



**VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ**

BRNO UNIVERSITY OF TECHNOLOGY

**FAKULTA INFORMAČNÍCH TECHNOLOGIÍ**

FACULTY OF INFORMATION TECHNOLOGY

**ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ**

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

**SEPARACE MLUVČÍCH V ČASOVÉ DOMÉNĚ**

TIME DOMAIN AUDIO SEPARATION

**BAKALÁŘSKÁ PRÁCE**

BACHELOR'S THESIS

**AUTOR PRÁCE**

AUTHOR

**JIŘÍ PEŠKA**

**VEDOUcí PRÁCE**

SUPERVISOR

**ing. KATEŘINA ŽMOLÍKOVÁ,**

**BRNO 2020**

## Zadání bakalářské práce



Student: **Peška Jiří**  
Program: Informační technologie  
Název: **Separace mluvčích v časové doméně pomocí neuronové sítě**  
**Time-Domain Neural Network Based Speaker Separation**  
Kategorie: Zpracování řeči a přirozeného jazyka

### Zadání:

1. Seznamte se s problémem separace mluvčích pomocí neuronových sítí.
2. Seznamte se s metodou TasNet pro jednokanálovou separaci signálu v časové doméně.
3. Implementujte danou metodu s využitím vhodného toolkitu (např. PyTorch, Keras).
4. Otestujte systém na vhodném datasetu. Zaměřte se na vyhodnocení vlivu velikosti sítě na přesnost.
5. Navrhněte a diskutujte možné zlepšení použité metody.

### Literatura:

- Luo, Yi, and Nima Mesgarani. "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.8 (2019): 1256-1266.
- dle doporučení vedoucího

Podrobné závazné pokyny pro vypracování práce viz <https://www.fit.vut.cz/study/theses/>

Vedoucí práce: **Žmolíková Kateřina, Ing.**

Vedoucí ústavu: Černocký Jan, doc. Dr. Ing.

Datum zadání: 1. listopadu 2019

Datum odevzdání: 14. května 2020

Datum schválení: 5. listopadu 2019

## Abstrakt

Práce se zabývá využitím konvolučních neuronových sítí pro automatickou separaci mluvčích v akustickém prostředí. Cílem je implementovat neuronovou síť podle architektury TasNet za použití frameworku pytorch, natrénovat síť s různými hodnotami hyperparametrů a porovnat kvalitu separací vzhledem k velikosti sítě.

Architektura oproti dosavadním metodám, které převáděly vstupní směs do časově-frekvenční reprezentace, používá konvoluční autoenkodér, který vstupní směs převádí do nezáporné reprezentace, která je optimalizovaná pro extrakci jednotlivých mluvčích. Samotné separace je docíleno aplikací masek, které jsou odhadnuty v separačním modulu. Modul tvoří opakující se posloupnost konvolučních bloků se zvyšující se dilatací, která pomáhá k modelování časových závislostí ve zpracovávané směsi.

K vyhodnocení přesnosti bylo použita metrika SDR (Sound to Distortion Ratio), která určuje poměr zastoupení šumu a zvuku v nahrávce. Natrénováním několika modelů s různými hodnotami hyperparametrů bylo možno zpozorovat závislost mezi velikostí sítě a hodnotou SDR. Zatímco menší síť dosahovala, po X epochách trénování, přesnosti XY, větší síť dosahovala až XX.

[[Jeste neco? 4. cast?]]

[[Doplnit SDR presnost do odstavce vyse.]]

## Abstract

[[Prelozit do anglictiny CZ abstrakt]]

## Klíčová slova

neuronové sítě, zpracování řeči, konvoluční neuronová síť, autoenkodér, separace mluvčích, strojové učení, tasnet, feed forward, hluboké učení

## Keywords

artificial neural networks, speech processing, convolutional neural networks, autoencoder, speech separation, machine learning, tasnet, feed forward, deep learning

## Citace

PEŠKA, Jiří. *Separace mluvčích v časové doméně*. Brno, 2020. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce ing. Kateřina Žmolíková,

# Separace mluvčích v časové doméně

## Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením ing. Kateřiny Žmolíkové. Další informace mi poskytli... Uvedl jsem všechny literární prameny, publikace a další zdroje, ze kterých jsem čerpal.

.....

Jiří Peška  
9. června 2020

## Poděkování

V této sekci je možno uvést poděkování vedoucímu práce a těm, kteří poskytli odbornou pomoc (externí zadavatel, konzultant apod.).

# Obsah

<b>1</b>	<b>Úvod</b>	<b>3</b>
<b>2</b>	<b>Separace mluvních</b>	<b>5</b>
2.1	Koktejl pártý efekt . . . . .	5
2.2	Metody pro separaci jednokanálových nahrávek . . . . .	5
2.2.1	CASA – computational auditory scene analysis . . . . .	5
2.2.2	NMF – non-negative matrix factorization . . . . .	6
2.2.3	Hluboké neuronové sítě . . . . .	6
2.3	Vyhodnocovací metriky . . . . .	7
2.3.1	Signal-to-noise ratio . . . . .	7
2.3.2	SDR – Source Distortion Ratio . . . . .	7
2.3.3	PESQ – Perceptual evaluation of speech quality . . . . .	7
2.3.4	STOI . . . . .	7
<b>3</b>	<b>Neuronové sítě</b>	<b>8</b>
3.1	Organizace feedforward sítí . . . . .	8
3.2	Umělý neuron . . . . .	9
3.3	Aktivační funkce . . . . .	11
3.4	Lineárně separovatelná data . . . . .	13
3.5	Konvoluční neuronové sítě . . . . .	15
3.5.1	Konvoluční vrstva . . . . .	15
3.6	Učení neuronových sítí . . . . .	15
3.6.1	Objektivní funkce . . . . .	16
3.6.2	Backpropagation . . . . .	16
3.6.3	Underfitting a generalizace . . . . .	17
3.6.4	Regularizace . . . . .	18
3.6.5	Residuální spojení . . . . .	18
3.6.6	Význam validační množiny v trénování . . . . .	18
<b>4</b>	<b>TasNet - Time-Domain Audio Separation Network</b>	<b>19</b>
4.1	Konvoluční auto-enzodér . . . . .	19
4.2	Separální modul . . . . .	20
4.2.1	Konvoluční bloky . . . . .	20
<b>5</b>	<b>Implementace sítě</b>	<b>22</b>
5.1	Implementace modelu . . . . .	23
5.2	Segmentace nahrávek . . . . .	24
5.3	Výpočet hodnoty loss za pomoci si-snr . . . . .	27

5.4	Pomocné skripty . . . . .	28
5.5	Trénovací prostředí . . . . .	29
<b>6</b>	<b>Experimenty a vyhodnocení</b>	<b>30</b>
6.1	Dataset . . . . .	30
6.2	Průběh trénování . . . . .	31
6.3	Experimenty s modely . . . . .	32
6.3.1	Vliv hyper-parametru X . . . . .	33
6.3.2	Vliv hyper-parametru R . . . . .	33
6.3.3	Vliv hyper-parametru délka-segmentu . . . . .	33
6.3.4	Vyhodnocení a porovnání modelů . . . . .	33
6.4	Možná rozšíření a navrhnutá vylepšení . . . . .	33
<b>7</b>	<b>Závěr</b>	<b>34</b>
	<b>Literatura</b>	<b>35</b>

# Kapitola 1

## Úvod

Zpracování řeči hraje v dnešní době důležitou roli v mnoha rozličných oborech. Mezi jeden z hlavních úkolů bezesporu patří separace zdrojů v zaznamenaném signálu, který může být složen ze signálů  $N$  mluvčích, ale i nechtěného hluku okolí. Vyřešení problému je předpoklad k dalším úkonům jako identifikace konkrétního mluvčího nebo třeba přepis konverzace na text. Se stále se zrychlujícím vývojem počítačů a s jejich zvyšujícím se výkonem se do popředí dostávají metody zpracování řeči založené na neuronových sítích, které v mnoha ohledech předčí ostatní algoritmy strojového učení.

Separace mluvčích v časové doméně dosahuje mimořádných výsledků v porovnání s dosavadními metodami založenými na převodu signálu z časové domény do frekvenční domény pomocí algoritmu STFT (Short-Time Fourier Transformation). Taková reprezentace signálu není optimalizovaná pro separaci řeči a nemusí pro tento úkol být optimální. V architektuře, která je navržena v referenční studii s názvem **TasNet: Surpassing Ideal Time-Frequency Masking for Speech Separation**[5], je vstupní signál převeden do nezáporné reprezentace, která je optimální pro extrakci jednotlivých mluvčích. Silnou stránkou systému je hluboká architektura sítě, která lépe modeluje dlouhodobé závislosti v signálu. Zároveň se ale musí vypořádat s problémy, které hluboké neuronové sítě mohou přinášet.

Téma v oblasti strojového učení a neuronových sítí mě zaujalo, protože tento obor zažívá obrovský rozmach a pomalu se stává součástí téměř všech odvětví, a je tudíž velmi perspektivní pro další výzkum. Právě bakalářskou práci jsem vyhodnotil jako dobrou příležitost k tomu, se seznámit se s neuronovými sítěmi a možnost vyzkoušet si, jak se s nimi pracuje, jak se implementují modely za pomoci frameworku a jak náročná je aplikace na nějaký reálný problém.

Úkolem této práce je nastudovat si problematiku neuronových sítí a jejich základní principy, seznámit se problémem separace mluvčích pomocí neuronových sítí a následně implementovat síť podle architektury TasNet pro separaci mluvčích v časové doméně, která byla navržena a popsána ve studii[5]. Poté tuto neuronovou síť natrénovat s různými kombinacemi hodnot hyperparametrů, které ovlivňují velikost sítě její vlastnosti, a nakonec porovnat přesnost a kvalitu separace mezi jednotlivými, různě velkými sítěmi a mezi výsledky studie. Přesnost a kvalita separace bude vypočítána pomocí metrik určených k hodnocení kvality řečového signálu. Síť budou testovány a vyhodnocovány na testovací množině jednokanálových směsí dvou mluvčích.

V první části práce, kterou pokrývá kapitola 3, jsou popsány základní prvky neuronových sítí, struktura umělého neuronu, jeho vstupy a výstupy, váhy a role aktivační funkce. V návaznosti na to jsou popsány neuronové sítě feedforward, které se skládají z vrstev neuronů. Dále je vysvětlen proces učení, který se skládá z několika kroků – výpočet výstupu sítě,

výpočet chyby a úprava vah, kde hraje podstatnou roli metoda backpropagation. Nakonec je zmíněn typ sítě feedforward – konvoluční neuronové sítě, které fungují na základě konvoluční operace. Konvoluční sítě se používají nejčastěji pro zpracování obrazu kvůli vlastnostem, které umožňují extrahovat příznaky s různou úrovní složitosti od základních útvarů jako úsečka, barva a podobně až po komplexnější příznaky - třeba celý obličej. Tohoto lze využít i při zpracování zvuku, kde jsou tyto extrahované příznaky jednorozměrné.

