TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií

Zlepšování řečových nahrávek pořízených v reálném prostředí

Diplomová práce

Studijní program: N2612 – Elektrotechnika a informatika
Studijní obor: 1802T007 – Informační technologie

Autor práce: **Bc. Tomáš Kounovský**
Vedoucí práce: Ing. Jiří Málek, Ph.D.

Liberec 2016

TECHNICAL UNIVERSITY OF LIBEREC
Faculty of Mechatronics, Informatics
and Interdisciplinary Studies

Enhancement of Real-world Speech Recordings

Diploma thesis

Study programme: N2612 – Electrical Engineering and Informatics
Study branch: 1802T007 – Information Technology

Author: **Bc. Tomáš Kounovský**
Supervisor: Ing. Jiří Málek, Ph.D.

Liberec 2016

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií
Akademický rok: 2015/2016

ZADÁNÍ DIPLOMOVÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Bc. Tomáš Kounovský**
Osobní číslo:	**M14000164**
Studijní program:	**N2612 Elektrotechnika a informatika**
Studijní obor:	**Informační technologie**
Název tématu:	**Zlepšování řečových nahrávek pořízených v reálném prostředí**
Zadávající katedra:	**Ústav informačních technologií a elektroniky**

Zásady pro vypracování:

1. Seznamte se s problematikou zlepšování jednokanálových řečových nahrávek (Speech Enhancement), zaměřte se na algoritmy používající odečítání spektra.
2. Seznamte se s problematikou trénování neuronových sítí, zaměřte se na sítě odstraňující šum (denoising autoencoders).
3. Natrénujte neuronovou síť umožňující odstranit z nahrávky řečového signálu některé vybrané druhy ruchů (například pouliční ruch) a vyhodnoťte úspěšnost odstranění pomocí objektivních kritérií.
4. Vytvořte aplikaci (případně na mobilním zařízení), která bude schopna nahrát a zlepšit řečovou nahrávku pořízenou v reálných podmínkách (případně v reálném čase).

Rozsah grafických prací:	**Dle potřeby dokumentace**
Rozsah pracovní zprávy:	**cca 40-50 stran**
Forma zpracování diplomové práce:	**tištěná/elektronická**

Seznam odborné literatury:

[1] **Boll, Steven F. "Suppression of acoustic noise in speech using spectral subtraction." Acoustics, Speech and Signal Processing, IEEE Transactions on 27.2 (1979): 113-120.**

[2] **Xu, Yong, et al. "An experimental study on speech enhancement based on deep neural networks." Signal Processing Letters, IEEE 21.1 (2014): 65-68.**

[3] **Xu, Yong, et al. "A regression approach to speech enhancement based on deep neural networks." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23.1 (2015): 7-19.**

[4] **Torch, Scientific computing for LuaJIT, [online 21.9.2015], http://torch.ch/.**

Vedoucí diplomové práce:	**Ing. Jiří Málek, Ph.D.** Ústav informačních technologií a elektroniky
Konzultant diplomové práce:	**Ing. Karel Paleček** Ústav informačních technologií a elektroniky
Datum zadání diplomové práce:	**14. září 2015**
Termín odevzdání diplomové práce:	**16. května 2016**

prof. Ing. Václav Kopecký, CSc.
děkan

Technická univerzita v Liberci - 14 -
L.S.

prof. Ing. Zdeněk Plíva, Ph.D.
vedoucí ústavu

V Liberci dne 14. září 2015

Prohlášení

Byl jsem seznámen s tím, že na mou diplomovou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

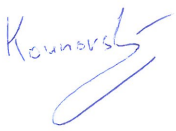
Beru na vědomí, že Technická univerzita v Liberci (TUL) nezasahuje do mých autorských práv užitím mé diplomové práce pro vnitřní potřebu TUL.

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

Diplomovou práci jsem vypracoval samostatně s použitím uvedené literatury a na základě konzultací s vedoucím mé diplomové práce a konzultantem.

Současně čestně prohlašuji, že tištěná verze práce se shoduje s elektronickou verzí, vloženou do IS STAG.

Datum: 16. 5. 2016

Podpis: 

Poděkování

Rád bych na tomto místě poděkoval svému vedoucímu, Ing. Jiřímu Málkovi, Ph.D., za vlídný přístup, trpělivost a odbornou pomoc při psaní této práce.

Abstrakt

Tato práce se zabývá problematikou zlepšování řečových nahrávek reálného charakteru odstraněním nežádoucího šumu v logaritmické výkonové spektrální oblasti pomocí hlubokých neuronových sítí. Velká databáze čistých řečových nahrávek je aditivně degradována několika druhy šumu na různých úrovních SNR a slouží jako trénovací sada pro dopřednou hlubokou neuronovou síť. Dvě sítě jsou natrénovány na odlišně sestavených trénovacích sadách tak, aby se naučily vztah mezi zašuměnou a čistou řečí. Jejich schopnosti odstranění šumu z nahrávek jsou otestovány ve známých i neznámých šumových podmínkách a porovnány s některými konvenčními metodami. Sítě jsou otestovány i v neznámých řečových podmínkách pomocí malé testovací databáze české řeči. Výsledky ukazují, že síť natrénovaná na databázi obsahující více druhů šumu je robustnější při zlepšování řeči degradované neznámým druhem šumu. Zároveň je tato síť schopna konkurovat konvenčním metodám při odstraňování šumu stacionárního charakteru a překonat je při odstraňování šumu silně nestacionárního.

Klíčová slova

Neuronové sítě, odstranění šumu, zpracování zvukového signálu, řečová nahrávka

Abstract

This thesis is focused on enhancement of real speech recordings by noise removal in the log-power spectrum using deep neural networks. Large database of clean speech is artificially corrupted by multiple noise types with multiple SNR levels and is used for training of a deep feedforward neural network. Two networks are trained in differently oriented training sets to learn the mapping from noisy to clean speech. The resulting network performance is then evaluated in matched and mismatched noise conditions and compared to some conventionally used methods. Both networks are also tested in mismatched speaker language conditions on a small czech speech database. Results show that a network trained on a database containing more noise types has more robust capabilities in speech enhancement by noise removal. Performance of this network is comparable with performance of conventionally used methods in stationary noise conditions and surpasses them when the noise is highly non-stationary.

Keywords

Neural networks, noise removal, audio signal processing, speech recording

Obsah

Seznam obrázků

Seznam grafů

Seznam tabulek

1 Úvod

1.1 Motivace a historie

V oblasti zpracování signálů je problematika zlepšování řečových nahrávek dlouho řešeným tématem. Aplikace vyžadující rychlé a spolehlivé odstranění šumu ze signálu jsou denně využívány velkou částí populace, například ve formě telefonních a radiových hovorů.

Takové aplikace často vyžadují zpracování jednokanálových signálů (nahrávání jedním mikrofonem). To má oproti vícekanálovým nahrávkám určité nevýhody - pokud jsou spektra řeči a šumu příliš podobná, pak je nelze dobře rozlišit [1]. Je tedy často nutné používat nějaké předpoklady (že šum zní chvíli na začátku nahrávky samostatně, že má nějaké rozložení pravděpodobnosti, jehož parametry je možné hledat apod.)

Odstranění šumu z jednokanálových nahrávek se obvykle řeší buď filtrováním v časové oblasti (např. pomocí lineárních časově invariantních filtrů), nebo ve spektrální (nebo časově-spektrální) oblasti, ve které má řeč určité charakteristické vlastnosti. K tomuto účelu byla vyvinuta řada algoritmů, z nichž většina se v určitých variantách používá dodnes.

Za jednu z nejdůležitějších metod se považuje metoda odečítání spektra [2], při níž se jednoduše odečítá od spektra zašuměné nahrávky spektrum odhadnutého šumu. Tato metoda je velice výpočetně nenáročná a vhodná pro potlačení šumu se silně stacionární povahou (např. zvuk rotoru helikoptéry). Pro nestacionární šum je však tato metoda nevhodná, neboť nedokáže dobře odhadnout dynamickou povahu šumu. Také tato metoda nebere v potaz spektrum řeči a pracuje pouze s odhadem spektra samotného šumu (které je v reálných podmínkách poměrně lehké získat, pokud je šum stacionární a nahrávka obsahuje neřečové segmenty), což vede ke značnému množství tzv. hudebních artefaktů - zdánlivě náhodně se vyskytujících krátkých tónů, které vznikají izolováním některých frekvenčních pásem v řečovém a dynamicky se měnícím šumovém spektru.

Problém hudebních artefaktů částečně řeší MMSE STSA [1] algoritmus [3]. Tento algoritmus vychází z Wienerova filtru a k odhadnutí spektra šumu a řeči využívá pravděpodobnostní modelování obou složek jako statisticky nezávislých Gaussových proměnných. Dosažené výsledky jsou poměrně dobré, nicméně tento algoritmus je výpočetně náročnější než jeho předchůdci.

Ani metoda odečítání spektra, ani MMSE STSA nedosahují dobrých výsledků při potlačování šumu v nahrávkách s úrovní SNR pod 0 dB [4].

[1] Minimum Mean Square Error Short-Time Spectral Amplitude - minimální čtvercová chyba pomocí krátkodobé spektrální amplitudy

Moderním zástupcem jednokanálových metod pro redukci šumu je OMLSA [5]. Jedná se o dva algoritmy, OM-LSA [2] pro odhad řeči a MCRA [3] pro odhad šumu. Optimální spektrální zesílení je získáno na základě váženého geometrického průměrování hypotetických zisků asociovaných s nejistotou přítomnosti řeči. Váhy těchto hypotetických zisků jsou vypočítány jako podmíněné pravděpodobnosti na základě čištěného signálu. Spektrum šumu je odhadováno pomocí průměrování předchozích umocněných spektrálních hodnot s vyhlazovacím parametrem, který je upravován pomocí pravděpodobnosti výskytu řeči. Tento algoritmus dosahuje vynikajících výsledků i pro nízká SNR, při přítomnosti dynamického šumu a s minimálním výskytem hudebních artefaktů. Nevýhodou je velmi vysoká výpočetní náročnost.

V poslední době se znovu vracejí do obliby umělé neuronové sítě (ANN - artificial neural network). Práce experimentující s využitím takových sítí k čištění řečových nahrávek mají značně rozdílné výsledky v závislosti na zvolených hyperparametrech sítě a použitých datech.

Obšírná studie [6] porovnávající různé přístupy k odstranění šumu v řečových nahrávkách poukazuje na výhody a nevýhody využití neuronových sítí k tomuto účelu (denoising autoencoders - autoenkodéry pro odstranění šumu). Sítě prakticky negenerují hudební artefakty a dosahují vysokých hodnot potlačení šumu (zlepšení klasifikace řeči pomocí jednoduchého klasifikátoru založeného na skrytých Markovových procesech až o 60 % po odstranění šumu z nahrávek). Jako nevýhodu sítí studie označuje silnou závislost na trénovacích datech a nutné předpoklady o stacionaritě signálu.

V současnosti se používají hlavně sítě s několika skrytými vrstvami (Deep Neural Networks - DNN), které mají oproti dříve používaným mělkým (Shallow Neural Network - SNN) sítím lepší schopnosti naučit se složité nelineární funkce při stejném počtu neuronů [7]. Jejich trénování je však komplikované a zájem o DNN obnovilo objevení DBN (Deep Belief Network) [8]. Jedná se o hluboké neuronové sítě, které jsou postaveny a předtrénovány pomocí skládání RBM (Restricted Boltzmann Machine) a algoritmu kontrastivní divergence. Těch využila nedávná studie [9] k optimální inicializaci vah robustního autoenkodéru pro odstranění šumu. Společně s následným trénováním pomocí klasické stochastické metody největšího spádu překována výsledná síť schopnosti konvenčních metod.

Další studie [10] poukazuje na skutečnost, že předtrénování nemusí být (alespoň pro úlohu autoenkodéru pro odstranění šumu) nutné. Typické problémy hlubokých neuronových sítí (nalezení pouze lokálního minima, zanikající gradient) lze vyřešit dostatečně velkou trénovací sadou (která snižuje šance uváznutí optimalizačního algoritmu sítě v lokálním minimu) a volbou vhodných hyperparametrů sítě (např. aktivační funkce).

[2] Optimally Modified Log-Spectral Amplitude - optimálně modifikovaná logaritmická spektrální amplituda
[3] Minima Controlled Recursive Averaging - rekurzivní průměrování na základě minima

Síť inicializovaná náhodně (pomocí inicializačního algoritmu) dosahuje srovnatelných výsledků se sítí předtrénovanou. Pokroky v oblasti výkonu počítačů také omezují dopad některých problémů, obvzláště s dlouhou dobou trénování větších sítí.

