řeči. Na těchto zvukových nahrávkách čtou různé osoby úryvky z knih. Textové pasáže se neopakují, pohlaví a věk mluvčího jsou nezávislé. Poté byla použita další veřejně dostupná online knihovna se šumy [10]. Vybrali jsme pouze dva typy nestacionárního šumu: z veřejného náměstí a ulice. Kombinovali jsme řeč a šum při šesti úrovních globálního SNR: nekonečno (kvůli technickým vlastnostem reprezentované hodnotou 50), 20, 10, 5, 0 a -5. Pomocí následujícího vzorce jsme vypočítali násobitel šumu na základě požadované globální SNR:



$$z = 10^{\frac{-SNR}{20}} * \sqrt{\frac{\sum_{i=1}^N s_i^2}{\sum_{i=1}^N v_i^2}}, \quad (3.1)$$

kde  $z$  je výsledným násobitelem šumu,  $SNR$  je požadovaná úroveň kvality výsledného signálu,  $N$  je délka signálu,  $s$  je signál s řečí a  $v$  je signál se šumem, který budeme přidávat.

Lokální referenční hodnoty SNR v každé sekundě signálu byly poté vypočteny podle vzorce:

$$ref = 10 * \log_{10} \frac{\sum_{i=1}^N s_i^2}{\sum_{i=1}^N v_i^2}, \quad (3.2)$$

kde  $ref$  je lokální hodnota SNR,  $s$  je signál s řečí a  $v$  je signál se šumem, který budeme přidávat.

V procesu tvorby datasetu byla data rozdělena na tři části - trénovací, development a testovací, aby bylo zajištěno, že každý z nich bude obsahovat reprezentativní vzorky signálů s rovnoměrně rozloženými úrovněmi globálního SNR. Tato metoda rozdělení datasetu byla zvolena, aby bylo zajištěno, že vytvořený model bude schopen pracovat s různými úrovněmi SNR.

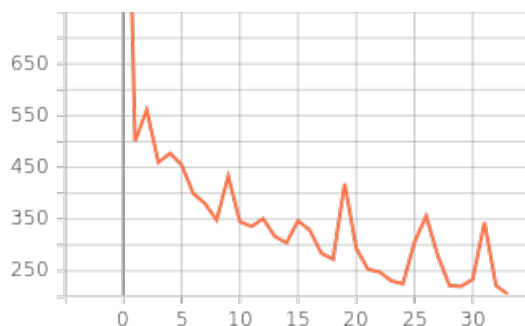
Aby bylo možné realizovat výše popsany algoritmus, byl použit programovací jazyk Matlab, který poskytuje potřebné nástroje a knihovny pro analýzu signálů. Tato volba umožnila snadnou implementaci a optimalizaci algoritmu pro rychlé a efektivní výpočty.

## 4 Experimentální ověření

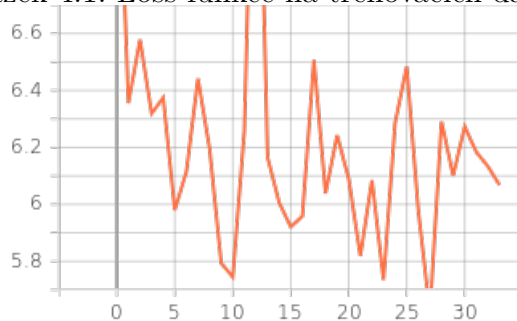
### 4.1 Ztrátová funkce na trénovacích a testovacích datech

Ztrátová funkce je silným nástrojem pro zobrazení toho, jak dobře neuronová síť pracuje a jak zvládá skrytá data. Loss funkce měří rozdíl mezi předpovědí sítě a zadanými vstupními daty. Úkolem trénování sítě je minimalizovat tuto funkci změnou vah a biasu.