Druhá část, kapitola 4, je věnována architektuře TasNet. Prvně je zmíněn konvoluční auto-ekodér a jeho princip. Dál je popsána podoba separačního modulu, jeho princip a konvoluční bloky, ze kterých je poskládán. Tento blok obsahuje konvoluční vrstvy, normalizace a aktivační funkce. Bloky jsou skládány za sebe se zvyšující se časovou dilatací, čemuž se říká TCN (Temporal Convolutional Network).

Kapitola 5 se zabývá implementací neuronové sítě a jejím trénováním. Je popsána a zdůvodněna volba frameworku, implementace sítě a struktura zdrojového kódu. Pro usnadnění často se opakujících úkonů jsem vytvořil pár pomocných scriptů, které jsou zde také popsány. Model prošel během implementace několika úpravami. Pro účely trénování a validace byly vstupní nahrávky rozdělovány na čtyřsekundové segmenty. Pro testování byly používány nahrávky celé. V této kapitole je popsán průběh trénování sítě, výsledky a použité nástroje.

V poslední části, která je pokryta v kapitole 6 jsou shrnuty experimenty s modelem a vyhodnocení výsledků, v rámci něhož je zkoumán vliv hyper-parametrů na učení sítě, na výsledky a přesnost separace v závislosti na zvolených parametrech nebo počtu konvolučních bloků. Výstup sítě v podobě separovaných mluvčích je porovnán také s výsledky referenční studie. Kvalita separace, neboli přesnost, je vypočítána pomocí si-snr metriky která udává poměr zastoupení chtěného signálu a hluku na pozadí. Je zde vyhodnoceno, jaký vliv má velikost sítě na přesnost separace a jsou zde shrnuty jednotlivé výsledky.



## Kapitola 2

# Separace mluvčích

Živočichové mají vrozenou schopnost zaměřit se na jeden konkrétní zvuk, zatímco všechny ostatní dokážou potlačit. Naučit ale této dovednosti počítač se ukázalo jako obtížný úkol, který je překážkou ASR (*automatic speech recognition*) systémům v rozsáhlejší nasazení do běžného života. Tento problém se nazývá koktejl párty. Vyřešení tohoto problému, tedy separování zdrojů, umožňuje širokou aplikaci v oblasti zpracování řeči, jako přepis konverzací na text, ovládání počítače hlasem, používání hlasových asistentů či identifikace konkrétního mluvčího v prostředí s mnoha současně mluvícími lidmi.

### 2.1 Koktejl párty efekt

Tento problém se nejlépe dá popsat jako schopnost soustředit se na jeden konkrétní zdroj zvuku z mnoha souběžných zdrojů jako hudba, konkrétní hudební nástroj, řeč dalších lidí či okolní zvuky, a přepínat mezi nimi dynamicky pozornost.

**[[vzniklo pro jeho vyreseni hafo altofirmu, ktere dosahovaly nejakyh vysledku. jeho vyresenim by se mohlo pouzivat asr na prepis reci an text a pod. Nejlepsich vysledku dosahly hluboke neuronove site.]]**

### 2.2 Metody pro separaci jednokanálových nahrávek

Pro separaci zdrojových signálů z nahrávek směsí existuje mnoho algoritmů založených na zpracování signálů, z nichž asi nejznámější jsou CASA (*computational auditory scene analysis*) **[[ref?]]** a NMF (*non-negative matrix factorization*) **[[ref?]]**, používané pro separaci jednokanálových nahrávek. V posledních letech se do popředí dostaly přístupy založené na hlubokých neuronových sítích, které předčily svou přesností a výkonností dosavadní algoritmy. Pro separaci vícekanálových nahrávek se používají metody *beamforming* nebo state-of-the-art *Multi-channel blind source separation*, které ale jsou nad rámec této práce. Tato kapitola shrnuje některé metody pro řešení separace, které jsou zmíněny v článku[6].

#### 2.2.1 CASA – computational auditory scene analysis

Metoda CASA je založena na procesu v lidském mozku, který separuje zdroje ze směsi mluvčích a simuluje vysokoúrovňové chování lidského sluchu. Pro separaci jsou většinou manuálně navrhnutá segmentační pravidla pro operování nad nízkourovňovými příznaky k odhadu časově frekvenční ( $T$ - $F$ ) masky, která izoluje komponenty signálu jednotlivých

mluvčích a následně je použita pro rekonstrukci všech zdrojů. **[[Hodit sem teda ty zdroje jako v 3.1 pro ASA, CASA a pod? jak jsou tam asi 3 po sobe?]]**

Přestože tato metoda byla i nadále rozvíjena, tak má mnoho nevýhod jako špatnou generalizaci v důsledku manuálního vytváření pravidel, nemožnost automatického učení z dat nebo nemožnost použití na separaci jiných zdrojů než je lidská řeč, které značně omezují její použití v mnoha reálných případech.

### 2.2.2 NMF – non-negative matrix factorization

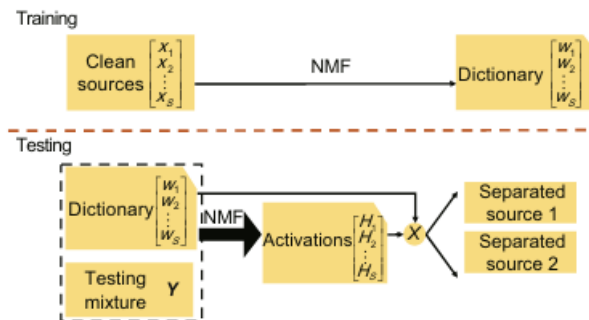
Nezáporná maticová faktorizace (NMF) **[[[(Lee and Seung, 2001)]]]** patří mezi daty řízené metody a je založena na předpokladu, že struktura spektrogramu může být reprezentována malým počtem bází.

Pro NMF platí

$$Y = \sum s W_s H_s \quad (2.1)$$

kde každý zdroj  $s$  je modelován aproximací nízkého řádu nezápornými maticemi  $W_s$ , která reprezentuje slovník a  $H_s$ , která reprezentuje aktivační funkci. Ty jsou sečteny a dohromady tvoří výslednou směs  $Y$ .

Metoda prochází trénovací a testovací fází. V trénovací fázi je každý zdroj dekomponován a mapován na množinu bází a aktivací a tím je zformován slovník  $W$  pro tento zdroj. V testovací fázi jsou naučené slovníky jednotlivých zdrojů  $s$  spojeny do jednoho, který dále není upravován. Následuje optimalizace aktivační matice  $H$  pro každý zdroj. Každý zdroj je následně ze směsi rekonstruován s použitím výsledných bází a aktivací.

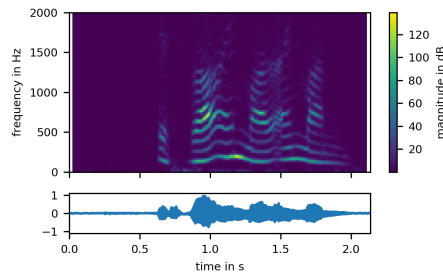


Obrázek 2.1: Znázornění trénovací a testovací fáze. Obrázek byl inspirován [6]

Stejně jako metoda CASA, i tato má mnoho nevýhod, které omezují její použití pro řešení reálných problémů separace.

### 2.2.3 Hluboké neuronové sítě

Nejlépe si při řešení problému koktejl párty vedou postupy založené na hlubokém učení a neuronových sítích. Metody převáděly směs do časově-frekvenční reprezentace znázorněné na obrázku 2.2 pomocí STFT (*short time fourier transformation*).



Obrázek 2.2: Časově–frekvenční reprezentace řeči, neboli spektrum

## 2.3 Vyhodnocovací metriky

Pro vyhodnocení kvality separace se používají různé metriky. Jejich volba závisí na řešeném problému.

### 2.3.1 Signal–to–noise ratio

Nezákladnější metrika pro určení kvality separace je signal–to–noise ratio (SNR), která porovnává zastoupení chtěného signálu a šumu v pozadí a je definována jako poměr energie originálního (ideálního) signálu a energie rozdílu mezi originálním a rekonstruovaným signálem.

$$SNR = \frac{E_s}{E_N} \quad (2.2)$$

Výsledný poměr často bývá vyjádřen v decibelech jako

$$SNR(dB) = 10 \log_{10} \frac{E_s}{E_N} \quad (2.3)$$

kde kladná SNR hodnota udává, že zastoupení signálu je větší, než zastoupení šumu [[citovat <https://labrosa.ee.columbia.edu/dpwe/pubs/Ellis04-sepeval.pdf>]].

### 2.3.2 SDR – Source Distortion Ratio

SDR je jedna z neznámějších metrik pro měření kvality řečového signálu.

### 2.3.3 PESQ – Perceptual evaluation of speech quality

Procentuální vyhodnocení kvality řeči (PESQ[7]) je standardizovaná metrika pro vyhodnocování kvality řeči. Jeho hodnota se pohybuje od -0.5 do 4.5 s tím, že vyšší hodnota znamená lepší procentuální kvalitu[4].

### 2.3.4 STOI

Bylo navrženo mnoho konvenčních přístupů a metod pro řešení problému koktejl párty a s ním spojené separace mluvčích, ale většina z nich nedosahuje dostatečné přesnosti nebo výkonu k použití v reálných situacích. Přístupy založené na hlubokém učení a neuronových sítích předčily dosavadní techniky jak v přesnosti, tak ve výkonu, což umožňuje jejich nasazení v běžném životě.

## Kapitola 3

# Neuronové sítě

V dnešní době zažívají neuronové sítě díky výkonosti počítačů velký rozmach. Jejich využití prostupuje skrze mnohé vědní obory a dokáží řešit celou řadu problémů, ve kterých dosahují výborných výsledků, které zdaleka předčily dosavadní postupy.

Neuronové sítě (*artificial neural networks*) jsou výpočetní model, který je inspirovaný strukturou lidského mozku, ve kterém je obrovské množství propojených a komunikujících neuronů. Ty se skládají ze vstupních dendridů, výstupních axonů a samotného těla neuronu. Na základě vnitřního potenciálu a vstupních hodnot je po přesažení prahové hodnoty neuron vybuděn a je vyslán signál na výstupní axon. Signál je nakonec předán dalším neuronům skrze jejich vstupní dendridy[3, p. 65–66].

Účelem neuronové sítě je naučit se plnit zadanou úlohu. Rozdíl oproti běžným algoritmům je ale ten, že způsob, jakým síť má problém řešit, není explicitně naprogramován, ale je postupně naučen. Základní způsoby učení jsou s učitelem (*supervised*) a bez učitele (*unsupervised*).

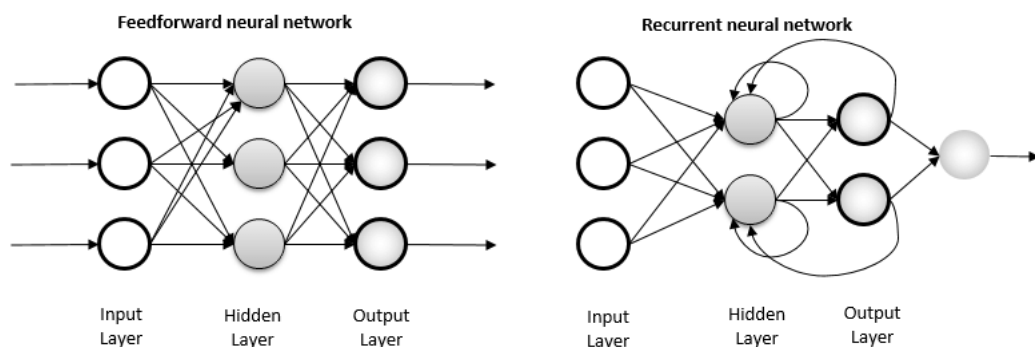
**Učení s učitelem**, pod které spadá tato práce, spočívá v **[[bla bla (dve tri vety)]]**.