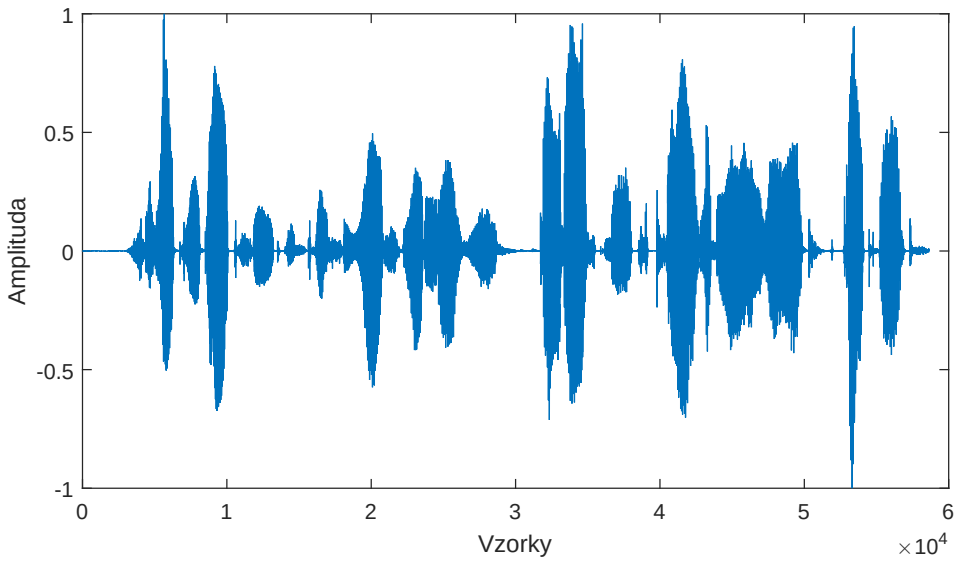
Výše uvedené poznatky jsou použity v této práci, jenž se snaží o vytvoření autoenkodéru pro odstranění šumu z řečových nahrávek pomocí hluboké neuronové sítě bez předtrénování. Výsledný autoenkodér by měl být schopen robustně odstraňovat z řečového signálu široké spektrum různých druhů šumu.

K porovnání robustnosti při odstraňování šumu jsou vytvořeny a porovnány dva autoenkodéry, první trénovaný pouze na jednom druhu šumu a druhý trénovaný na širší množině druhů šumu. Výsledky obou autoenkodérů jsou porovnány se standardními metodami popsanými dříve (odečítání spektra, MMSE STSA, OMLSA). Robustnější verze autoenkodéru je následně využita při implementaci aplikace, která umožňuje odstraňovat šum z řečových nahrávek.

1.2 Zpracování signálů a zlepšování řeči

1.2.1 Signál v časové oblasti

Signál je definován jako zaznamenané hodnoty změny nějaké fyzikální veličiny v čase. V oblasti zpracování řeči se nejčastěji jedná o hlasitost zvuku, kterou snímá jeden či více mikrofonů s danou frekvencí (tzv. vzorkovací frekvence, sampling rate - pak se jedná o tzv. diskrétní signál), viz graf 1.



Graf 1: Signál řeči v časové oblasti

Ze signálu v časové oblasti jsme schopni zjistit hodnoty hlasitosti (amplituda), času, a po jejich kombinaci rychlosti průchodů nulou (ZCR - Zero Crossing Rate), která zhruba odpovídá základní frekvenci tónu, resp. základní hlasivkové frekvenci řeči. Signál v časové oblasti lze zpracovávat a odstraňovat z něj šum [6], nicméně většina informačního obsahu řečového signálu se nachází ve frekvenční a časově-frekvenční oblasti.

1.2.2 Signál ve spektrální a časově-spektrální oblasti

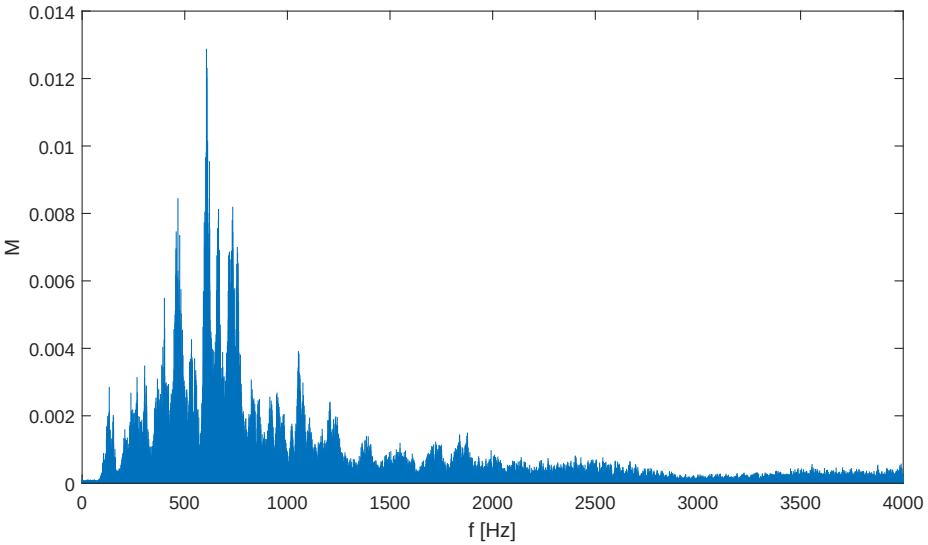
Signál v časové oblasti lze rozložit na množinu kosinových funkcí s různými frekvencemi, amplitudami a fázovými posuny. Takto rozložený signál se nazývá spektrum a lze ho získat pomocí diskrétní Fourierovy transformace (DFT) dle vzorce 1.

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-\imath 2\pi kn} \tag{1}$$

kde

$x[n]$ značí n-tý vzorek signálu
N je délka Fourierovy transformace v počtu vzorků
$k \in \{0, 1, \ldots, N-1\}$ je index frekvenčního koeficientu

DFT počítá magnitudu (míru aktivity) a fázi (posun) jednotlivých frekvenčních složek signálu. Magnitudové spektrum obvykle nese většinu užitečné informace řečového signálu a lze z něj dobře odhadnout informace o signálu, viz graf 2 - dominantní frekvence kolem 500 Hz a následné harmonické frekvence poukazují na lidskou řeč.



Graf 2: Magnitudové spektrum krátkého úseku řečového signálu.

Délka DTF N určuje, mimo jiné, také frekvenční rozlišení r_f (rovnice 2), tj. vzdálenost v Hz mezi jednotlivými frekvenčními složkami udávanými DFT.

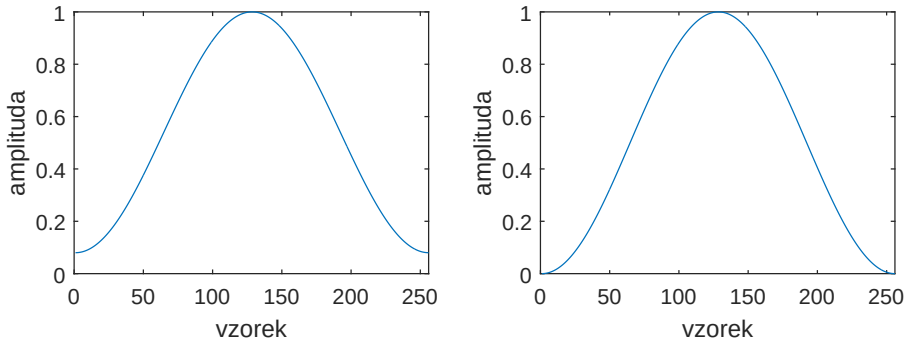
$$r_f = \frac{F_s}{N} \tag{2}$$

kde

F_s je vzorkovací frekvence signálu v Hz

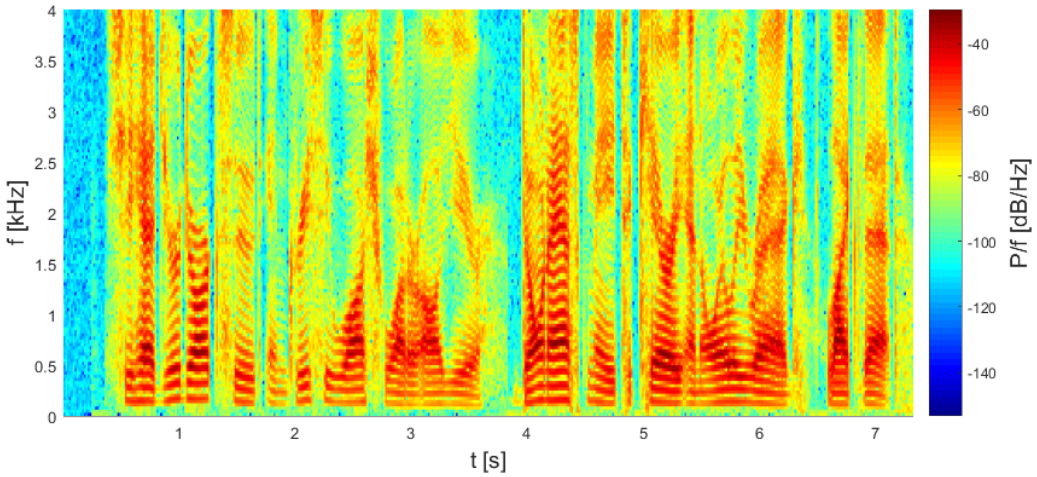
Čím lepší frekvenční rozlišení je třeba získat, tím větší musí být délka DFT. Tím se však snižuje časové rozlišení, tj. přesnost lokalizace nalezené spektrální aktivity v časové oblasti - pokud bude DFT aplikována na celou nahrávku, pak bude ve spektru možno rozlišit nejvíce frekvenčních aktivit, ale nebude možno určit kde v nahrávce se nacházejí. V praxi se tedy obvykle volí určitý kompromis - k získání informace o změně spektra v čase se implementuje tzv. krátkodobá Fourierova transformace (STFT, Short-Time Fourier Transform.)

Při výpočtu STFT se signál v časové oblasti rozdělí do rámců, na které je pak nezávisle aplikována DFT. Signál v jednotlivých rámcích je vynásoben vhodnou okénkovací funkcí - tj. symetrická funkce v oboru hodnot <0,1>, jenž nějak váží určité části signálu v rámci. Mezi nejznámější a v oblasti zpracování řeči nejpoužívanější patří Hammingovo a Hannovo okénko, viz obrázek 1.



Obrázek 1: Hammingovo (vlevo) a Hannovo (vpravo) okénko

Pro zajištění dobrého časového rozlišení a možnosti perfektní rekonstrukce signálu do časové oblasti se obvykle využívá překryv jednotlivých rámců, standardně mezi 25 a 75 % délky rámce dle využité okénkovací funkce. Spočtená magnitudová spektra jednotlivých rámců jsou pak naskládána za sebe, čímž vzniká tzv. spektrogram - graf udávající změnu spektra v čase (graf 3). Na ose y hovoříme o frekvenčních pásmech, na ose x o rámcích, jedno frekvenční pásmo v jednom rámci je tzv. frekvenční bin (koš).



Graf 3: Spektrogram řeči (délka rámce 256 vzorků, délka překryvu 128 vzorků)

Ze spektrogramu lze poměrně dobře identifikovat řeč a její prvky - souhlásky, samohlásky, formanty, sykavky, explozivy atd. Řeč je v čase tzv. kvazi-stacionární, tzn. že v dostatečně krátkém časovém úseku můžeme řeč považovat za stacionární signál. Tento úsek se obvykle udává kolem 30 ms a často se okolo něj staví délka rámce STFT (např. pro vzorkovací frekvenci 16 kHz bude ideální délka rámce 480 vzorků, pravděpodobně zaokrouhleno na nejbližší mocninu dvou, tedy 512).

V praxi se obvykle pracuje pouze s magnitudovou částí spektra [11], nejčastěji po umocnění na druhou (tzv. power, výkonové spektrum). Pro oblast zpracování řeči se často pracuje s logaritmickou verzí výkonového spektra, která poskytuje určité výhody - logaritmováním se částečně normalizuje vzdálenost jednotlivých hodnot, které mohou ve výkonovém spektru být vzdáleny o několik řádů. Lidské ucho také vnímá zvuk logaritmicky, použití logaritmického výkonového spektra se může systém přibližovat psychoakustickému modelu vnímání zvuku člověkem [12].