Jak jsme očekávali, ztrátová funkce na trénovacích datech klesá. Skoky této funkce během trénování znamenají, že model buď může být příliš citlivý na změny v datech, nebo jsme měli příliš velký skok gradientu. I když je na obrázku 4.1 vidět, že trénovací loss bude klesat, trénování bylo pozastaveno. Kvůli průběhu ztrátové funkce na validační sadě dat můžeme předpokládat, že se síť začíná přeučovat. Zapamatovali jsme epochu s nejlepšími výsledky.



Obrázek 4.1: Loss funkce na trénovacích datech



Obrázek 4.2: Loss funkce na validačních datech

## 4.2 Jak byla vypočítaná přesnost?

Výpočet přesnosti modelu byl proveden v rozmezích, které byli pro vytvoření souboru dat zásadní. Šest úrovní globálního SNR při generování datasetu se stali hraničními SNR v rozmezích.

$$acc = \frac{HIT}{HIT + MISS} \cdot 100 \quad (4.1)$$

Kde  $acc$  je přesnost modelu v procentech % ve zvoleném rozmezí,  $HIT$  jsou správné předpovědi a  $MISS$  je suma všech špatných předpovědí.

## 4.3 Jak byla vypočítaná RMSE?

Místo celkového výpočtu RMSE jsme volili cestu selektivního spočítání v rámci stejně definovaných intervalů jako při určování přesnosti. RMSE se obvykle počítá ve stejných jednotkách jako data použitá k jejímu výpočtu. Pokud jsou například data měřena v metrech, bude rovněž v metrech.

$$RMSE = \sqrt{\frac{1}{I} \sum_{i=0}^I (\hat{s}_i - s_i)^2} \quad (4.2)$$

Kde  $RMSE$  je vyjádřena v decibelech,  $\hat{s}_i$  je predikovaná hodnota SNR v  $i$ té vteřině a  $s_i$  je očekávaná hodnota SNR.

## 4.4 Analýza výsledků provedení testů

Počítali jsme celkovou přesnost a průměrnou RMSE porovnáním odhadnuté a skutečné hodnoty SNR. Celková přesnost je 89,5%, průměrná RMSE je 3,9 a průměrná RMSE v užitečném rozsahu je 3,0.

Použití rozsahů s přesnými předpověďmi při výpočtu přesnosti poskytuje spolehlivější odhad výkonnosti modelu. Zohledňuje nejen to, zda je předpověď modelu správná, ale také to, jak blízko je skutečné hodnotě. Definování rozmezí přijatelné úrovně přesnosti umožňuje zohlednit variabilitu a šum v datech a umožňuje lépe pochopit výkonnost modelu.

Tabulka 4.1: Tabulka záměn a přesnosti klasifikaci v rámci intervalu

	$(\infty, 20)$	$\langle 20, 10 \rangle$	$\langle 10, 5 \rangle$	$\langle 5, 0 \rangle$	$\langle 0, -5 \rangle$	$\langle -5, -\infty \rangle$
$(\infty, 20)$	31264	1889	47	12	5	5
$\langle 20, 10 \rangle$	848	18958	1364	176	25	18
$\langle 10, 5 \rangle$	4	658	16807	899	151	61
$\langle 5, 0 \rangle$	0	50	1071	18293	821	357
$\langle 0, -5 \rangle$	1	10	44	1350	17802	1180
$\langle -5, -\infty \rangle$	1	20	32	104	1467	13998
přesnost	97,3%	87,8%	86,8%	87,8%	87,8%	89,6%
RMSE	3,9	3,9	2,5	2,4	2,5	8,6

Kromě toho umožňuje také nastavit práh přijatelné efektivní hodnoty kvadratické chyby a všechny předpovědi, které se vejdu do tohoto rozmezí, budou považovány za správné předpovědi, což zvyšuje celkovou přesnost. Byla definovaná přijatelná úroveň odchylky - 2.5 dB. Tento přístup umožňuje realističtější posouzení výkonnosti modelu v reálných scénářích, kde jsou dokonalé předpovědi vzácné a malé odchylky v předpovědích jsou tolerovatelné.