Mezi problémy, které se dají řešit neuronovými sítěmi patří klasifikační a regresní problémy. Konkrétní příklad z oblasti klasifikace může být rozpoznávání objektů na obraze, psaného písma nebo detekce obličejů na videu, ale i mnohé aplikace ve zpracování řeči. Některé z problémů lze řešit sítí, která se skládá z jednotek neuronů, zatímco složitější problémy vyžadují mnohem větší kapacitu sítě a ta pak může obsahovat i tisíce neuronů.

Upravovat samotnou strukturu a chování neuronové sítě lze pomocí jejích hyperparametrů, což jsou parametry určující nastavení neuronové sítě a trénovacího algoritmu. Tyto parametry musí být určeny před začátkem trénování a většinu nelze později měnit. Mohou určovat kapacitu modelu, velikost záběrného pole, velikost filtrů, ale i regulovat samotný proces učení nastavením počtu epoch, po které se model má učit, nebo také počet dat předaný v jedné dávce (tzv. *minibatches*) během učení sítě.

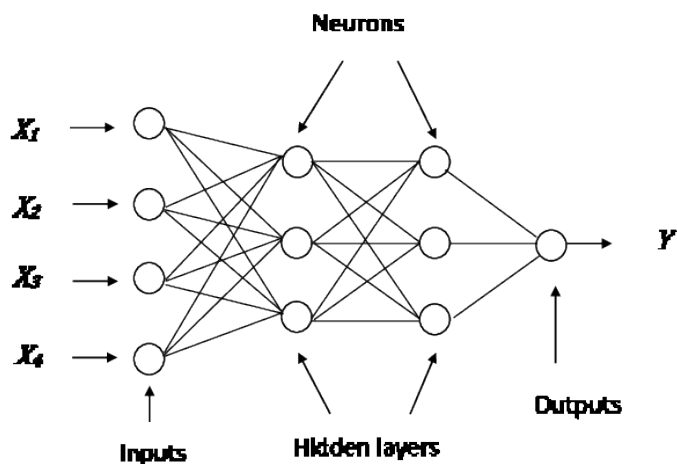
### 3.1 Organizace feedforward sítí

Feed forward neuronové sítě (*Multi Layer Perceptron – MLP*) jsou typ umělých neuronových sítí, kde se nevyskytují cykly ve výpočetním grafu, tedy zpětná propojení vrstev, takže informace se pohybuje pouze jedním směrem, od vstupní vrstvy přes skryté vrstvy až po vrstvu výstupní. Sítě, které obsahují cykly, se nazývají rekurentní. Rozdíl znázorňuje obrázek 3.1.



Obrázek 3.1: Příklad grafu feed forward sítě a rekurentní neuronové sítě. Lze si všimnout orientace šipek u feed forward sítě, které směřují pouze jedním směrem, zatímco u rekurentní sítě šipky směřují i k předešlým uzlům grafu.

Struktura neuronové sítě je organizována do vrstev, které se skládají z neuronů. Feed-forward síť je tvořena třemi typy vrstev (viz obrázek 3.2). Každá vrstva může obsahovat až  $n, n \in \mathbb{N} \setminus \{0\}$  neuronů. Vstupní vrstva slouží k předání hodnot do sítě, ale nijak tyto hodnoty nemodifikuje. Nezměněné jsou zkopírovány do první skryté vrstvy. Následují skryté vrstvy, z nichž poslední je napojena na výstupní vrstvu. Ta má obvykle méně neuronů než předešlé vrstvy a hodnoty na výstupu mohou představovat třídy, do kterých má být klasifikován vstup. S počtem jednotlivých vrstev souvisí pojem hloubka sítě, která je rovna počtu všech vrstev neuronové sítě od vstupní až po výstupní vrstvu. Pojmem hluboká neuronová síť se označuje taková síť, která má dvě nebo více skrytých vrstev. Vrstvy se ještě dále rozdělují na plně propojené, pooling, s přeskočením či na konvoluční vrstvy.



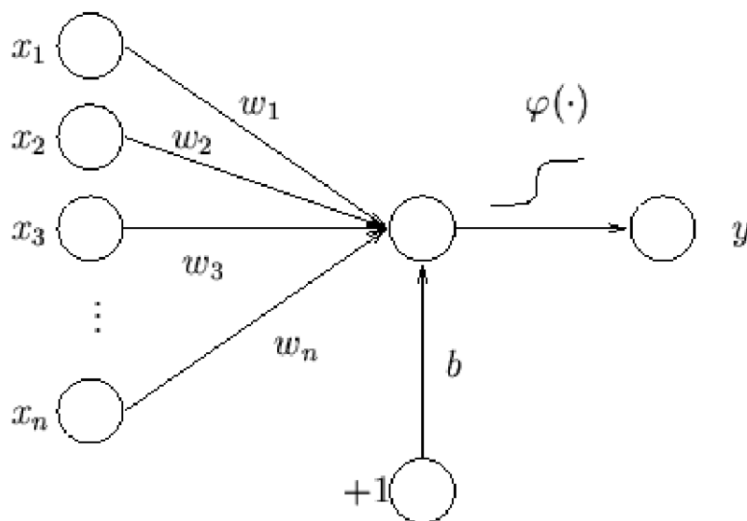
Obrázek 3.2: Schéma neuronové sítě, která má 2 skryté vrstvy

## 3.2 Umělý neuron

Základní stavební jednotka neuronových sítí je umělý neuron (*artificial neuron*) (viz obrázek 3.3). Tento model je založen na principu reálných neuronů, které se nacházejí v organizmech. Umělý neuron obsahuje libovolně mnoho vstupních propojení, přes které se mu předávají

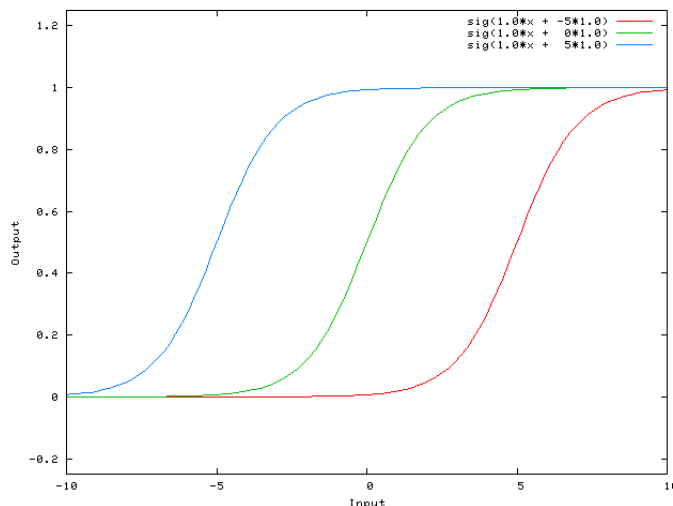
data v podobě vstupního vektoru  $\vec{x} = [x_1, x_2, \dots, x_n], x_n \in \mathbb{R}$ . Sám neuron obsahuje hodnotu bias  $b \in \mathbb{R}$  a vektor vah  $\vec{w} = [w_1, w_2, \dots, w_n], w_n \in \mathbb{R}$ , jenž je upravován během trénování neuronu.

Výstupní hodnota závisí na vstupních datech, aktuálním vnitřním stavu (hodnoty vah a biase) a na zvolené aktivační funkci. Vstupní hodnoty jsou váhovány, což znamená, že každá vstupní hodnota je vynásobena s váhou na daném vstupním spojení. S použitím definovaných vektorů lze napsat, že vstupní vektor je vynásoben s vektorem vah.



Obrázek 3.3: Schéma umělého neuronu

Hodnota bias  $b$ , která je přičtena k sumě násobků vah a vstupních hodnot, je prahová hodnota modifikující dobu, kdy se aktivuje neuron a změní svůj výstup. Matematicky to znamená, že s grafem aktivační funkce horizontálně pohybuje doleva nebo doprava v závislosti na tom, je-li hodnota biasu pozitivní nebo negativní. Toto posunutí je znázorněno na obrázku 3.4. V závislosti na řešeném problému může být žádoucí, aby i hodnota bias byla modifikována během učení společně s ostatními váhami. V opačném případě je hodnota nastavena pevně na nějakou konstantní hodnotu, obvykle na hodnotu jedna.



Obrázek 3.4: Vliv hodnoty bias na aktivační funkci

Výstup neuronu se vypočítá jako

$$y = f\left(\left(\sum_{k=1}^n w_k x_k\right) + b\right) \quad (3.1)$$

kde  $f$  je nějaká aktivační funkce,  $x_k \in \mathbb{R}$  je vstupní hodnota,  $w_k \in \mathbb{R}$  je váha, kterou se vstupní hodnota vynásobí a  $b \in \mathbb{R}$  je hodnota bias, která je přičtena k celkové sumě předtím, než je výsledek předán aktivační funkci.

Pokud neuron obsahuje skokovou aktivační funkci, tak se jedná o speciální typ neuronu, kterému se říká perceptron.

### 3.3 Aktivační funkce

Aktivační, neboli prahová funkce určuje výstupní hodnotu neuronu. Funkce se vybírá na základě problému, který se má neuronová síť naučit řešit. Správná volba prahové funkce vede k lepší konvergenci učení sítě. Naopak špatná volba může vést ke stále větší odchylce od správného řešení – může divergovat. Povaha problému může vyžadovat specifické vlastnosti aktivační funkce - lineární nebo nelineární. Pro nestandardní problémy je obvykle potřeba experimentálně zjistit, která funkce bude nejlépe vyhovovat danému problému.

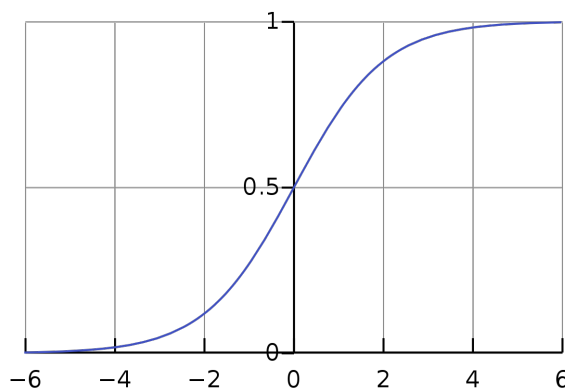
Pokud by veškeré aktivační funkce v modelu byly lineární, tak celkové mapování sítě by bylo omezeno pouze na lineární mapování vstupu na výstup. Reálné problémy ale lineární obvykle nejsou a v případě pokusu modelovat takovým modelem nelineární vztahy by vedlo k velice nepřesným výsledkům, který by byl zapříčiněn podučením (*underfitting*), což znamená, že model, který se učí zakódovat nějaký vzor v datasetu, je příliš jednoduchý. Proto je potřeba zavést do modelu i nelineární aktivační funkce, které tento problém řeší[3, p. 77–78].