1.2.3 Zlepšování řeči

Obecný systém pro zlepšování řečové nahrávky ve spektrální oblasti obvykle počítá s tím, že pozorovaný signál je suma nekorelovaných signálů šumu a řeči dle rovnice 3.

$$y[n] = x[n] + s[n] \tag{3}$$

kde

$y[n]$ je n-tý vzorek pozorovaného signálu
$x[n]$ je n-tý vzorek řečového signálu
$s[n]$ je n-tý vzorek šumového signálu

Po výpočtu STFT lze rovnici 3 zapsat jako rovnici 4.

$$Y_k[m] = X_k[m] + S_k[m] \tag{4}$$

kde

$Y_k[m]$ je k-té frekvenční pásmo spektra signálu y v rámci m, podobně pro signály x a s

Vylepšené spektrum $\hat{X}_k[m]$ se získá dle rovnice 5

$$\hat{X}_k[m] = G_k[m] \cdot Y_k[m] \tag{5}$$

kde

$G_k[m]$ je odhadnutý spektrální zisk

Ideální spektrální zisk se dá získat ze spektra pozorovaného a šumového pomocí rovnice 6, v praxi však pro jednokanálové metody obvykle nemáme k dispozici skutečné šumové spektrum $S_k[m]$. Různé metody ho tedy aproximují ze známých údajů, obvykle získaných z neřečové oblasti nahrávky označené pomocí detektoru aktivity řeči (VAD - voice activity detector) [13].

$$G_k[m] = 1 - \frac{S_k[m]}{Y_k[m]} \tag{6}$$

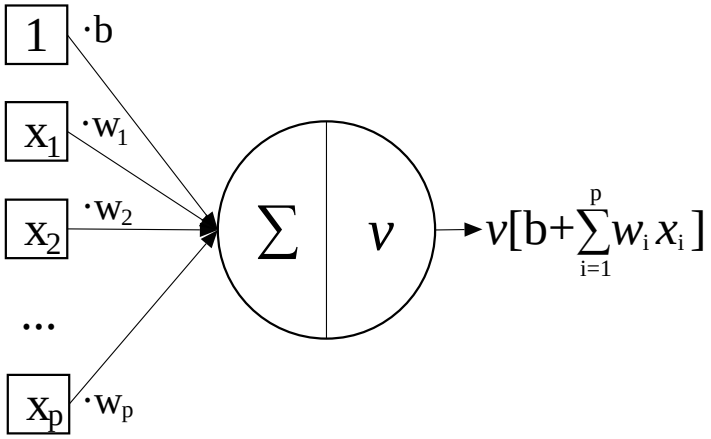
1.3 Umělé neuronové sítě

1.3.1 Historie

Vývoj neuronových sítí začal ve 40. letech 20. století, kdy byl vytvořen první výpočetní model neuronové sítě [14]. V roce 1958 byl vytvořen jednoduchý perceptron. Výzkum pokračoval až do roku 1969, kdy bylo ukázáno, že perceptrony nejsou schopny simulovat logickou funkci XOR, a také že tehdejší počítače neměly dostatečný výpočetní výkon k trénování velkých neuronových sítí.

Druhý „boom“ zažily neuronové sítě s objevením algoritmu zpětné propagace chyb (back-propagation of errors) [15], který umožnil rychle a efektivně trénovat vícevrstvé sítě, čímž byl vyřešen problém simulace XOR funkce. V 80. letech s příchodem paralelních výpočtů vzrostl výkon počítačů a s ním i zájem o neuronové sítě. Se stálým zvětšováním sítí však brzy ani to nestačilo a neuronové sítě vytlačily ze zájmu vědců výpočetně méně náročné metody.

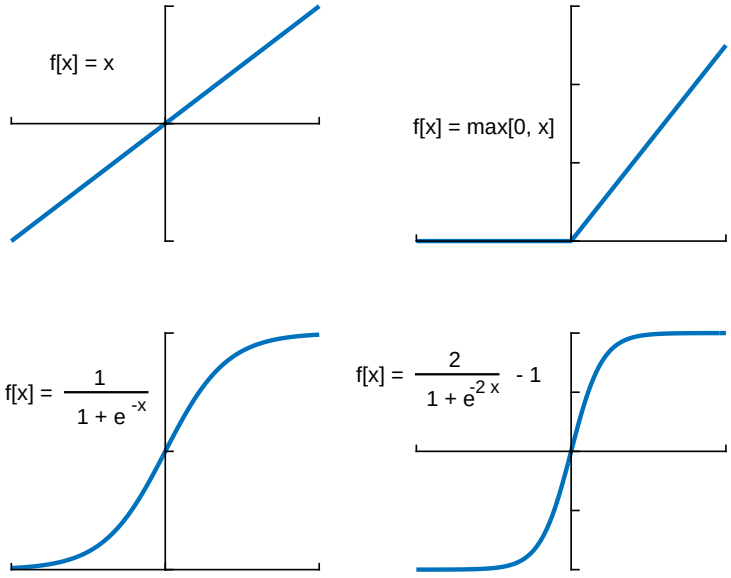
Poslední renesance neuronových sítí přišla v roce 2006 [16]. Kombinace objevu efektivity předtrénování sítě s vysokým paralelním výpočetním výkonem grafických karet (GPGPU) přinesla možnost trénovat složitější a větší sítě než kdy předtím, obvzláště v oboru hlubokého učení (deep learning). V posledních letech neuronové sítě dominují v soutěžích rozpoznávání vzorů (pattern recognition) a strojového učení (machine learning).



Obrázek 2: Vnitřní stavba umělého neuronu

1.3.2 Neuron

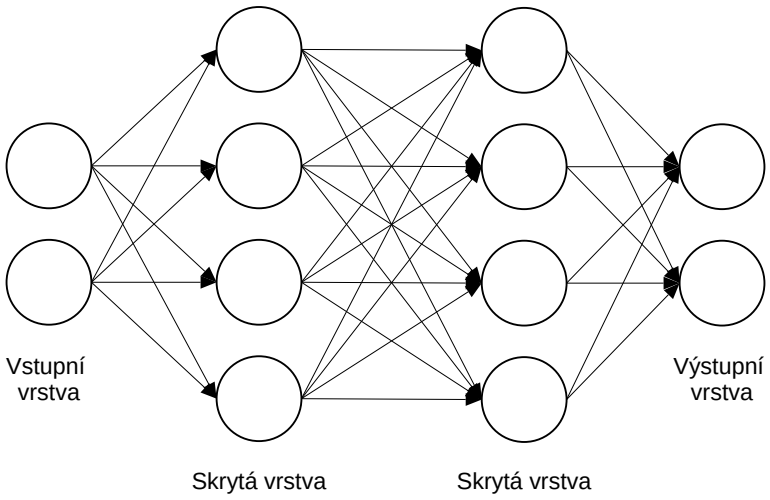
Na obrázku 2 je znázorněn základní stavební prvek umělé neuronové sítě, neuron. Lze ho rozdělit na tři logické celky. První část neuronu váží vstupní signály $x_1, x_2, \ldots, x_p$ váhami $w_1, w_2, \ldots, w_p$. Vstupní signály mohou přicházet přímo ze vstupu sítě, nebo od předchozích neuronů. Druhou částí je sumační funkce Σ, jenž sčítá vážené signály $w_i x_i$ společně s biasem b. Váhy a biasy neuronů jsou obvykle zpočátku nastaveny buď náhodně, nebo podle vhodného algoritmu (často používaná metoda je např. Nguyen-Widrow [17]), a v průběhu učení sítě se mění tak, aby výstup sítě byl optimální z hlediska kriteriální funkce. Třetí částí je úprava výsledné hodnoty z druhé části pomocí tzv. aktivační funkce. Často používané aktivační funkce jsou zobrazeny na obrázku 3. Pro trénování neuronové sítě pomocí algoritmu zpětné propagace chyb (popsán později) je nutné, aby aktivační funkce měla derivaci.



Obrázek 3: Často používané aktivační funkce. (Vlevo nahoře) lineární funkce, (vpravo nahoře) rektifikovaná lineární funkce, (vlevo dole) logistická funkce, (vpravo dole) hyperbolická tangenta

1.3.3 Topologie

První neuronové sítě měly pouze jednu vrstvu perceptronů (neuronů, jenž se chovají jako binární klasifikátory). Objev zpětné propagace chyb (error backpropagation) umožnil trénovat neuronové sítě s více vrstvami. Takové sítě obvykle obsahují vstupní vrstvu, 1 až K tzv. skrytých vrstev a výstupní vrstvu (viz obrázek 4).



Obrázek 4: Dopředná topologie neuronové sítě, v tomto konkrétním případě 2 vstupní signály, 2 skryté vrstvy po 4 neuronech, 2 neurony ve výstupní vrstvě

Sítě s jednou skrytou vrstvou se obvykle nazývají mělké (shallow), sítě s více skrytými vrstvami hluboké (deep). Teoreticky jsou mělké a hluboké sítě schopny řešit stejné problémy se stejnými výsledky, ukazuje se však, že hluboké sítě jsou se stejným počtem neuronů schopny modelovat složitější vztahy [7].

Dále se neuronové sítě dělí podle návrhu spojení mezi jednotlivými neurony. Často používané jsou tzv. dopředné (feedforward, dosl. „krmit dopředu“) neuronové sítě, ve kterých se signál propaguje výhradně směrem od vstupu na výstup. Tyto sítě se často používají pro úlohy regrese a hledání křivek (function fitting). Existuje velké množství různých architektur neuronových sítí, mezi další varianty se například řadí sítě cyklické nebo rekurentní.

1.3.4 Trénování

Nová neuronová síť má obvykle inicializovány váhy všech neuronů víceméně náhodně. Má-li síť sloužit k nějakému užitečnému účelu, musí být pro daný úkol nejprve natrénována.

Trénování neuronové sítě může probíhat buď s učitelem (supervised learning), kdy jsou síti předložena jak zdrojová, tak cílová data, nebo bez učitele (unsupervised learning), kdy jsou síti předložena pouze zdrojová data. V prvním případě je obvykle úkolem neuronové sítě naučit se vztah mezi zdrojovými a cílovými daty (regrese, klasifikace), ve druhé případě se obvykle jedná o úlohu shlukování či hledání skryté reprezentace dat. V této práci se jedná o úlohu regrese a tedy učení s učitelem.

K nalezení vztahu mezi vstupem a výstupem je nutno nejprve definovat kritérium, kterým bude síť hodnotit přesnost svého odhadu (performance/error function - výkonová/chybová funkce). Tímto kritériem bývá obvykle vzdálenost mezi reálným a očekávaným výstupem sítě. Pro regresní úlohy se nejčastěji používají různé varianty výpočtu chyby - odmocnina ze čtvercové chyby, suma absolutních chyb, průměrná absolutní chyba atd., v zásadě však může být použita jakákoliv funkce, která má derivaci.

Samotné trénování je úlohou numerické optimalizace - váhy a biasy neuronů jsou upravovány za účelem minimalizace chybového kritéria a vytvářejí multidimenzionální chybový prostor. Cílem trénování je pak najít takové nastavení vah a biasů, pro které se v kriteriální funkce v chybovém prostoru co nejvíce blíží globálnímu minimu.

Jedná se o iterativní proces, kdy v každé iteraci (nazývané epocha) síť na základě trénovacích dat shromažďuje údaje o změnách všech vah a biasů. Trénovací data mohou být zpracována buď po jednotlivých vzorcích (online trénování), nebo dávkově (mini-batch - dosl. malá dávka; offline trénování). V praxi se nejčastěji používá dávkové trénování s vhodně zvolenou velikostí dávky - čím větší dávka, tím rychlejší výpočty (pomocí knihoven (např. BLAS) implementujících rychlé maticové operace, oproti násobení matic a vektorů), ale tím více dat je nutno shromáždit pro stejné snížení chyby [18]. Zmenšuje-li se kriteriální funkce v průběhu trénování, pak říkáme, že síť konverguje.

K vypočtení ideální změny vah a biasů se nejčastěji využívá stochastické metody největšího spádu (stochastic gradient descent - SGD) v kombinaci se zpětnou propagací chyb [18]. Samotná změna váhy a biasu neuronu j je pak definována v rovnici 7.

$$(w_{ij}, b_j) = (w_{ij}, b_j) - \lambda \cdot \frac{\partial E}{\partial (w_{ij}, b_j)} \tag{7}$$

kde

w_{ij} je i-tá váha neuronu j
b_j je bias neuronu j
λ je vhodně zvolená délka gradientního kroku (learning rate, dosl. „rychlost učení“)
$\frac{\partial E}{\partial (w_{ij}, b_j)}$ je gradient chyby vůči váze w_{ij}, resp. biasu b_j

Gradient $\frac{\partial E}{\partial (w_{ij}, b_j)}$ značí, jak moc zvýšení dané váhy (nebo biasu) ovlivňuje zvýšení chybového kritéria. Tento gradient lze vypočítat pomocí řetězového pravidla a je definován různě podle toho, zda se neuron j nachází ve výstupní, nebo ve skryté vrstvě. Pokud je neuron j ve výstupní vrstvě, pak je třeba spočítat propagaci gradientu chyby pouze skrz samotný neuron dle rovnice 8.

$$\frac{\partial E}{\partial (w_{ij}, b_j)} = \frac{\partial E}{\partial v_j} \cdot \frac{\partial v_j}{\partial \Sigma_j} \cdot \frac{\partial \Sigma_j}{\partial (w_{ij}, b_j)} \tag{8}$$

kde

v_j je aktivační funkce neuronu j
Σ_j je sumační funkce neuronu j

Pokud se však neuron j nachází ve skryté vrstvě, pak mezi ním a chybovým kritériem stojí minimálně jedna další vrstva neuronů. Gradient $\frac{\partial E}{\partial v_j}$ tudíž nelze přímo spočítat, neboť výstupní signál v_j neovlivňuje přímo chybové kritérium E, ale pouze neurony v další vrstvě. Řešením tohoto problému je nejprve spočítat gradienty všech vah v následující vrstvě, které výstup v_j přímo ovliňuje. Sumu těchto gradientů lze následně substituovat za chybové kritérium E v rovnici 8. Tento postup lze rekurzivně aplikovat na jakýkoliv počet skrytých vrstev. Ve výsledku se tak chybové kritérium postupně propaguje nazpátek skrze síť, odkud pochází název tohoto algoritmu.