Jako alternativní vyhodnocení jsme se rozhodli vypočítat přesnost modelu pro dílčí intervaly, abychom určili, zda síť funguje pro všechny úrovně SNR zhruba stejně. Pro lepší názornost jsme vytvořili tabulku záměn.

I když přesnost není tak vysoká, RMSE ukazuje, že rozdíl mezi predikovanými a skutečnými hodnotami není tak velký, což může být důležité pro určité aplikace, kde malé odchylky jsou přijatelné.

## 5 Závěr

Cílem tohoto projektu bylo vyvinout síť pro určení poměru signálu k šumu daného signálu bez použití referencí. Síť byla natrénována na souboru dat signálů s různými úrovněmi SNR a vyhodnocena. Výsledky ukázaly, že vyvinutá síť je schopna předpovídat SNR signálů s vysokou přesností. Celkově lze říci, že tento projekt úspěšně vyvinul síť schopnou v reálném čase detekovat SNR signálu, kterou lze použít v různých oblastech, jako jsou telekomunikace a zpracování zvuku. Co se týče budoucí práce, plánuji pokračovat ve vývoji této sítě a rozšířit její možnosti v rámci své bakalářské práce. Rozšíříme síť tak, aby mohla odhadovat další metriky, a budeme porovnávat s existujícími přístupy k odhadu SNR bez reference.

## Použitá literatura

- [1] GAMPER, Hannes et al. Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. 2019, s. 85–89. Dostupné z DOI: [10.1109 / WASPAA .2019 . 8937202](https://doi.org/10.1109/WASPAA.2019.8937202).
- [2] RIX, Antony et al. Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In: 2001, sv. 2, 749–752 vol.2. ISBN 0-7803-7041-4. Dostupné z DOI: [10.1109/ ICASSP.2001.941023](https://doi.org/10.1109/ICASSP.2001.941023).
- [3] BEERENDS, John et al. Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I-Temporal Alignment. *AES: Journal of the Audio Engineering Society*. 2013, roč. 61, s. 366–384.
- [4] SMYTH, Tamara. *Music 175: Time and Space*. San Diego (UCSD), Department of Music, University of California, 2016.
- [5] GAZOR, S. a Wei ZHANG. A soft voice activity detector based on a Laplacian-Gaussian model. *IEEE Transactions on Speech and Audio Processing*. 2003, roč. 11, č. 5, s. 498–505. Dostupné z DOI: [10.1109/TSA.2003.815518](https://doi.org/10.1109/TSA.2003.815518).
- [6] MUŽÍČEK, Michal. *Robustní odhad odstupů řeči od šumu pomocí hlubokých neuronových sítí*. Liberec, 2016. Diplomová práce. Technická univerzita v Liberci.
- [7] JAT, Dharm Singh, Anton Sokamoto LIMBO a Charu SINGH. Chapter 6 - Voice Activity Detection-Based Home Automation System for People With Special Needs. In: DEY, Nilanjan (ed.). *Intelligent Speech Signal Processing*. Academic Press, 2019, s. 101–111. ISBN 978-0-12-818130-0. Dostupné z DOI: <https://doi.org/10.1016/B978-0-12-818130-0.00006-4>.
- [8] *How to Easily Draw Neural Network Architecture Diagrams* [online]. 2021. [cit. 2023-01-31]. Dostupné z: [https://miro.medium.com/v2/resize:fit:1400/ format:webp/1\\*kQtbGWZgi3n35Qojkg8cFw.png](https://miro.medium.com/v2/resize:fit:1400/format:webp/1*kQtbGWZgi3n35Qojkg8cFw.png).
- [9] *Libre speech* [online]. [cit. 2022-05-13]. Dostupné z: [https://www.openslr.org/ 12/](https://www.openslr.org/12/).
- [10] *Acoustic scene classification* [online]. 2019. [cit. 2023-02-07]. Dostupné z: <https://dcase.community/challenge2019/task-acoustic-scene-classification>.