Z pohledu učení je také důležité, aby aktivační funkce byla diferencovatelná. To umožňuje použití učících metod založených na výpočtu gradientu, jako je algoritmus backpropagation.

## Sigmoid

$$f(x) = \frac{1}{1 + \exp(-z)} \quad (3.2)$$

[[popsat]]

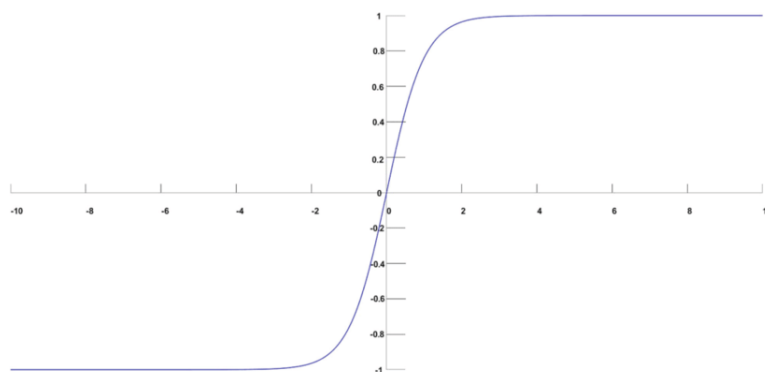


Obrázek 3.5: Graf aktivační funkce sigmoid

## Softmax

$$f(x) = \frac{1}{1 + \exp(-z)} \quad (3.3)$$

[[popsat]]



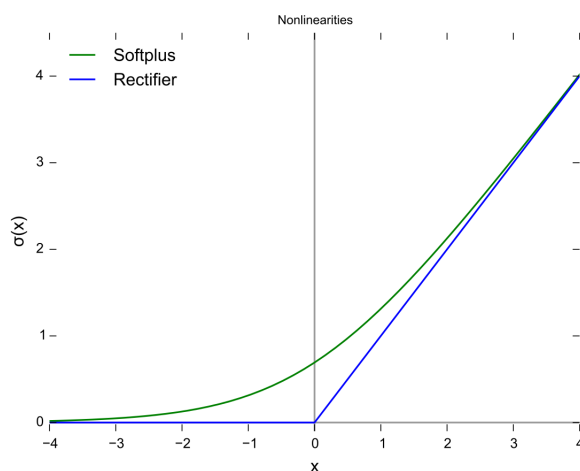
Obrázek 3.6: Graf aktivační funkce softmax

## ReLU

Rectified Linear Unit je nejčastěji používaná aktivační funkce. Vyžaduje-li neuronová síť nějakou nelinearitu, je ReLU pro většinu případů ideální. Pro každou zápornou hodnotu  $x$  vrací 0 a pro kladnou hodnotu  $x$  vrací tutěž hodnotu  $x$ , jak udává rovnice

$$f(x) = \max(0, x) \quad (3.4)$$



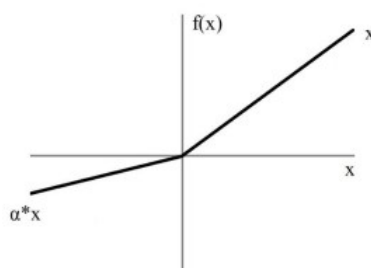


Obrázek 3.7: Graf aktivační funkce ReLU

## PReLU

Parametrizovaná ReLU je nelineární aktivační funkce, která se používá v případě, že chceme produkovat na výstup malý nenulový gradient i v případě záporné vstupní hodnoty  $x$ . V tom případě je vstupní hodnota vynásobena parametrem  $\alpha$  a to představuje výsledek. Parametr  $\alpha$  se společně s ostatními váhami učí během učícího procesu.

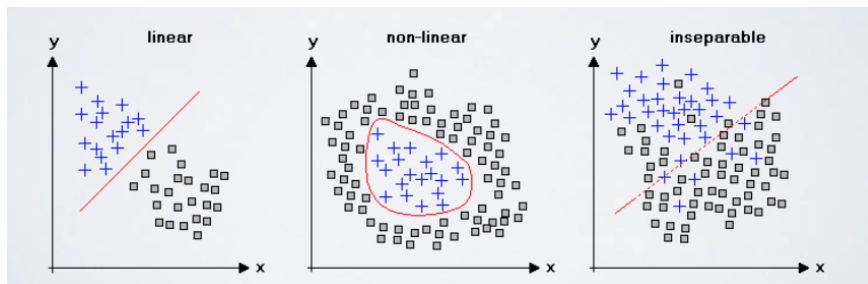
$$f(x) = \begin{cases} x & \text{if } x \geq 0 \\ \alpha x & \text{if } x < 0 \end{cases} \quad (3.5)$$



Obrázek 3.8: Graf aktivační funkce PReLU

## 3.4 Lineárně separovatelná data

Síť složená z jediného neuronu by byla omezena pouze na klasifikaci lineárně separovatelných dat. To jsou taková data, která lze rozdělit, existuje-li alespoň jedna přímka v rovině taková, že na jedné straně přímky jsou všechna data klasifikována do první třídy, a na druhé straně data klasifikována do třídy druhé, jak lze vidět na obrázku 3.9.



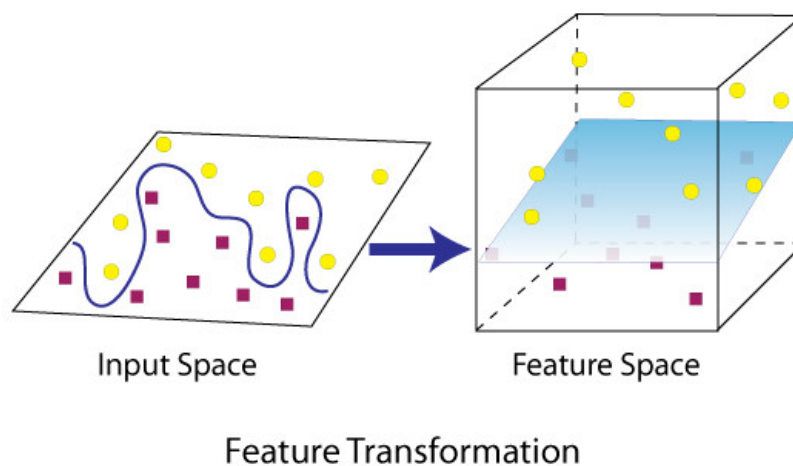
Obrázek 3.9: Lineárně separovatelná (vlevo) a neseparovatelná (vpravo) data

Mějme množinu dat, kterou tvoří dvě rozdílné třídy objektů  $\Omega_1, \Omega_2 \subset \mathbb{R}^m$ . Třídy jsou lineárně separovatelné, jestliže existuje  $w \in \mathbb{R}^m$  a  $\theta \in \mathbb{R}$  takové, že

$$\forall x \in \Omega_1 : x^T w > \theta \quad (3.6)$$

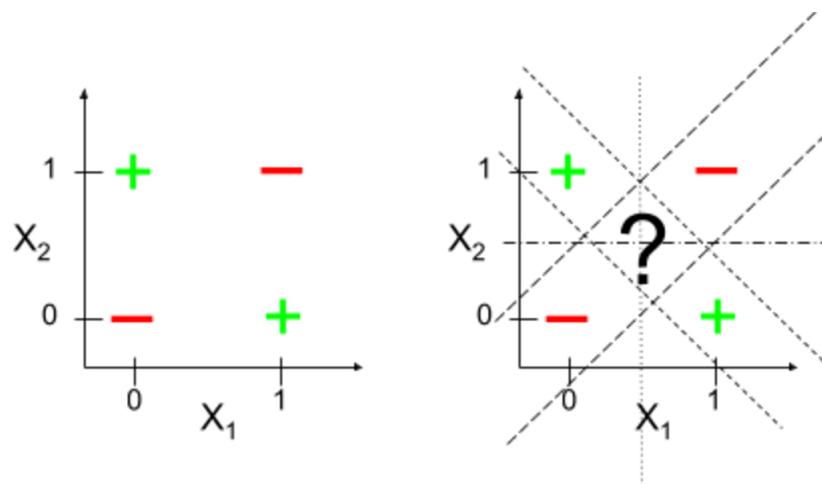
$$\forall x \in \Omega_2 : x^T w < \theta \quad (3.7)$$

Problém lze zobecnit tak, že řešením bude separace tříd pomocí hyper roviny v prostoru  $\mathbb{R}^m$ , jak zachycuje obrázek 3.10 [1].



Obrázek 3.10: Separace pomocí hyper roviny

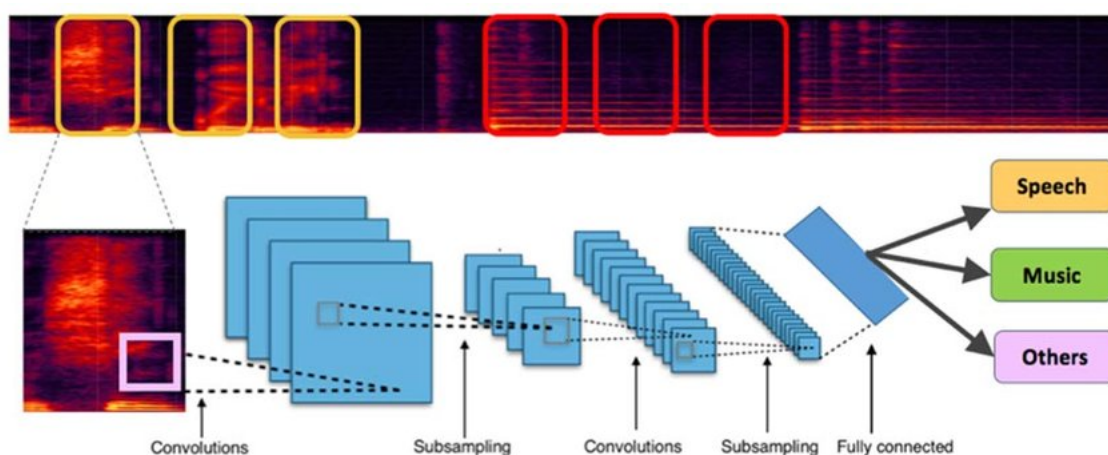
Neuron se nedokáže ale naučit klasifikovat třeba funkci XOR, protože ta není lineárně separovatelná. Byly by zapotřebí minimálně dvě přímky pro rozdělení dat, jak zobrazuje obrázek 3.11[[citovat xor]]. K překonání tohoto problému se neurony spojují do podoby neuronové sítě, která dokáže řešit libovolně komplexní problémy.



Obrázek 3.11: Graf funkce XOR, jejíž výstup není lineárně separovatelný

### 3.5 Konvoluční neuronové sítě

**[[Nebo to pojmenovat jako Konvolucni vrstva?]].** Konvoluční neuronové sítě jsou typ feedforward sítí, které neobsahují pouze plně propojené vrstvy, ale také vrstvy konvoluční.