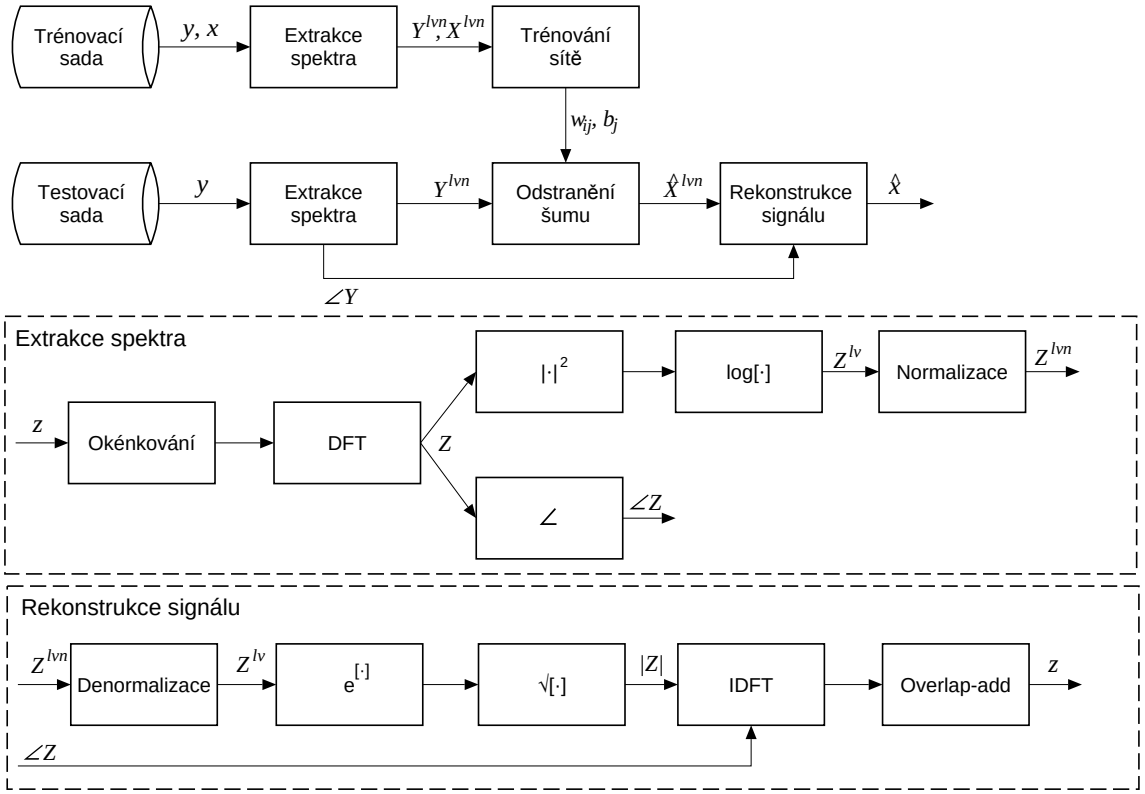
Konvergence sítě trénované algoritmem SGD silně závisí na parametru λ - příliš malá hodnota způsobí pomalou konvergenci, příliš velká hodnota naopak může konvergenci znemožnit. Další nevýhodou tohoto algoritmu je volený směr kroku v chybovém prostoru - čím větší je parametr λ, tím více má algoritmus tendenci vychylovat se z optimální cesty, což může vést ke zdlouhavé konvergenci. Existují však algoritmy, které dokážou vhodnou délku a směr kroku odhadnout samy [19].

2 Návrh systému a trénování sítě

Veškerá práce byla prováděna v prostředí Matlab R2015a.

2.1 Popis navrhovaného systému

Systém navržený v této práci je navržen podobně jako systém využitý v [9], viz obrázek 5. Zatímco blokově jsou oba systémy podobné, jednotlivé části se mezi sebou liší - trénovací a testovací sady jsou jiné a trénování sítě probíhá zásadně jiným způsobem.



Obrázek 5: Blokový diagram navrhovaného systému

Odstranění šumu ze signálu probíhá následovně: vstupní signál y je transformován pomocí krátkodobé Fourierovy transformace do logaritmického výkonového spektra Y^{lv} (fázové spektrum $\angle Y$ je uchováno pro pozdější rekonstrukci). To je normalizováno pomocí stejných hodnot jako trénovací data (Y^{lvn}) a předloženo síti jako sekvence rámců a jejich kontextu (viz kapitola 2.3.1). Výstupní spektrum $\hat{X}^{lvn}$ je denormalizováno, transformováno do magnitudové domény a spojeno s fázovým spektrem původního signálu. Výsledné spektrum $\hat{X}$ je pak rekonstruováno pomocí zpětné Fourierovy transformace a algoritmu overlap-add do časové oblasti ($\hat{x}$). Trénování sítě k účelu popsanému výše je rozvedeno v kapitole 2.3.2.

2.2 Popis a příprava dat

2.2.1 TIMIT

Jako zdroj čisté řeči slouží databáze anglické řeči TIMIT [20]. Tato databáze byla pod záštitou agentury DARPA vytvářena ústavy SRI International (SRI), Texas Instruments (TI) a Massachusets Institute of Technology (MIT) v roce 1993. Celkem se jedná o 630 řečníků různých pohlaví, věků a dialektů, z nichž každý čte 10 krátkých (2 - 3 sekundových) vět - 2 bohaté na dialekty, 5 foneticky kompaktních a 3 foneticky různorodé (z celkového korpusu 2342 různých vět). Všechny nahrávky obsahují textový a fonetický přepis, v této práci jsou však využita pouze zvuková data.

TIMIT celkem obsahuje 6300 nahrávek, z čehož 6256 (celkem 4 hodiny, 17 minut a 13 sekund řeči) bylo v této práci k dispozici. Korpus byl rozdělen v poměru 6000 (95.9 %) nahrávek pro trénování a 256 (4.1 %) nahrávek pro testování.

TIMIT je navzorkován vzorkovací frekvencí 16 kHz, pro tuto práci byly všechny nahrávky decimovány na 8 kHz. To slouží hlavně ke snížení potřebných výpočetních zdrojů, přičemž řečové signály by neměly být příliš poškozeny - většina energie řečového signálu je soustředěna do frekvenčního pásma cca 300 - 4000 Hz (horní hranice není pevně dána, 4000 Hz je často volený kompromis). Odstraněním komponent ležících mimo toto pásmo se ztrácí malá část sykavek a barvy hlasu, nicméně řeč zůstává stále srozumitelná.

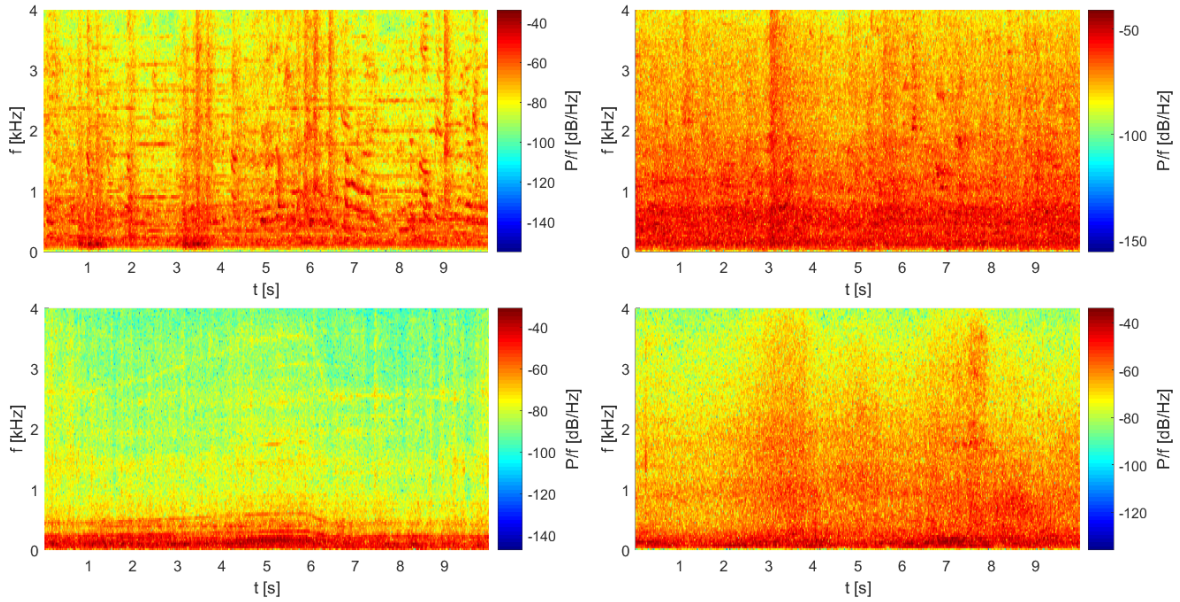
2.2.2 CHiME

Jako zdroj šumu slouží databáze šumu CHiME Challenge [21]. CHiME Challenge je jednou z několika soutěží v oblasti zpracování audio signálů, jazyka a řeči a strojového učení. Obvyklým úkolem pro soutěžící je například rozpoznávání klíčových slov v zašuměných nahrávkách pomocí separace a rozpoznání řeči. Databáze šumu, která je k těmto úkolům poskytnuta, je ideální pro tuto práci.

Celkem obsahuje 50 hodin šumových nahrávek, jenž jsou rovnoměrně rozloženy do 4 druhů různých šumů (kavárna, autobus, lidé a ulice), z nichž každý má jiné vlastnosti, viz obrázek 6.

Šum Kavárna (cafeteria) je silně dynamický a obsahuje hlavně promíchané hlasy lidí, hudbu a zvuky kolidujících talířů a sklenic. Hlasy lidí ruší primárně nízké frekvence, ve vyšších frekvencích převládá hudba a nádobí.

Šum Lidé (pedestrians) je podobný kafeterii, rušení je však více rozprostřené přes celé spektrum, dominují mu více promíchané hlasy pochodujících lidí a neobsahuje žádnou hudbu a náhlé kolize nádobí.



Obrázek 6: Spektrální charakteristiky použitých druhů šumu; Kavárna (vlevo nahoře), Lidé (vpravo nahoře), Autobus (vlevo dole) a Ulice (vpravo dole)

Šum Autobus (bus) je většinou stacionární a hlavní rušení nastává v nízkých frekvencích, kde převládá zvuku motoru, a v úzkých pásmech na frekvencích vyšších (harmonické frekvence motoru).

A nakonec šum Ulice (street), jenž obsahuje hlavně lehce dynamické zvuky projíždějících motorových vozidel s občasným klaxonem, skřípáním pohybujících se vrat a lehkých úderů do mikrofonu. Hlavní rušení nastává, podobně jako u autobusu, v nízkých frekvencích, rovnoměrně je však zarušeno celé spektrum.

Pro trénování specifické varianty autoenkodéru pro odstranění šumu byl zvolen šum kavárna, pro trénování obecné varianty byly zvoleny šumy kavárna, lidé a autobus. Pro účely testování v neznámých podmínkách byl ponechán šum ulice.

Všechny nahrávky šumů byly, stejně jako u databáze řeči, decimovány z původních 16 kHz na 8 kHz.

2.2.3 Výroba trénovacích a testovacích dat

V reálném světě se často setkáváme s nekorelovaným signálem řeči a šumu, což znamená že řeč a šum spolu nijak nesouvisí, nepocházejí ze stejného zdroje. Z tohoto důvodu je možno vytvořit zašuměný signál reálných charakteristik prostým sčítáním signálu řeči se signálem šumu vynásobeným koeficientem k k vytvoření požadovaného globálního SNR dle rovnice 9.

$$y[n] = x[n] + k \cdot s[n] \tag{9}$$

kde

k je koeficient šumu vypočítaný dle rovnice 10

$$k = 10^{\frac{SNR}{-20}} \cdot \sqrt{\frac{E[x]}{E[s]}} \tag{10}$$

kde

SNR je konstanta udávající požadovanou úroveň globálního SNR v dB

$E[x]$ a $E[s]$ je globální energie řečového signálu x, resp. šumového signálu s, spočtená dle rovnice 11

$$E[z] = \sum z[n]^2 \tag{11}$$

Pomocí těchto vzorců byly v kombinaci se šumem Kavárna aditivně degradovány všechny nahrávky trénovací řečové databáze na úrovně SNR = 0, 5 a 10 dB. Tím trénovací databáze dosáhla délky 12 hodin, 51 minut a 39 sekund zašuměné řeči. Výsledná databáze slouží k trénování specifické varianty autoenkodéru pro odstranění šumu.

Pro účely trénování obecné varianty autoenkodéru pro odstranění šumu byla databáze čisté řeči zduplikována, rozdělena na tři stejně velké části (aby byl každý druh šumu zastoupen stejnou mírou) a degradována pomocí šumů Kavárna, Lidé a Autobus s úrovněmi SNR = 0, 5 a 10 dB. Výsledná databáze obsahuje 25 hodin, 43 minut a 18 sekund zašuměné řeči.

Podobným způsobem byla degradována i testovací data. Jednotlivé čisté nahrávky však byly nejprve po dvojicích spojeny za sebe (zatímco řečník zůstal nezměněn) a na začátek jim bylo přidáno 0,25 sekund ticha. Důvodem je předpoklad některých konvenčních algoritmů, se kterými jsou srovnávány výsledky experimentů. Tyto algoritmy vyžadují na začátku zpracovávaného signálu neřečový segment k vytvoření dobrého odhadu charakteristik šumu. Delší trvání nahrávky také umožňuje těmto algoritmům lépe adaptovat odhad šumu a dosáhnout lepších výsledků [13].

Pro účely testování v neznámých šumových podmínkách byl ponechán šum Ulice. Taktéž byly z testovací sady vytvořeny degradované nahrávky s neznámými úrovněmi SNR = -3, 2 a 7 dB.

2.2.4 Logaritmická výkonová spektra

Na zašuměná a čistá data byla aplikována krátkodobá Fourierova transformace. Pro vážení rámců bylo zvoleno Hammingovo okénko o délce 256 vzorků, což odpovídá délce rámce 32 ms. Délka Fourierovy transformace byla zvolena stejně. Překryv byl zvolen 50% (128 vzorků), což umožňuje perfektní rekonstrukci signálu ze spektra pomocí metody overlap-add. Výsledná spektra jsou matice typu 129xM, kde 129 odpovídá počtu frekvenčních binů a M počtu rámců nahrávky dle rovnice 12.

$$M = \left\lfloor \frac{l[z] - l[w]}{Q} \right\rfloor + 1 \tag{12}$$

kde

$l[z]$, $l[w]$ je funkce udávající délku signálu z, resp. okénkové funkce w ve vzorcích
Q je délka překryvu rámců ve vzorcích

Z výsledného krátkodobého spektra byla extrahována magnituda, jelikož je práce zaměřena na potlačení šumu pomocí magnitudové filtrace (fázové spektrum bylo v této části zahozeno, neboť je potřeba až pro rekonstrukci testovaného signálu). K zajištění dobrého škálování byla magnitudová spektra převedena na logaritmická výkonová spektra dle vzorce 13.

$$Z_k^{lv}[m] = \log |Z_k[m]|^2 \tag{13}$$

2.2.5 Normalizace

K zajištění dobré konvergence při trénování bylo nutné trénovací data normalizovat. Byla použita standardní normalizace na nulovou střední hodnotu a jednotkový rozptyl dle rovnice 14

$$X_k^{lvn}[m] = \frac{X_k^{lv}[m] - \overline{X}_k^{lv}}{\sigma_k} \tag{14}$$

kde

$\overline{X}_k^{lv}$ je průměrná hodnota magnitudy k-tého frekvenčního pásma
σ_k je standardní odchylka magnitudy k-tého frekvenčního pásma

Všechny hodnoty $\overline{X}_k$ a σ_k jsou uchovány a použity pro normalizaci dat při testování.

2.2.6 Česká řečová databáze

Pro otestování robustnosti vůči neznámému jazyku řečníka byla autorem sestavena a nahrána malá databáze českých, foneticky bohatých vět. Jsou předpokládány horší výsledky po

odstranění šumu navrženými systémy, neboť čeština obsahuje unikátní, v angličtině se nevyskytující, foném (ř) a fonetickou stavbu vět. Schopnost odstraňovat šum i za přítomnosti neznámého jazyka může poukazovat na robustnost aplikace neuronových sítí pro odstranění šumu z řečových nahrávek.

Je vhodné uvést, že vyhodnocení experimentů na této databázi může být částečně zkresleno, neboť řečová databáze byla nahrávána v domácích podmínkách a samotné referenční (čisté) signály tak obsahují určité množství šumu. Tím pádem bude stejný šum bude očekáván i v nahrávce zpracované, ve které se již vyskytovat nemusí. Pro omezení zmíněných nechtěných efektů byly nahrávky před použitím filtrovány standardním hornopropustním filtrem (zádržné pásmo 0 - 100 Hz, propustné pásmo 100 - 4000 Hz, úbytek v zádržném pásmu -30 dB).

Tato autorem vytvořená databáze je k dispozici na přiloženém médiu.

2.3 Neuronová síť a trénování

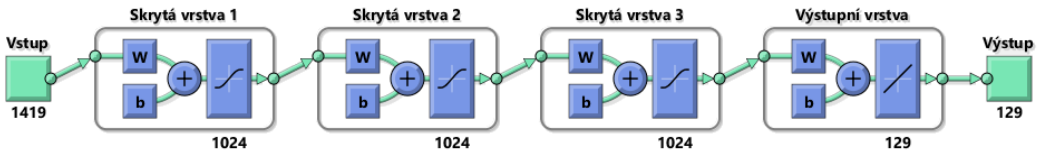
2.3.1 Volba hyperparametrů

Volba hyperparametrů sítě (topologie, počet vrstev, počty neuronů, aktivační funkce, trénovací algoritmy atd.) je velmi důležitou součástí návrhu neuronové sítě [18]. Klasická topologie pro úlohu autoenkodéru pro odstranění šumu je dopředná, která byla použita i zde.

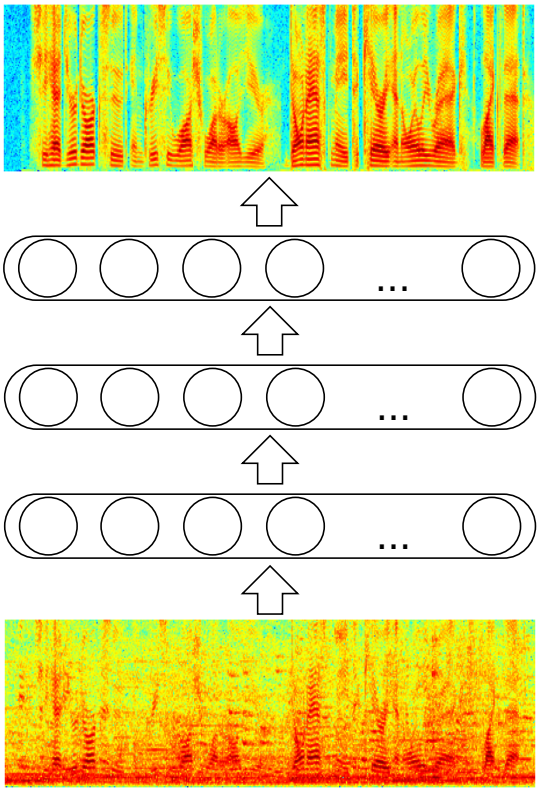
Ukazuje se [9], že výsledky autoenkodérů pro odstranění šumu je možné zlepšit, má-li síť vedle aktuálního příznakového vektoru také informace o vývoji příznaků v čase. Jednou z možností jak tyto informace síti poskytnout je rozšíření příznakového vektoru o příznaky vypočtené z okolních rámců, což označujeme jako přidání kontextové informace. Ke každému rámci vstupních dat byla přidána kontextová data v rozsahu +- 5 rámců. Rámce na začátku a na konci nahrávky byly doplněny nulami, výsledné matice byly transformovány do vektorové podoby (11 rámců, 129 příznaků na rámec, tedy 1419x1) a opět poskládány za sebe. Velikost vstupu a výstupu sítě je tedy dána jako 1419 vstupních signálů a 129 výstupních neuronů.

Hloubka sítě byla zvolena jako 3 skryté vrstvy. Počet neuronů ve skrytých vrstvách byl zvolen 1024, což dává dohromady 3 685 505 měnitelných parametrů (vah a biasů). Váhy a biasy sítě byly inicializovány algoritmem Nguyen-Widrow [17]

Aktivační funkce ve skrytých vrstvách byly zvoleny jako hyperbolická tangenta (tanh). To je hlavně z toho důvodu, že tanh pracuje v oboru hodnot (-1,1). To zhruba odpovídá logaritmickému výkonovému spektru normalizovanému na nulovou střední hodnotu a jednotkovou varianci, které bude síti předkládáno. Aktivační funkce ve výstupní vrstvě je lineární. Blokový diagram výsledné sítě je k vidění na obrázku 7.



Obrázek 7: Blokový diagram navržené sítě



Obrázek 8: Neuronová síť při procesu trénování, vstupní spektrogram zašuměného a cílový spektrogram neporušeného řečového signálu

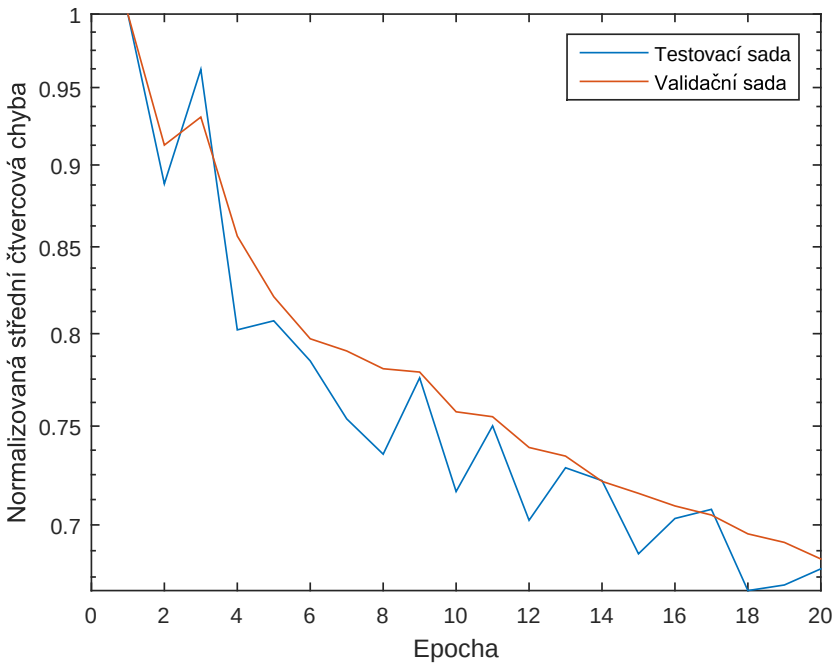
2.3.2 Trénování

Proces trénování je znázorněn na obrázku 8. Trénovací algoritmus byl zvolen jako SCG (Scaled Conjugate Gradient), neboť poskytuje dobrý kompromis mezi rychlostí trénování a paměťovými požadavky (např. defaultní algoritmus Levenberg-Marquardt nebylo možno kvůli velikosti navržené sítě vůbec použít, nezávisle na velikosti trénovací mini-dávky nebo parametru k redukci paměťových nároků). Parametry *lambda* a *sigma* byly ponechány defaultní, tedy $lambda = 5{\times}10^{-7}$ a $sigma = 5{\times}10^{-5}$.

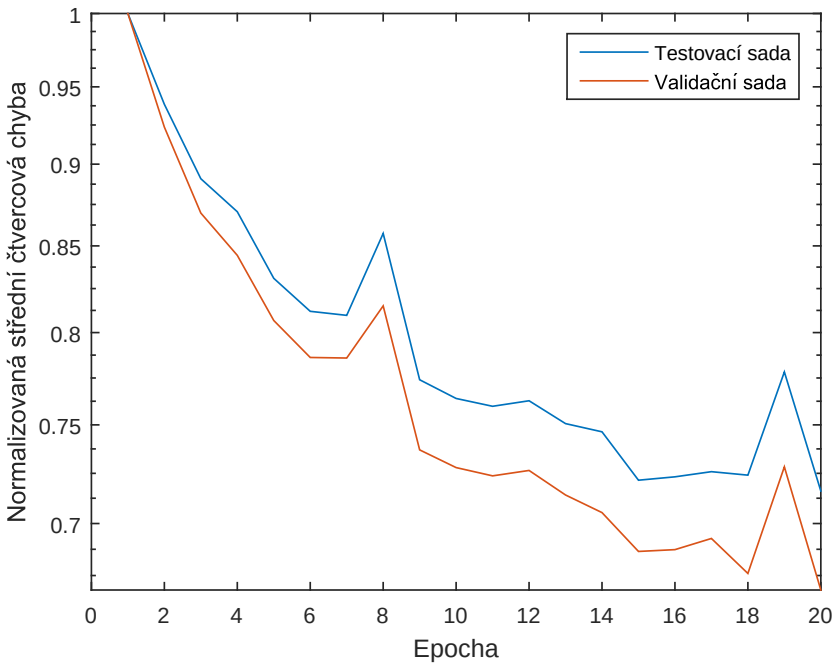
Matlab sám o sobě neimplementuje trénování neuronových sítí pomocí dávkové metody nad velkými daty (36 GB v paměti pro specifický autoenkodér a 72 GB pro obecný autoenkodér), proto bylo nutné takový tréning simulovat. Trénovací data byla uměle rozdělena na malé části a funkce *train* byla volána v cyklu. Síť byla trénována pomocí knihovny Cuda na grafické kartě, nicméně velikost paměti použité grafické karty (Nvidia GeForce GTX 550 Ti, 1 GB VRAM) omezovala maximální použitelnou velikost mini-dávky, která byla nakonec zvolena jako 500 rámců (což zhruba odpovídá 2,5 nahrávkám). Mimo to bylo kvůli omezené velikosti paměti RAM (8 GB) nutno často přesouvat velká množství dat mezi pamětí a pevným diskem. Všechna tato fakta negativně ovlivňují dobu trénování, která se pohybovala kolem 4 hodin na epochu u specifického autoenkodéru a 7,4 hodin na epochu u obecného autoenkodéru.