Obrázek 3.12: Konvoluční neuronové sítě

#### 3.5.1 Konvoluční vrstva

Diskrétní konvoluce je definována jako

kde  $\star$  je konvoluční operátor,  $f$  je funkce signálu, funkce  $g$  je konvoluční jádro a  $f_i$  a  $g_i$  jsou hodnoty funkce na indexu  $i$ .

### 3.6 Učení neuronových sítí

Cílem trénování je naučit neuronovou síť vykonávat nějaký úkol. Trénování založené na metodě učení s učitelem vyžaduje dostatečně velký dataset obsahující dvojice vstup–výstup a obnáší volbu objektivní funkce, kterou se během trénování snažíme minimalizovat.

Formálně je cílem učení aproximovat nějakou funkci  $f^*$ . Síti je předána vstupní hodnota  $x$ , pro kterou síť definuje mapování na výstupní hodnotu jako  $y = f(x; \theta)$ , kde  $\theta$  je parametr, který se síť učí tak, aby dosáhla nejlepší aproximace funkce. [2, p. 163].

### 3.6.1 Objektivní funkce

[[cost funkce a popis, co to je, k čemu to je, proč to je]]

#### MSELoss

- vzoreček

#### Cross Entropy

- vzoreček

### 3.6.2 Backpropagation

Backpropagation (backprop) je algoritmus typu učení s učitelem pro učení neuronových sítí. Algoritmus zpětně prochází neuronovou sítí a na základě hodnoty cost funkce  $J(\theta)$  počítá její gradient metodou *gradient descent*. Cílem je upravit všechny váhy v síti na základě toho, jak moc každá váha přispívá k celkové chybě. Algoritmus backpropagation počítá řetězové pravidlo (*chain rule*).

Algoritmus se skládá ze 3 opakujících se kroků. Prvním krokem je forward fáze, ve které se počítá pro každou dvojici vstupu a referenčního výstupu

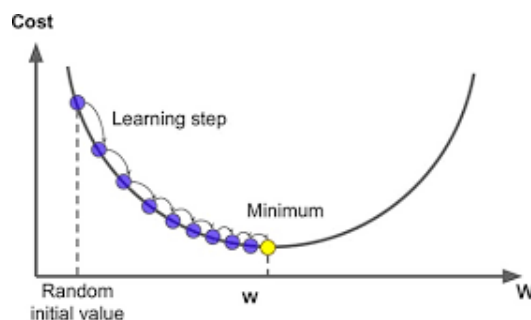
[?] [?, p 197]

- zpětné šíření chyby - adaptační algoritmus, podíl neuronu na chybě, - 3 opakující se fáze učení:

1) feedforward - dopředu 2) zpětné šíření chyby - Backpropagation 3) úprava vah a biasu na základě chyby pomocí gradient descent - chain rule

### Gradient descent

Gradient descent je iterativní optimalizační algoritmus pro hledání lokálního minima diferencovatelné objektivní funkce  $J(\theta)$ , kde  $\theta \in \mathbb{R}^d$  reprezentuje učené parametry modelu, úpravou těchto parametrů v opačném směru, než je hodnota gradientu. Výpočet je prováděn postupným pohybem ve směru největšího klesání, které je určeno zápornou hodnotou gradientu. Rychlost pohybu závisí na velikosti kroku, což udává *learning rate*. Správná volba *learning rate* ovlivňuje rychlost, jakou je nalezeno minimum funkce. Při nízké hodnotě bude výsledek přesnější, ale nalezení minima bude výpočetně náročnější, protože v každém kroku se počítá nová hodnota gradientu. Při větší hodnotě je riziko, že minimum bude přeskočeno[8].



Obrázek 3.13: Ukázka algoritmu gradient descent při hledání minima dané funkce postupným přibližováním dle nastavené velikosti kroku

[[popsat více vliv LR na loss]]

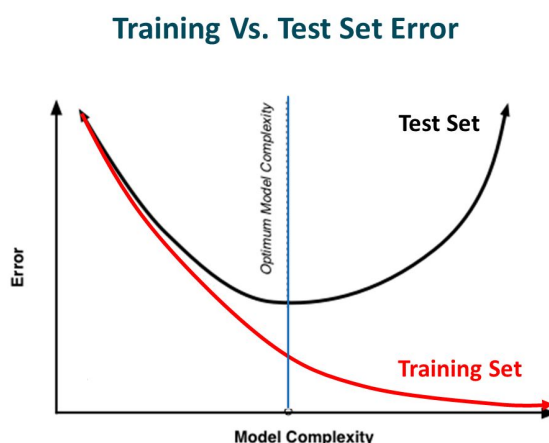
### 3.6.3 Underfitting a generalizace

Generalizace, neboli *test error*, je vlastnost modelu, která udává, jak dobře model pracuje s dříve neviděnými daty, tedy s daty, které nebyly použity pro jeho učení. Během učení vzniká učicí chyba, která je trénováním postupně minimalizována, ale kvalita modelu je pak měřena na testovacích datech. Testovací chyba by měla být stejná, jako *learning error* [[mam psat learning error nebo ucičí chyba???]]. Z rozdílu mezi těmito chybovými hodnotami se dají diagnostikovat problémy s naučením sítě – podučení (*underfitting*) a přeučení (*overfitting*).

*Underfitting* nastane v případě, že model se již nedokáže v důsledku jeho kapacity zmenšit dostatečně jeho chybovou hodnotu na trénovacím datasetu.

*Overfitting* nastane, když rozdíl mezi trénovací chybou a testovací chybou je příliš velký.

Kapacita je vlastnost modelu určující, kolik se toho dokáže model naučit. V případě nízké kapacity se model nedokáže naučit všechny příznaky z trénovacích dat a v případě vysoké kapacity se mohou přeučit zapamatováním si jejich příznaků a tím sice dosáhnou nízké trénovací chyby, ale špatných výsledků během testování[2, p107].



Obrázek 3.14: Vliv kapacity modelu trénovací a testovací chybu

### 3.6.4 Regularizace

Regularizace je souhrn postupů, které říkají, jak modifikovat učící algoritmus tak, aby se zredukovala chyba generalizace, zatímco testovací chyba zůstane stejná[2, p117].

### 3.6.5 Residuální spojení

[[**Toto chce obrazek residualního spojení. Popsat. Citovat odněkud, proč jsou residualní spojení dobrá pro trénování DeepNN a v jednom odstavci to nějak shrnout**]].

### 3.6.6 Význam validační množiny v trénování

[[**Mozná ne jako subsekcí ale jen jako jeden odstavec**]] Většina algoritmů strojového učení má nějakou sadu hyperparametrů, kterou je upravováno chování algoritmu. Hodnoty hyperparametrů obvykle bývají nastavovány ručně ještě před spuštěním procesu učení a hodnota se v průběhu nemění, protože hodnoty by bylo obtížné optimalizovat. Některá nastavení se nicméně mohou stát hyperparametrem a být upravována během trénování, ale není vhodné je měnit na základě výsledku učení na trénovací sadě, protože by mohlo dojít k přetrénování (overfitting) v důsledku **[[CEHO??]]**. Pro tento případ potřebujeme validační sadu, která je odlišná od trénovací sady. Po každém zpracování trénovací sady následuje validační sada, po jejímž skončení jsou optimalizovány hyperparametry. **[[[kniha 117-118] kap 5.3 = Hyperparameters and validation set]]** **[[najít jeste nejakej zdroj s popisem a pripadne nejaký zajímavější info.]]**  
[[today mini zaver teto cele kapitoly]]

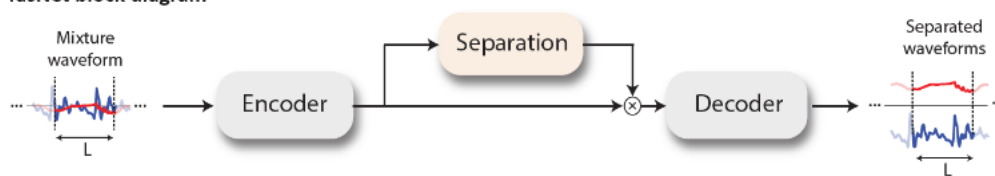
## Kapitola 4

# TasNet - Time-Domain Audio Separation Network

[[uvod do kapitoly]]

[[Architektura full – obrázek, bloky... Na začátku zmínit, že kapitola vychází ze studie ref studie. Kapitola NEBUDE moc dlouhá, cca 5-6 stran max.]]

A. TasNet block diagram

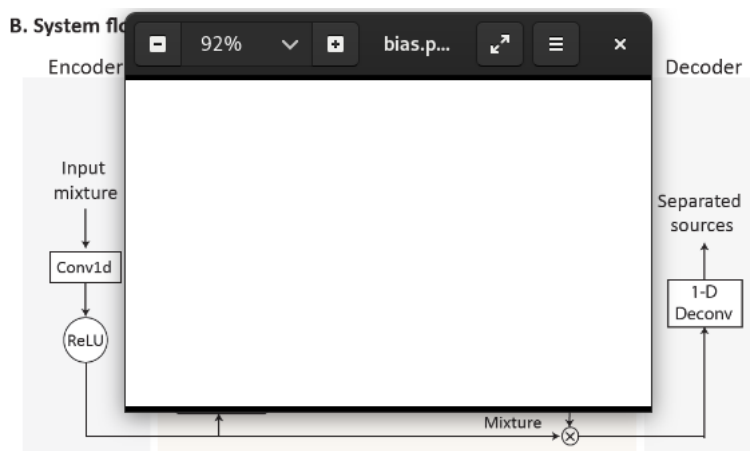


Obrázek 4.1: Zjednodušený model architektury

### 4.1 Konvoluční auto-ekodér

[[Konvoluční autoenkodér, vstup, výstup...]]

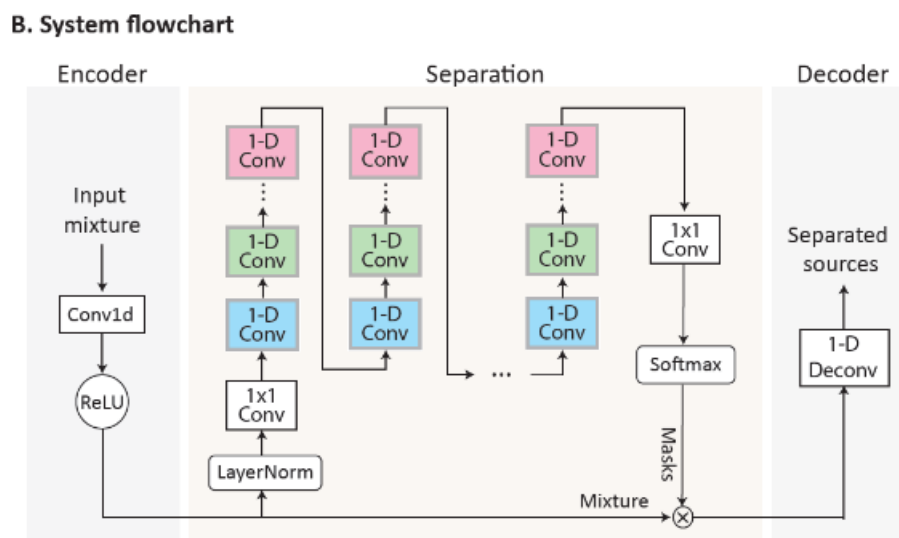
- schema bez separačního modulu - non negative representation of audio



Obrázek 4.2: Schéma konvolučního autoenkodéru

## 4.2 Separační modul

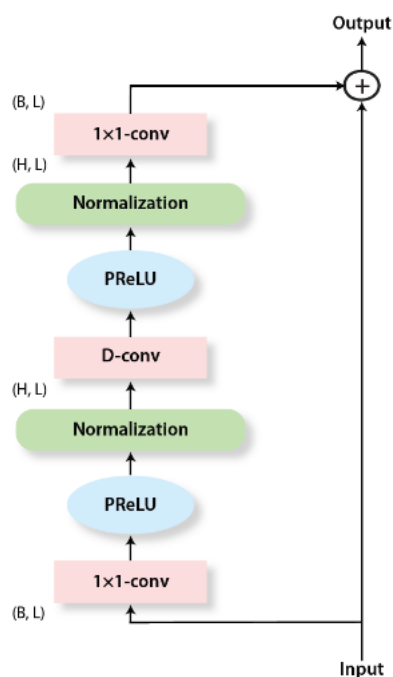
- odhad masek pro jednotlivé mluvčí - schema se separacním modulem



Obrázek 4.3: Schéma architektury TasNet

### 4.2.1 Konvoluční bloky

- Z čeho se skládá – konvoluční vrstvy, normalizace - diagram konv bloku. - Možná: Dilatace a time perception



Obrázek 4.4: Jeden konvoluční blok



[[zaver kapitoly]]

## Kapitola 5

# Implementace sítě

Pro implementaci neuronových sítí vzniklo mnoho frameworků jako pytorch, tensorflow, keras a další, které umožňují síť poskládat z již předdefinovaných modulů. Jejich vlastnosti lze modifikovat dle potřeby pomocí argumentů při jejich instanciaci nebo při jejich použití. Kromě těchto bloků a mnoha dalších možností frameworky obsahují i metody pro práci s daty, funkce pro vykreslování grafů a pro matematické operace.

Mým úkolem bylo implementovat neuronovou síť podle architektury TasNet pro separaci mluvčích v časové doméně. V rámci zadání jsem si zvolil použít framework s názvem pytorch<sup>1</sup>, který má kvalitní dokumentaci a aktivní uživatelskou základnu. Pytorch funguje nad jazykem Python<sup>2</sup> a v celé práci používám jeho aktuálně nejnovější verzi (Python 3.8.2) a mimo jiné podporuje práci s daty, včetně implementace vlastního nahrávání a manipulace s daty. Samotná implementace probíhala na systému Kali linux (4.19.0-kali5-amd64), který je založený na systému Debian. Aplikace byla vyvíjena v textovém editoru Vim<sup>3</sup> a byla během implementace spouštěna a testována v příkazovém procesoru Bash<sup>4</sup> (Bourne Again SHell) verze 5.0.16. Bash jsem zvolil i pro implementaci některých pomocných skriptů kvůli jednoduchému spouštění z terminálu a možnostem automatizace některých opakujících se příkazů. Pro správu verzí jsem používal verzovací systém Git<sup>5</sup>.

Základní jednotkou, se kterou síť pracuje, je tensor. Ve frameworku pytorch je **torch.Tensor** třída reprezentující multidimenzionální matici obsahující data jednoho typu. Na tensoru jsou převáděna data během načítání datasetu a následně předávána samotné síti ke zpracování.

**[[sem pridat to trenovani podle poznamky od vedouci, bude se to hodit i proto ze tu mam ten algoritmus trenovani.]]**

---

<sup>1</sup><https://pytorch.org/>

<sup>2</sup><https://www.python.org/>

<sup>3</sup><https://www.vim.org/>

<sup>4</sup><https://www.gnu.org/software/bash/>

<sup>5</sup><https://git-scm.com/>

```

    parsování argumentů programu;
    instanciaci modelu TasNet;
    if args.checkpoint! = NULL then
        | načtení checkpointu;
    end
    instanciaci třídy SegmentDataset a dataloaderu pro trénovací data;
    instanciaci třídy SegmentDataset a dataloaderu pro validační data;
    for epocha v args.Epochs do
        /* trénování */
1       loss = 0;
2       for směs_mluvčích z trénovacího datasetu do
            /* cyklus načítá z dataloaderu segmenty nahrávek, které jsou
               generovány v třídě SegmentDataset */
3           odhadovan_separace = TasNet(sms_mluvch);
4           loss = sisnr(optimln_separace, odhadovan_separace);
5           propagace chyby a úprava vah;
6       end
        /* validace */
7       vypnout gradient;
8       for směs_mluvčích z validačního datasetu do
9           odhadovan_separace = TasNet(sms_mluvch);
10          validacni_loss = sisnr(optimln_separace, odhadovan_separace);
11          if hodnota validační loss neklesla potřeží v řadě then
12              | learning_rate = learning_rate/2;
13          end
14      end
15      na konci epochy uložit checkpoint;
    end

```

**Algoritmus 1:** Zjednodušený algoritmus běhu programu pro trénování sítě

## 5.1 Implementace modelu

**[[Pozn: pytorch, scripty, tridy, moduly, parametry a volby spusteni]]** **[[ (nacistani dat a jejich transformace, cyklus, main.py, parametry, .py soubory) ]]** Model se skládá ze separační části, která se skládá z konvolučních bloků, dále z...

**[[miniuvod k implementaci modelu]]**

### Třídy SegmentDataset a AudioDataset

**[[zpracovani datasetu, jedna segmentuje, druha ne]]** Blizsi popis segmentace v podkapitole 5.2.

### Třída Tasnet

Třída Tasnet.py reprezentuje model neuronové sítě. Atributy třídy představují jednotlivé vrstvy neuronové sítě. Síť obsahuje konvoluční vrstvy, enkodér, dekodér a *temporal convolutionan network*, která je tvořena sekvencí konvolučních bloků, které jsou zmíněny dále. V této třídě jsou také inicializovány váhy algoritmem xavier **[[citovat]]**. Ve funkci *forward()*,

která je volána, když jsou instanci sítě předána data ke zpracování, jsou sekvenčně volány jednotlivé vrstvy. Během zpracování předaných dat je v této sekvenci vypočítána maska a aplikována na zpracovávanou směs. Tím vzniknou dvě separované nahrávky, které jsou předány na výstup. **[[zminat stride, padding, a možna obrazek co to je a možna trochu prepsat (ktere, ktera, ktere).]]**

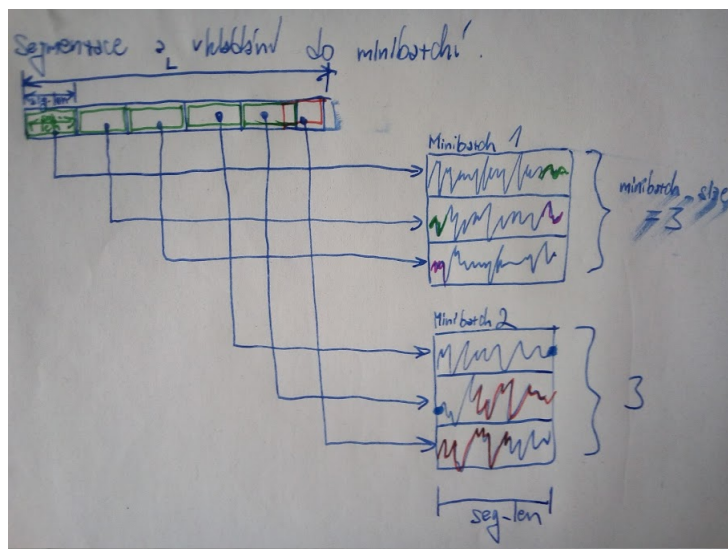
## Třída ResBlock

Třída ResBlock reprezentuje jeden konvoluční blok, který je opakován se zvyšující se dilatací v *temporal convolutional network* v separační části určené k odhadu masek. Podobně jako třída Tasnet, i tato třída obsahuje inicializaci a zřetězení vrstev. V inicializaci jsou nastaveny hodnoty pro konvoluční operace jako počet konvolučních jader, časová dilatace a počet kanálů. Tyto operace jsou ve funkci *forward* zřetězeny do sekvence konvolučních operací a aktivačních funkcí, podle vzorové architektury TasNet. Třída ResBlock je specifická svým residuálním spojením, kde dochází k rozdělení dat při zpracování, kdy jedna kopie je transformována operacemi v bloku a na konci sečtena s daty, které transformacemi neprošly a tento součet je výstupem bloku. Residuální spojení jsou podrobněji popsána v kapitole 3.

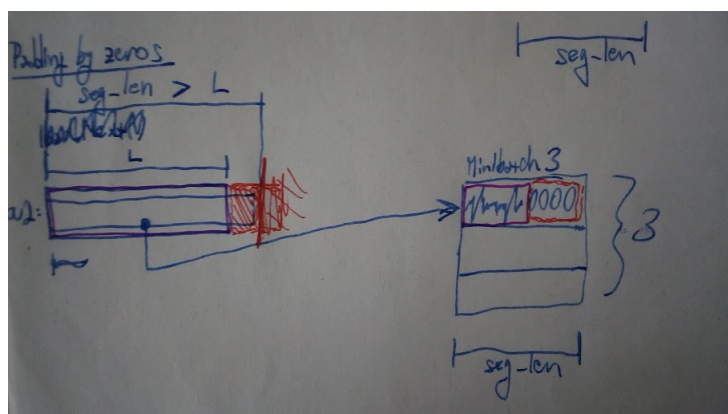
## 5.2 Segmentace nahrávek

Pro účely trénování sítě byly vstupní nahrávky rozdělovány na segmenty o délce 4, případně 2 sekund. Délku segmentů  $L_s$  lze nastavit přes argument `--segment_length` při volání skriptu *train.py*. Tato hodnota reprezentuje jeden z hyper-parametrů sítě. Výchozí hodnota je 4 sekundy. Při nižších hodnotách se prodlužuje délka trénování, protože se z jedné vstupní nahrávky o délce  $L$  vygeneruje až  $L/L_s$  segmentů a tím narůstá počet dat, který síť musí zpracovat.

Poslední segment by byl kratší, pokud by délka nahrávky byla kratší než délka segmentu. V takovém případě je segment doplněn nulami do délky ostatních segmentů jak lze vidět na obrázku 5.2. Toto doplnění je nutné kvůli dávkovému zpracování (**batches**), které neumožňuje rozdílnou délku elementů v dávce. Pokud je délka nahrávky delší než délka segmentu, tak se poslední segment vezme od konce bez nutnosti ho pak nulami doplňovat. Skládání segmentů do minibatche zobrazuje obrázek 5.1. Při vzorkovací frekvenci  $f_s = 8000\text{Hz}$ , obsahují segmenty  $f_s * L_s$  vzorků vstupní nahrávky.