Při trénování neuronových sítí v mini-dávkách je zvykem trénovací data náhodně permutovat. To zamezuje tomu, že se síť bude učit nežádoucím způsobem, např. pokud bude síť dostávat nahrávky stále od jednoho řečníka, tak se naučí efektivně rozpoznat a zachovat právě jeho hlas, tedy bude konvergovat k nevhodnému lokálnímu minimu (může se stát částečně závislou na řečníkovi.) Proto byla trénovací data pro každou epochu náhodně permutována v rámci jednotlivých nahrávek.

Obě sítě byly trénovány pomocí kritéria MSE (mean square error - střední čtvercová chyba) po předem neurčenou dobu. Grafy 4 a 5 ukazují kritérium sítě v závislosti na čase (počtu epoch). Trénování je obvykle zastaveno na základě poklesu tohoto kritéria - tendence validační křivky bude po většinu času klesající, nicméně tendence testovací křivky nikoliv. Pokud se chyba na testovací sadě zvětšuje, zatímco chyba na validační sadě zmenšuje, dochází k tzv. přeučení (overfitting) a zhoršování schopností sítě v neznámých podmínkách. S dostatečně velkou a rozmanitou trénovací sadou však problém přeučení není tak významný a trénování bylo zastaveno z důvodů časových po 20. epoše pro obě sítě.



Graf 4: Vývoj optimalizačního kritéria v průběhu trénování specifického autoenkodéru



Graf 5: Vývoj optimalizačního kritéria v průběhu trénování obecného autoenkodéru

3 Experimenty a výsledky

3.1 Použité metriky

3.1.1 PESQ

PESQ (Perceptual Evaluation of Speech Quality - vjemová evaluace řečové kvality) je standard v oblasti telekomunikace pro automatické vyhodnocení kvality řečového signálu [22]. PESQ funguje na základě porovnávání referenční (čisté) a zašuměné nahrávky. Algoritmus nejprve provede časové zarovnání a následně porovnává nahrávky vzorek po vzorku.

PESQ je automatickou aproximací standardu MOS (Mean Opinion Score - Střední posudkové skóre), ve kterém lidé manuálně hodnotí kvalitu řeči na stupnici od 1 (nejhorší) do 5 (nejlepší). Výsledky PESQ testu jsou stejného charakteru na stupnici od 1 do 4,5. To je dáno tím, že PESQ využívá statistik, ze kterých vyplývá, že účastníci MOS testu málokdy hodnotí nejlepší možnou známkou (i při naprosto čisté nahrávce) a nejlepší průměrně očekávatelné skóre je právě 4,5.

3.1.2 LSD

LSD (log-spectral distance, log-spectral distortion) je objektivní míra odlišnosti signálů udávající logaritmickou vzdálenost ve výkonovém spektru [23]. Výsledek v dB je vypočítán dle rovnice 15. Jedná se o vzdálenost mezi pozorovaným a ideálním spektrem, tedy čím menší výsledek, tím lepší.

$$LSD = \frac{1}{M} \cdot \sum_{m=0}^{M-1} \left\{ \frac{1}{\frac{N}{2}+1} \cdot \sum_{k=0}^{\frac{N}{2}} \left[10 \cdot \log \frac{|X_k[m]|}{|\hat{X}_k[m]|} \right]^2 \right\} \tag{15}$$

kde

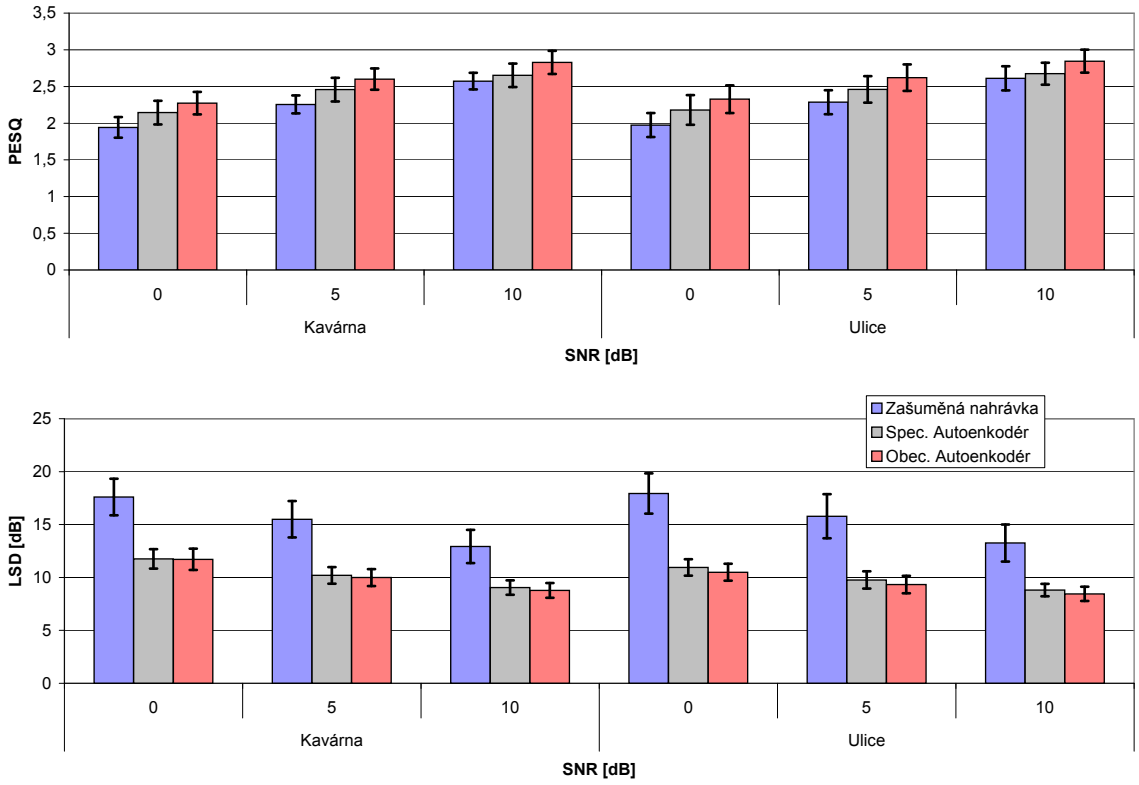
M je počet rámců

N je délka Fourierovy transformace

X je referenční spektrum a $\hat{X}$ spektrum porovnávané (zašuměné, odhadnuté...)

3.2 Porovnání specifického a obecného autoenkodéru

Nejprve byly na všech testovacích sadách porovnány autoenkodéry natrénované v této práci. Pro účely porovnání jsou uvedeny výsledky PESQ a LSD ve známém (Kafeterie) a neznámém (Ulice) šumovém prostředí v grafu 6. Sloupce uvádějí střední hodnotu a rozšířené úseky standardní odchylku.



Graf 6: Porovnání výsledků PESQ a LSD obecné a specifické varianty autoenkodéru pro odstranění šumu ve známých a neznámých šumových podmínkách

Výsledky jasně ukazují, že obecná verze autoenkodéru je schopna o něco efektivněji odstraňovat šum (nižší LSD znamená, že zpracovaný signál je podobnější nezašuměnému referenčnímu signálu) a zároveň zlepšovat srozumitelnost a kvalitu řečového signálu (vyšší PESQ skóre.)

Zajímavé je, že obecný autoenkodér měl proti specifickému autoenkodéru pro trénování k dispozici méně nahrávek obsahujících šum kavárna (12 hodin, 51 minut a 39 sekund pro specifickou variantu a 8 hodin, 34 minut a 26 sekund pro obecnou), i přesto však dosáhl při odstraňování šumu Kavárna lepších výsledků. Lepší výsledky PESQ u obecné varianty značí, že dokáže lépe vystihnout charakteristiku řečového signálu, který nebude při zpracování tolik poničen. Autor se domnívá, že tento fakt lze vysvětlit dvojnásobným množstvím řečových nahrávek při trénování a tedy lepší možností naučit se obvyklé spektrální charakteristiky řeči. Výsledky LSD jsou pro obě varianty velmi blízké, což znamená, že zpracované signály jsou srovnatelně podobné signálu referenčnímu.

Výsledky v neznámém prostředí (Ulice) předurčují k využití v reálném prostředí obecnou variantu (důvod - víc šumů při trénování znamená lepší robustnost v neznámých podmínkách), v dalších experimentech tedy již nebude uváděna specifická varianta.

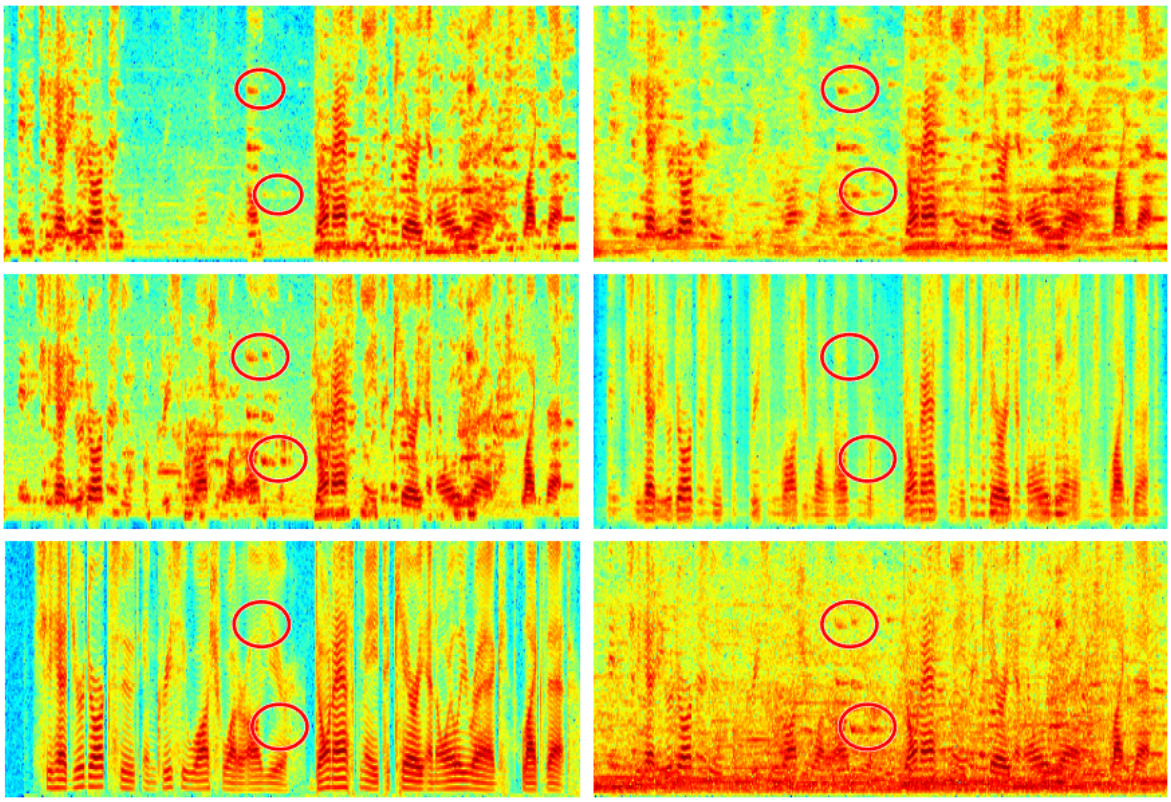
3.3 Porovnání autoenkodéru s konvenčními metodami

Tato kapitola porovnává výsledky obecného autoenkodéru s výsledky dříve uvedených konvenčně používaných metod (Odečítání spektra, MMSE STSA, OMLSA). Grafy opět uvádějí střední hodnotu a standardní odchylku. Medián byl pro všechny výsledky odchýlený od střední hodnoty do 5 %, v grafech tedy není pro přehlednost úváděn.

Graf 8 ukazuje výsledky jednotlivých metod při zpracovávání signálu poškozeného šumem Kavárna. Tento šum je vysoce nestacionární a velká část jeho energie vykazuje ve spektrální oblasti podobnost řečovému signálu. Pro většinu metod je velice obtížné rozlišit tento šum od řeči a efektivně jej odstranit.

Metoda odečítání spektra byla navržena pro stacionární šumy, což odrážejí její výsledky PESQ. Zpracované signály obsahují spoustu šumových residuí a hudebních artefaktů, které poškozují srozumitelnost řeči. Kladný efekt na srozumitelnost má tato metoda až pro vysoká počáteční SNR. MMSE STSA dosahuje lepších výsledků než odečítání spektra, při zpracování nevzniká tolik hudebních artefaktů, výsledná nahrávka je srozumitelnější a odstranění šumu začíná být efektivní už od nižších SNR. OMLSA zachovává řečový signál lépe než ostatní konvenční metody, nicméně vzhledem k charakteristice šumového prostředí spolu s řečí zůstává v signálu spousta šumových residuí, viz obrázek 9. Výsledky LSD všech konvenčních metod jsou na podobné úrovni.