Obrázek 5.1: Dělení nahrávky na segmenty o délce `segment_len` a jejich následné vkládání do minibatche. Lze si všimnout, že pokud poslední segment není dostatečně dlouhý, tak se vezme od konce nahrávky a jeho počáteční vzorky budou duplicitní s posledními vzorky předešlého segmentu



Obrázek 5.2: Je-li zpracovávána nahrávka, jejíž délka je kratší než je délka segmentu, je tento segment doplněn z prava nulami do požadované délky

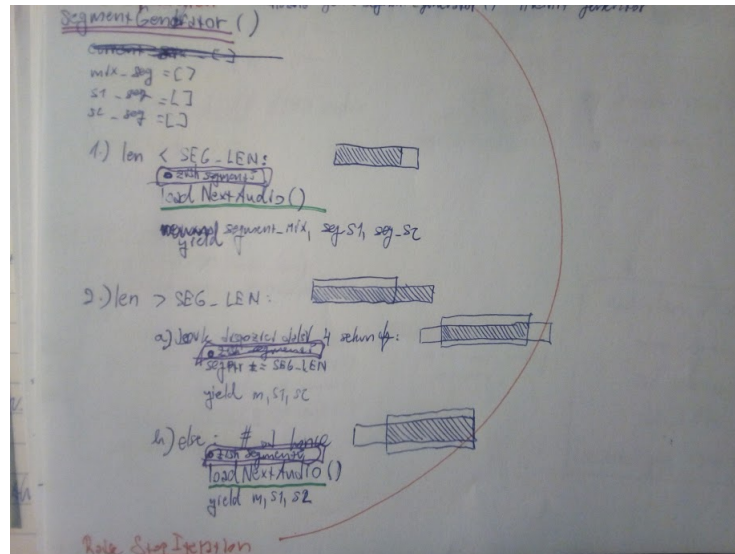
**[[možná tento odstavec ještě nějak upravit a doplnit obrázek kde je nati-  
tani asi 3 audio nahravek]]** Algoritmus 2 a obrázek 5.3 zobrazují proces segmentace. Funkce `segmentGenerator()`, která slouží jako generátor, pracuje současně se třemi nahrávkami – se směsí dvou mluvčích, s nahrávkou prvního mluvčího a s nahrávkou druhého mluvčího, které jsou do atributů `current_mixture`, `current_source1` a `current_source2` třídy `SegmentDataset` nahrávány funkcí `loadNextAudio()` poté, co z předchozí trojice nahrávek již nelze vygenerovat další segmenty.

```

Function segmentGenerator() : int[] is
    s1_segment = [];
    segptr = 0;
    while je dostupná další trojice nahrávek do
        if current_mixture_len < SEGMENT_LENGTH then
            /* aktuální nahrávka je kratší než délka segmentu */
1         s1_segment = current_source1[:];
2         yield s1_segment;
        else
            if segptr + SEGMENT_LENGTH < current_mixture_len then
                /* z aktuální nahrávky vzít segment */
3         s1_segment = current_source1[(segptr+SEGMENT_LENGTH)];
4         segptr += SEGMENT_LENGTH;
5         yield s1_segment;
            else
                /* nelze již načíst celý segment, takže se vezme od konce */
6         start_index = (current_mixture_len - SEGMENT_LENGTH);
7         s1_segment =
            current_source1[start_index:current_mixture_len];
8         yield s1_segment;
            end
        end
    end
end

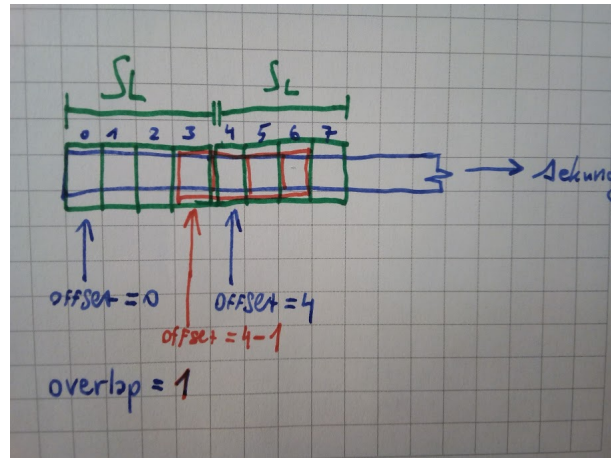
```

**Algoritmus 2:** Algoritmus segmenace nahrávek používá výhody python generátoru, který, narozdíl od běžných funkcí, po vrácení hodnoty příkazem `yield` neztrácí svůj vnitřní stav a při jeho dalším zavolání pokračuje tam, kde skončil. V uvedeném příkladě je znázorněna pouze segmentace nahrávky jednoho mluvčího, ale analogicky se segmentují i nahrávky směsi mluvčích a nahrávky druhého mluvčího



Obrázek 5.3: Způsob segmentace nahrávek

Kód segmentace také umožňuje určit délku, o kterou se segmenty budou překrývat. Těto funkcionality je docíleno zavedením proměnné **overlap**, která obsahuje délku překrytí a při segmentaci je od proměnné **offset**, který určuje začátek následujícího segmentu, tato hodnota odečtena, jak ukazuje obrázek 5.4. Při trénování je překrytí nulové, protože by to zvýšilo počet trénovacích dat jejich částečnou duplikací, což by negativně ovlivnilo délku trénování.

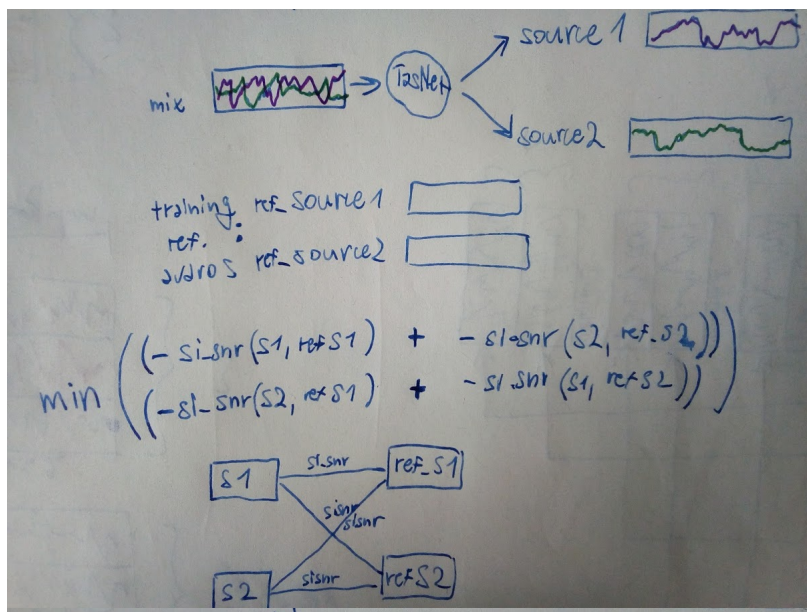


Obrázek 5.4: Segmentace nahrávek s nenulovým překrytím

### 5.3 Výpočet hodnoty loss za pomoci si-snr

[[popsat implementaci sisnr a vypocet loss za pomoci cross vypoctu, obrazek]]





Obrázek 5.5: Implementace a zjištění správné pozice korespondujících nahrávek pro výpočet hodnoty loss

## 5.4 Pomocné skripty

Během implementace se vyskytla potřeba zjednodušit si některé repetitivní úkony a jelikož upřednostňuji práci z terminálu, tak jsem si pro ně naprogramoval několik skriptů za použití jazyků python3 a bash.

Skript **gupload.sh** pro odesílání zdrojových souborů na *Google Drive*<sup>6</sup>, ze kterého *Google Colab*<sup>7</sup> načítá data a kód, který lze následně po částech spouštět z grafického rozhraní. K interakci s Google Drive jsem použil nástroj *rclone*<sup>8</sup>, který nabízí mnoho operací pro vzdálenou práci s Google Drive z terminálu. S pomocí *rclone* lze stahovat či nahrávat data, vzdáleně měnit adresářovou strukturu, mazat, vytvářet a upravovat soubory a podobně.

Skript **gdownload.sh** je určený pro snadné stahování dat z Google Drive. Stejně jako předchozí skript, používá operace poskytované nástrojem *rsync*. Při zavolání stáhne adresář obsahující data z posledního trénování, nebo adresář, jehož název je předán jako argument skriptu v příkazové řádce. Stažený adresář obsahuje checkpointy, soubory obsahující výstup trénování, soubor obsahující dvojici hodnot  $[zpracovno_nahrvek, loss]$ , které lze jednoduše vykreslit do grafu, a nakonec soubor s výsledky testování obsahující hodnoty metrik pro testovaný checkpoint.

Dále jsem vytvořil několik jednoduchých skriptů pro rychlé spuštění trénování a testování modelu s různými přednastavenými parametry. Bylo tak možno mít mezi skripty rozdílné cesty k adresářům pro účely trénování a ladění chyb nebo třeba nastavení času, po kterém se proces má sám ukončit. Pro spuštění trénování modelu tak lze zavolat skript **nntrain.sh**, pro testování skript **nnctest.sh** a pro inferenci (separaci mluvčích na předané nahrávce) skript **nninference.sh**.

<sup>6</sup><https://www.google.com/drive/>

<sup>7</sup><https://colab.research.google.com/notebooks/intro.ipynb>

<sup>8</sup><https://rclone.org/>



## 5.5 Trénovací prostředí

Trénování sítě probíhalo z počátku na CPU *intel i5* mého osobního notebooku *Lenovo Y50-70*, což se ale ukázalo jako naprosto nevhodné vzhledem k náročnosti výpočtu kvůli nedostatečnému výkonu. Začal jsem tedy používat službu *Google colab*, která poskytuje na omezenou dobu (12 hodin) stroje, které navíc obsahují technologii *cuda*, což mi umožnilo výpočet provádět na GPU, čímž se trénování zrychlilo několikanásobně. Ještě více pomohlo předplacení služby, čímž se mi prodloužila doba, po kterou jsem mohl síť trénovat, na 24 hodin a zvýšila se šance na získání silnějších strojů. Modely byly trénovány na GPU *T80* a *T100* v závislosti na tom, která z nich mi zrovna byla přidělena pro dané sezení (*session*).

Implementace samotného modelu ve frameworku *pytorch* nebyla náročná. Vzhledem k netriviálnímu úkolu jsem musel implementovat některé části, jako třeba segmentace nahrávek, sám. To přineslo mnoho potíží při odstraňování chyb v samotném modelu. Síť jsem prvně trénoval a ladil na CPU svého notebooku, takže probíhalo velmi pomalu (v řádu desítek hodin i při nízké velikosti modelu), a některé chyby se objevily až v průběhu trénování nebo dokonce na jeho konci. Situaci razantně zlepšil přechod na Google Colab, kde jsem síť trénoval na poskytovaných GPU a trénování tak zabralo **["pouze" (opravit uvozovky)]** několik hodin, takže se případná chyba objevila dříve, ale i přesto mi lazení sítě zabralo mnoho času. Pro další usnadnění samotné implementace jsem vytvořil několik skriptů, které mi usnadnily spouštění a lazení sítě.

## Kapitola 6

# Experimenty a vyhodnocení

- trenovani s ruznymi hyperparametry, uspesnost a tabulky s hyper parametry a dosazenymi vysledky a hodnotami sisnr, sdr atd. - model size comparison. - porovnaní s vysledky ze studie - obrazky separovanych mluvčích - signalu. - spektra - grafy trenovani loss a vysledkuu. - pametova narocnost modelu

Neuronové sítě představují mocný nástroj mnoha využití, který ale musí být naučen, jak danou činnost provádět. Trénování sítí je výpočetně náročný úkol, jehož náročnost se, krom dalších parametrů, může lišit v závislosti na velikosti datasetu a sítě. Obecně čím větší dataset a rozměr sítě, tím déle trénování trvá. Je proto podstatné, na jakém stroji je síť učena a od toho se odvíjí čas trénování.