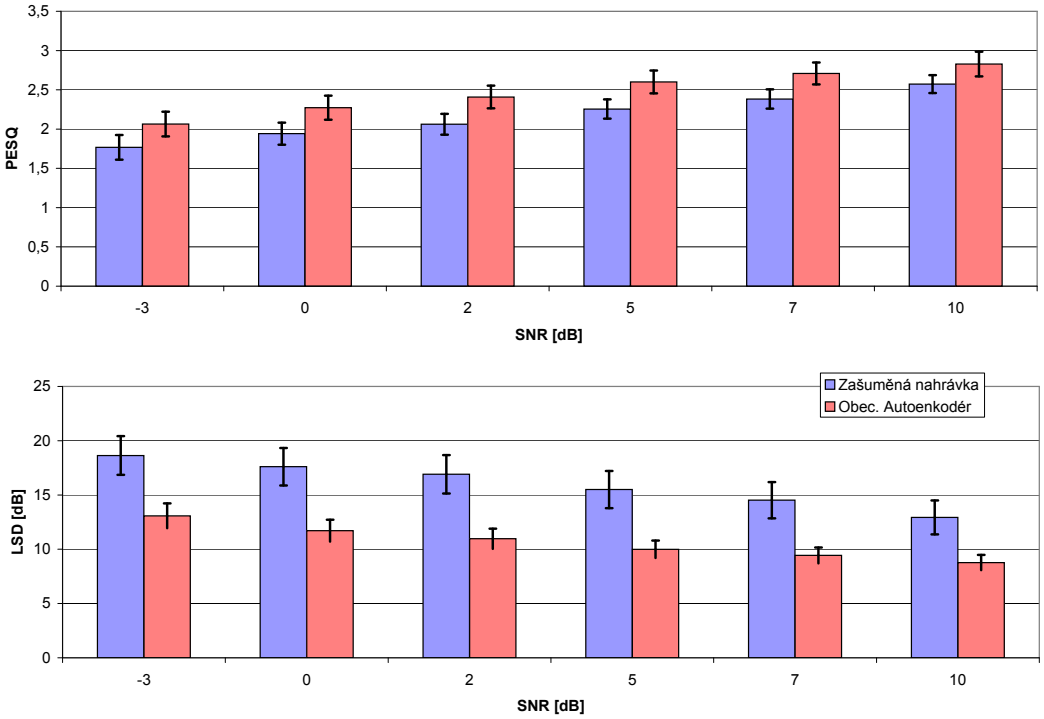
Pro obecný autoenkodér se jedná o známé prostředí, i přes to se však nejedná o triviální úlohu. Autoenkodér konzistentně dosahuje lepších výsledků PESQ než konvenční metody. Zpracované nahrávky neobsahují žádné hudební artefakty a téměř žádná šumová residua, jak je vidět na obrázku 9. Výsledky LSD jsou konzistentně lepší než u konvenčních metod.



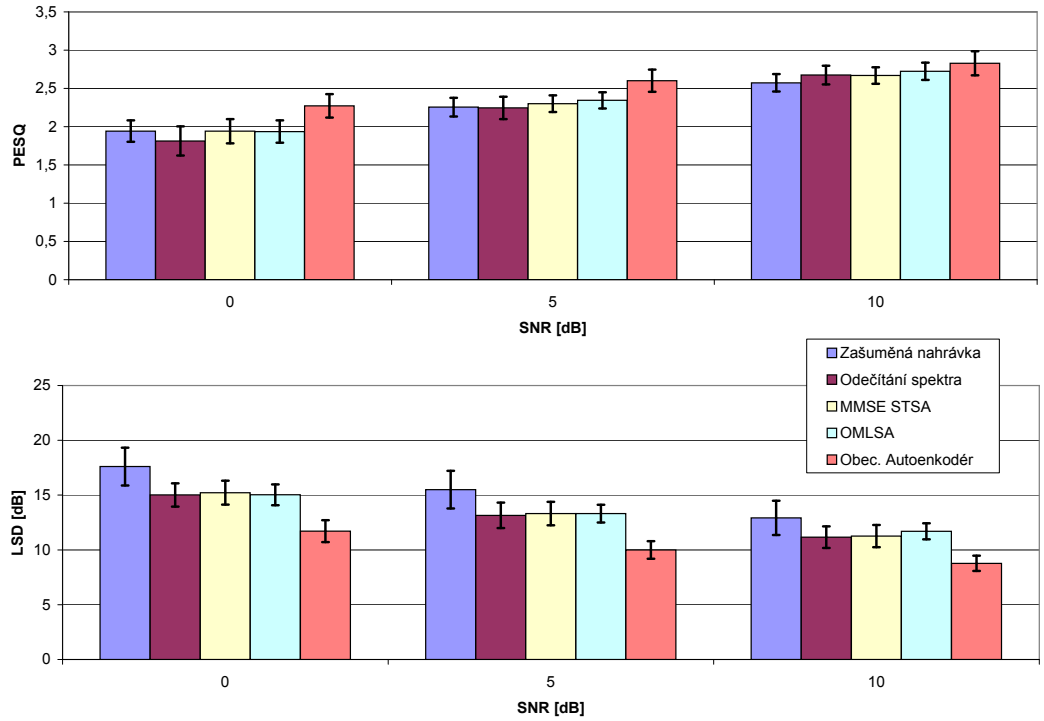
Obrázek 9: Spektrogramy signálu čisté řeči (vlevo dole), zašuměné řeči (vpravo dole), zpracovaného pomocí odečítání spektra (vlevo nahoře), MMSE STSA (vpravo nahoře), OMLSA (vlevo uprostřed) a obecného autoenkodéru (vpravo uprostřed). Znázorněné jsou pozice některých výrazných šumových residuí.

Také je nutno podotknout, že obecný autoenkodér dokázal zpracovat signály ve známých i neznámých úrovních SNR bez výrazného poklesu na efektivitě, jak je vidět v grafu 7. Tento trend byl pozorován u všech kombinací známých a neznámých šumových a jazykových podmínek. To naznačuje, že je možné očekávat adekvátní výsledky i přes ostatní neznámé úrovně SNR, které se v nahrávkách pořízených v reálném prostředí objevují.

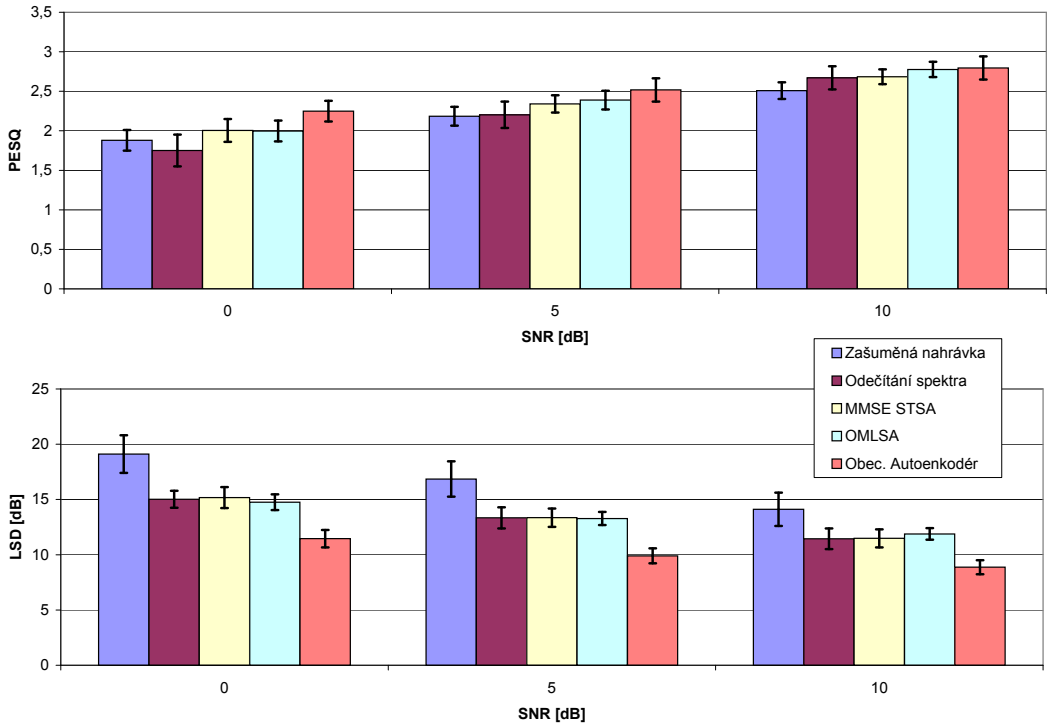
Podobné výsledky byly zaznamenány i pro šum Lidé, zobrazeny v grafu 9. Tento šum je také nestacionárního charakteru, nicméně neobsahuje tak extrémní krátkodobé jevy jako šum Kafeterie. To zvyšuje efektivitu konvenčních metod, obvzláště odečítání spektra, která začíná zlepšovat srozumitelnost řečového signálu od SNR = 5 dB a výše. Metoda OMLSA opět dosahuje nejlepšího zachování řečového signálu, ovšem se značným množstvím šumových residuí a v některých případech dochází izolováním ojedinělých řečových komponent ve vyšších frekvencích k nepříjemným hudebním artefaktům.



Graf 7: Výsledky obecného autoenkodéru pro známé a neznámé úrovně SNR



Graf 8: Výsledky PESQ a LSD, šum Kavárna



Graf 9: Výsledky PESQ a LSD, šum Lidé

Pro obecný autoenkodér se opět jedná o známé prostředí a výsledky PESQ a LSD i zde nasvědčují, že neuronová síť byla schopna naučit se správné vztahy mezi zašuměnou a čistou řečí. Nahrávky zpracované touto metodou opět obsahují nejméně artefaktů a zbytkového šumu, nicméně často dochází k lehkému poškození řečového signálu.

Graf 10 znázorňuje výsledky jednotlivých metod pro šum Autobus. Tento šum je nejvíce stacionární a soustředěný hlavně v nízkých frekvencích. Zdánlivě se jedná o jednoduchý šum k odstranění, nicméně právě soustředění většiny jeho energie do nízkých frekvencí způsobuje, že signál řeči je v těch samých frekvencích silně zastíněn (velmi nízké lokální SNR) a může být těžké ho rekonstruovat. Naopak vysoké frekvence neobsahují tolik šumu a zlepšující metody v nich mohou způsobit více škody na srozumitelnosti zkreslením užitečného řečového signálu než užitku odstraněním nežádoucího šumu.

Konvenční metody, zvláště pak metody vytvářející si odhad spektra na základě prvotního ticha (initial silence; úsek na začátku nahrávky, kde se vyskytuje pouze šum), tj. odečítání spektra a MMSE STSA, na tomto typu šumu fungují nejlépe ze všech ostatních šumů, neboť jejich odhad šumu se po celou nahrávku příliš neliší od skutečného šumu. To se odráží ve výsledcích, zvláště metoda odečítání spektra funguje velice dobře. Žádná z konvenčních metod nebyla schopná při nízkých hodnotách globálního SNR dobře rekonstruovat nejnižší řečové

frekvence.

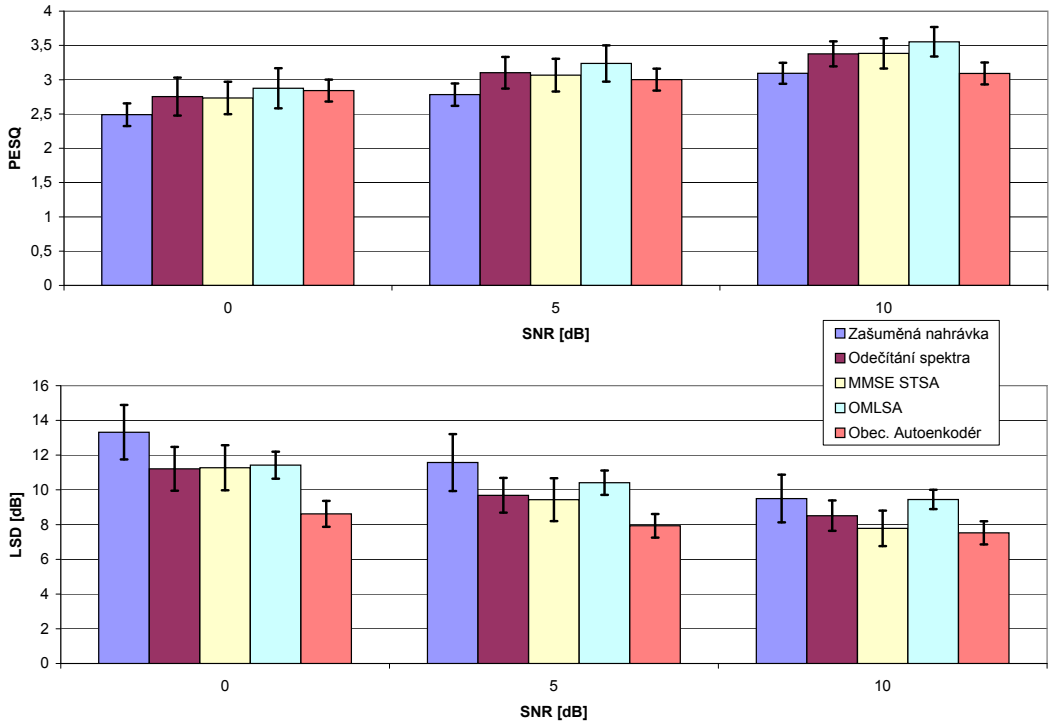
Zajímavé je, že metoda OMLSA sice dosahuje nejlepších výsledků PESQ, ale nejhorších výsledků LSD. To je dáno tím, že OMLSA je tzv. non-distortion (nezkreslující) metodou - snaží se zachovávat spektrální charakteristiky řeči i za cenu nižší redukce šumu.