Po dostatečném natrénování probíhá vyhodnocení sítě. Výsledky trénování jsou měřeny metrikami, které se volí na základě problému, který se síť učila řešit. Výsledky testování sítě jsou následně porovnávány v závislosti na velikosti daného modelu. Velikost modelu je ovlivněna hyper-parametry, které určují například počet vrstev sítě, velikost segmentace a další vlastnosti.

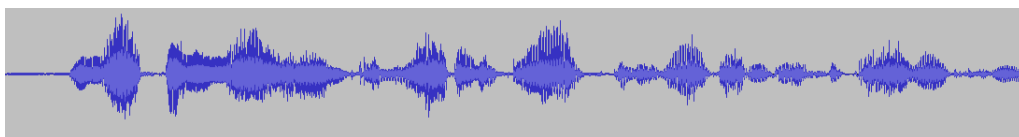
Cílem experimentů je zjistit, které z hyper-parametrů mají největší vliv na kvalitu separace a tedy na výsledky vyhodnocení, a jak moc lze model zmenšit, aby stále dával dostatečně dobré výsledky separace mluvčích. Tato kvalita je měřena primárně metrikami STOI, SI-SNR a PESQ, které byly popsány v kapitole 2. [\[\[ref na zdroje u metrik\]\]](#)

### 6.1 Dataset

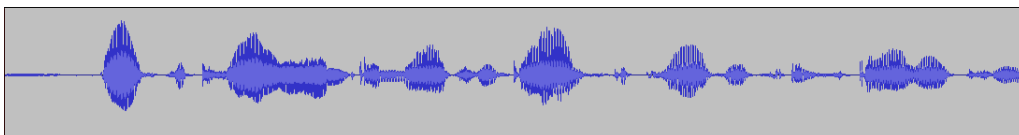
Dataset je množina dat, na kterých je síť učena a testována. Dataset se dá rozdělit na 3 podmnožiny, které jsou navzájem exkluzivní. Trénovací dataset je určen k trénování sítě a bývá největší. Validační dataset se prochází po zpracování trénovacího datasetu v průběhu trénování a slouží pro ověření, že hodnota loss na dosud neviděných datech stále klesá a nedochází tak k přetrénování. Při zpracování tohoto datasetu je síti zabráněno v učení. Poslední je testovací dataset, na kterém probíhá vyhodnocení sítě, tedy výpočet hodnot jednotlivých metrik.

Trénování a vyhodnocení modelu proběhlo na množině jednokanálových nahrávek směsí dvou mluvčích. Množina byla vygenerována náhodným výběrem různých mluvčích z Wall Street Journal (WSJ0) a vytvořením směsi mluvčích z těchto dvou náhodně vybraných nahrávek. Celková délka trénovacích dat je přes 10 hodin a přes 6 hodin validačních dat. Nahrávky jsou převzorkovány na 8kHz a během trénování zarovnány na zero means a jednot-

kovou varianci [[[studie str 5 Dataset]][49 - ze studie odkaz na script na generování a popis na netu]]].



Obrázek 6.1: Ukázka nahrávky směsi dvou mluvčích



Obrázek 6.2: První mluvčí ze směsi



Obrázek 6.3: Druhý mluvčí ze směsi

Lze si všimnout, že sečtením signálů separovaných mluvčích na obrázku 6.2 a 6.3 dostaneme přesně signál směsi, což lze vyjádřit vztahem

$$x[t] = \sum_{i=1}^C s_i[t] \quad (6.1)$$

, kde  $x[t] \in \mathbb{R}^{1 \times T}$  je diskrétní signál směsi a  $s_i[t] \in \mathbb{R}^{1 \times T}$ , kde  $i = 1, \dots, C$ , je jeden z  $C$  zdrojů[cite studie str3 vlevo].

[[doplnit info o zero means a jendotkove varianci]]

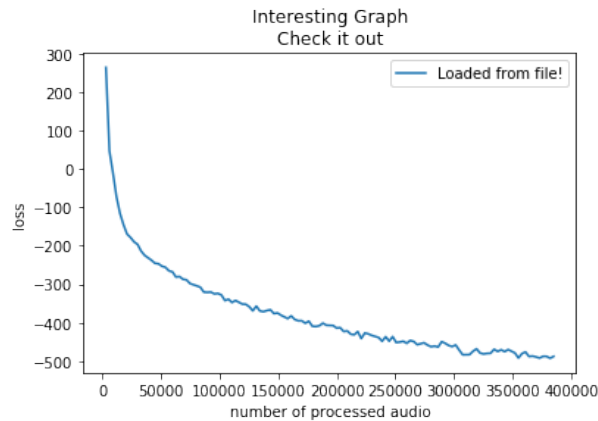
## 6.2 Průběh trénování

[[miniÚvod - konkrétní hyperparams, výkon a čas trénování, seg-len, bylo natrénováno několik modelů ... konkrétní hyperparams, výkon a čas trénování, seg-len]]. Bylo natrénováno několik, různě velkých modelů, které se lišily velikostí jejich hyper-parametrů, konkrétně počtem konvolučních bloků, počtem jejich opakování a velikostí segmentů zpracovávaných nahrávek z datasetu.

[[https://www.tablesgenerator.com/]].

Tabulka 6.1: Hodnoty hyperparametrů trénovaných sítí a jejich přesnost

	X	R	segment length sisnr
2	2	4	8.3
3	2	4	9.45
4	3	4	10.2
8	4	2	13.6



Obrázek 6.4: Příklad grafu loss hodnoty během učení



Obrázek 6.5: Hodnota loss při trénování modelů s různou velikostí hyperparametrů

### 6.3 Experimenty s modely

[[rozdelit tabulku podle HP a spis mensi tabulku ukazujici vysledku pro konkretni hyper parametry. experimenty: sisnr, reportovat ruzne metriky - stoi, pesq, prozkoumat hyper parametry a jak ktere jsou dulezite. Kdyz budu zmensovat sit tak jak se bude zhorsovat.]]

[[miniÚvod]] Pro vyhodnocení úspěšnosti a kvality modelů byly použity metriky si-snr vyjadřující poměr XXX, pesq říkájící YYY a stoi vyjadřující ZZZ [[mozna ref]].

### 6.3.1 Vliv hyper-parametru X

[[zmensovani a zvetsovani X, vysledky testovani, loss pripadne nejake fajnové grafy porovnani vysledky mezi jednotlivymi X, obrazky, grafy]]

### 6.3.2 Vliv hyper-parametru R

[[zmensovani a zvetsovani Y, vysledky testovani, loss pripadne nejake fajnové grafy porovnani vysledky mezi jednotlivymi Y, obrazky, grafy]]

### 6.3.3 Vliv hyper-parametru délka-segmentu

[[zmensovani a zvetsovani SL, vysledky testovani, loss pripadne nejake fajnové grafy porovnani vysledky mezi jednotlivymi SL, obrazky, grafy]]

### 6.3.4 Vyhodnocení a porovnání modelů

[[shrnutí a porovnani jednotlivých zmen mezi sebou komplexne. Pripadne porovnani vysledku s ref studií, obrazky, grafy]]

## 6.4 Možná rozšíření a navrhnutá vylepšení

- variabilnější dataset, mikrofony, šum a bordel prostředí - separace více mluvčích - hlučné prostředí - identifikace konkrétního řečníka - realtime separace

## Kapitola 7

# Závěr

**[[co jak dopadlo, vysledky a vyhodnoceni velikosti modelu a jaky byl nejlepsi]]**

Cílem práce bylo implementovat síť podle architektury TasNet pro separaci mluvčích v časové doméně a porovnat vliv velikosti sítě na kvalitu separace. Síť byla implementována za pomoci frameworku pytorch a jazyku python a natrénována na datasetu obsahujícím jednokanálové směsi dvou mluvčích. Trénování proběhlo na **[[X]]** modelech, které se od sebe lišily počtem opakujících se konvolučních bloků, velikostí časové dilatace a délkou vstupních segmentů směsí. Pro účel vyhodnocení modelů byla použita metrika si-snr, která udává poměr chtěného signálu ku šumu na pozadí, tedy obecně kvalitu separace.

Experimenty ukázaly, že během testování nejlépe dopadla síť, která měla 8 konvolučních bloků po 4 opakováních, s délkou vstupního segmentu  $L = 2$  sekundy. Tento model dosáhl po 100 epochách trénování hodnoty až **[[13,4]]** a tím se stal nejúspěšnějším modelem. Při fyzickém poslechu separovaných nahrávek bychom neslyšeli téměř žádný náznak druhého mluvčího. Oproti tomu, nejméně přesný model měl pouze 4 konvoluční bloky, 2 opakování a při délce segmentů  $L = 4$  sekundy dosahoval hodnoty SDR pouze **[[9.2]]**.

Zkoušel jsem separovat také nahrávky, které byly úplně mimo dataset, ale výsledek se nedá hodnotit jako úspěšný, jelikož hraje velkou roli prostředí, mikrofon, šum v pozadí a další vlivy, na které byla neuronová síť naučena. Tento problém by se dal překonat rozšířením trénovacího datasetu o větší škálu nahrávek mluvčích, které by byly pořízeny z různých zařízení v různě rušném prostředí.

**[[Doplnit ještě neco eh]]** Možna.

# Literatura

- [1] BAUCKHAGE, C. a CREMERS, O. Lecture Notes on Machine Learning: Linear Separability. Únor 2019.
- [2] IAN GOODFELLOW, A. C. *DEEP LEARNING*. MIT Press, 2017. ISBN 9780262035613.
- [3] KELLEHER, J. D. *DEEP LEARNING / John D. Kelleher*. MIT Press, 2019. ISBN 9780262537551.
- [4] KIM, J., EL-KHAMY, M. a LEE, J. End-to-End Multi-Task Denoising for joint SDR and PESQ Optimization. *CoRR*. 2019, abs/1901.09146. Dostupné z: <http://arxiv.org/abs/1901.09146>.
- [5] LUO, Y. a MESGARANI, N. *Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation*. 2018.
- [6] QIAN, W. C. C. X. e. a. Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers Inf Technol Electronic Eng* 19. Únor 2019, s. 40–63. Dostupné z: <https://doi-org.ezproxy.lib.vutbr.cz/10.1631/FITEE.1700814>.
- [7] RIX, B. J. G. H. M. P. a HEKSTRA, A. P. Perceptual evaluation of speech quality (pesq)-anew method for speech quality assessment of telephonenetworks and codecs. *Acoustics, Speech, and SignalProcessing*. 2001, sv. 2, s. 40–63. Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. Perceptual evaluation of speech quality (pesq)-anew method for speech quality assessment of telephonenetworks and codecs. In *Acoustics, Speech, and SignalProcessing*, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, volume 2, pp. 749–752. IEEE, 2001.
- [8] RUDER, S. *An overview of gradient descent optimization algorithms*. 2016.