Obecný autoenkodér pro odstranění šumu má opačný problém než konvenční metody. Nízké frekvence jsou i přes vysokou úroveň šumu dobře rekonstruovány, nicméně vyšší frekvence jsou lehce poškozeny zkreslením. Tento nedostatek je patrný z výsledků PESQ, kdy pro vysoká počáteční SNR obecný autoenkodér nedosahuje žádného zlepšení i přes dobré potlačení šumu. Tomuto nedostatku se více věnuji v kapitole 3.4. Pro nízká počáteční SNR však stále dosahuje obecný autoenkodér nejlepších výsledků PESQ a celkově přes všechna testovaná počáteční SNR i nejlepších výsledků LSD.

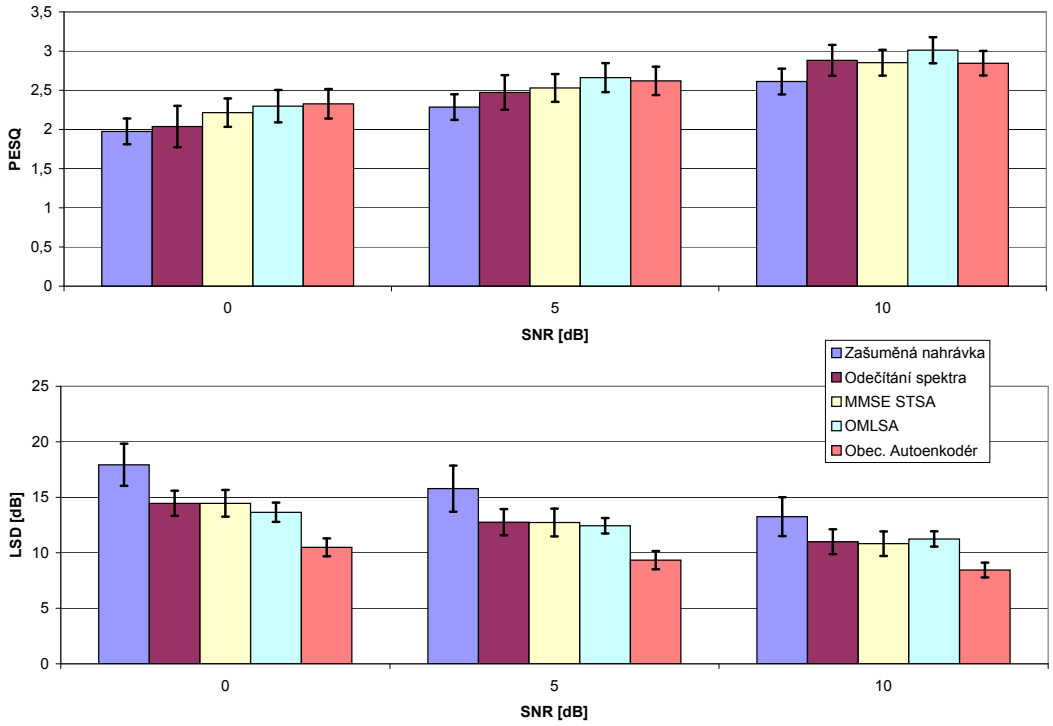
Posledním testovaným šumem byl šum Ulice, jehož výsledky jsou zobrazeny v grafu 11. Jedná se o šum podobný šumu Autobus, energie v šumu Ulice je však více rovnoměrně rozprostřena přes všechny frekvence. Projížděním okolo mikrofonu způsobují motorové dopravní prostředky silné kolísání energie šumového spektra a tím získává šum středně nestacionární charakter.

Konvenční metody zde dosahují o něco horších výsledků než u šumu Autobus. Metoda odečítání spektra je založena na předpokladu o stacionaritě šumu, který šum Ulice nesplňuje, a nedokáže tak pro nízká počáteční SNR (0 dB a níže) příliš zlepšovat poslechovou kvalitu signálu. Zbylé dvě metody se umí lépe adaptovat změnám v šumovém spektru, zvláště metoda OMLSA dosahuje lepších výsledků PESQ než obecný autoenkodér pro SNR = 5 dB a výše.

Pro obecný autoenkodér se jedná o neznámé šumové prostředí, nicméně výsledky LSD ukazují, že systém dokázal efektivně redukovat šum v řečové nahrávce lépe než konvenční metody a pro SNR $<$ 5 dB i lépe zvyšovat srozumitelnost a kvalitu řeči.



Graf 10: Výsledky PESQ a LSD, šum Autobus

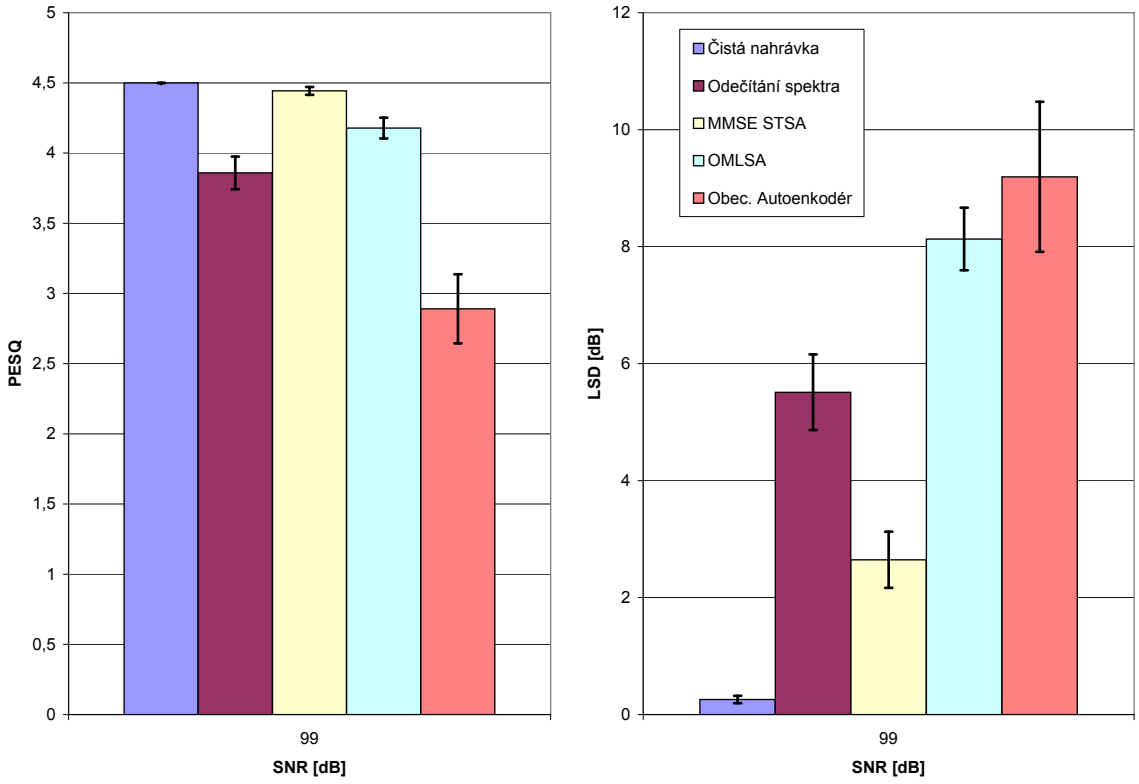


Graf 11: Výsledky PESQ a LSD, šum Ulice

3.4 Poškození čistého řečového signálu

Všechny testované metody nějakým způsobem poškozují čistý řečový signál. K evaluaci míry poškození byla vytvořena speciální testovací sada, na kterou byl symbolicky aplikován šum Autobus s úrovní SNR = 99 dB. To je nutné kvůli implementaci některých kritérií, která nebyla navržena na porovnávání signálu sama se sebou, nicméně šum je v takových nahrávkách prakticky neregistrovatelný objektivně ani subjektivně, jak je vidět v grafu 12.

Výsledky potvrzují jevy pozorované při zpracovávání nahrávek s vysokám počátečním SNR, zvláště při odstraňování šumu Autobus - obecný autoenkodér natrénovaný v této práci poškozuje čistou řečovou nahrávku nejvíce ze všech testovaných metod. Dosažené výsledky PESQ a LSD zhruba odpovídají maximálně dosaženým výsledkům v kapitole 3.3, což poukazuje na určitý limit schopností obecného autoenkodéru pro modelování charakteristik čistého řečového signálu.

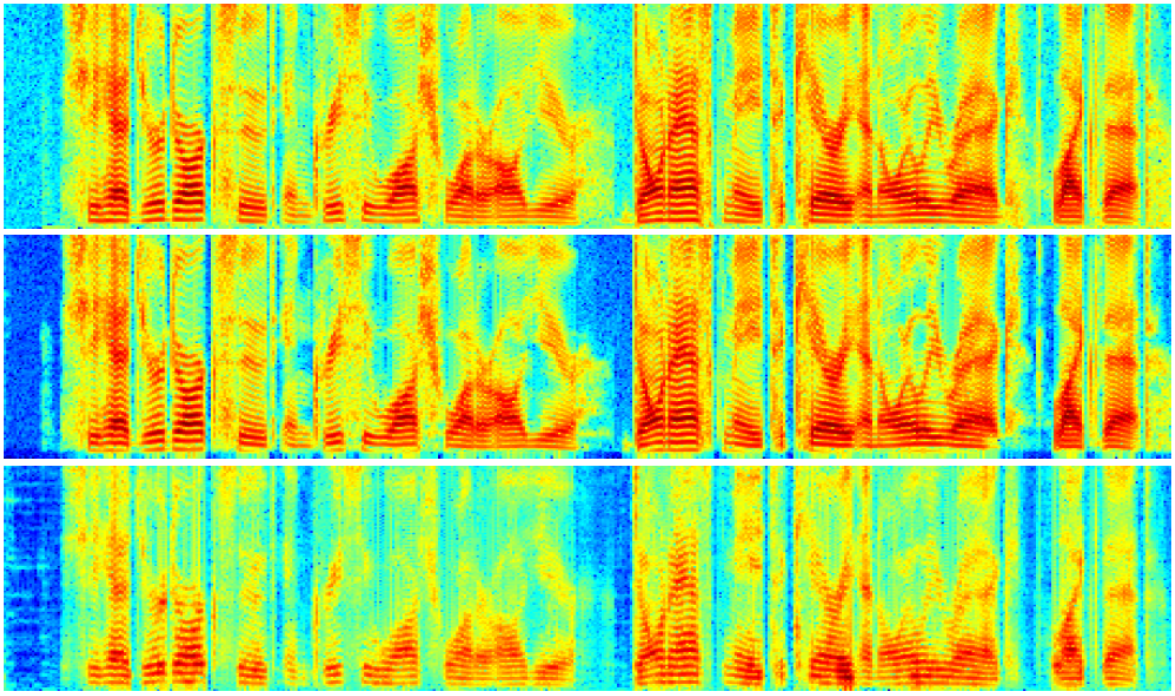


Graf 12: Míra poškození čistého řečového signálu po zpracování různými metodami

Na obrázku 10 je vidět, že narozdíl od nezkreslující metody OMLSA obecný autoenkodér čisté řečové spektrum zkresluje, a to hlavně v jeho nejslabších částech. Potlačení těchto slabých řečových komponent vede ke změně barvy hlasu a ke ztrátě srozumitelnosti, což koreluje s nízkým zvýšením PESQ skóre u některých druhů šumu při vysoké počáteční úrovni SNR.

Tento nedostatek by nejspíše šlo zmírnit přidáním dvojic s čistou řečí na vstupu i výstupu sítě do trénovací sady. Zvětšení trénovací sady o nahrávky s vysokou úrovní SNR (15 dB, 20 dB a výše) by také mohlo mít kladný efekt na schopnosti zachování řečového signálu.

Nejmenšího poškození dle měřených metrik dosahuje metoda MMSE STSA, a to v obou kritériích, prakticky však při poslechu není mezi konvenčními metodami téměř žádný rozdíl.



Obrázek 10: Spektrogramy signálu řeči zašuměné na úroveň SNR = 99 dB (nahoře), po zpracování metodou OMLSA (uprostřed) a po zpracování obecným autoenkodérem (dole)

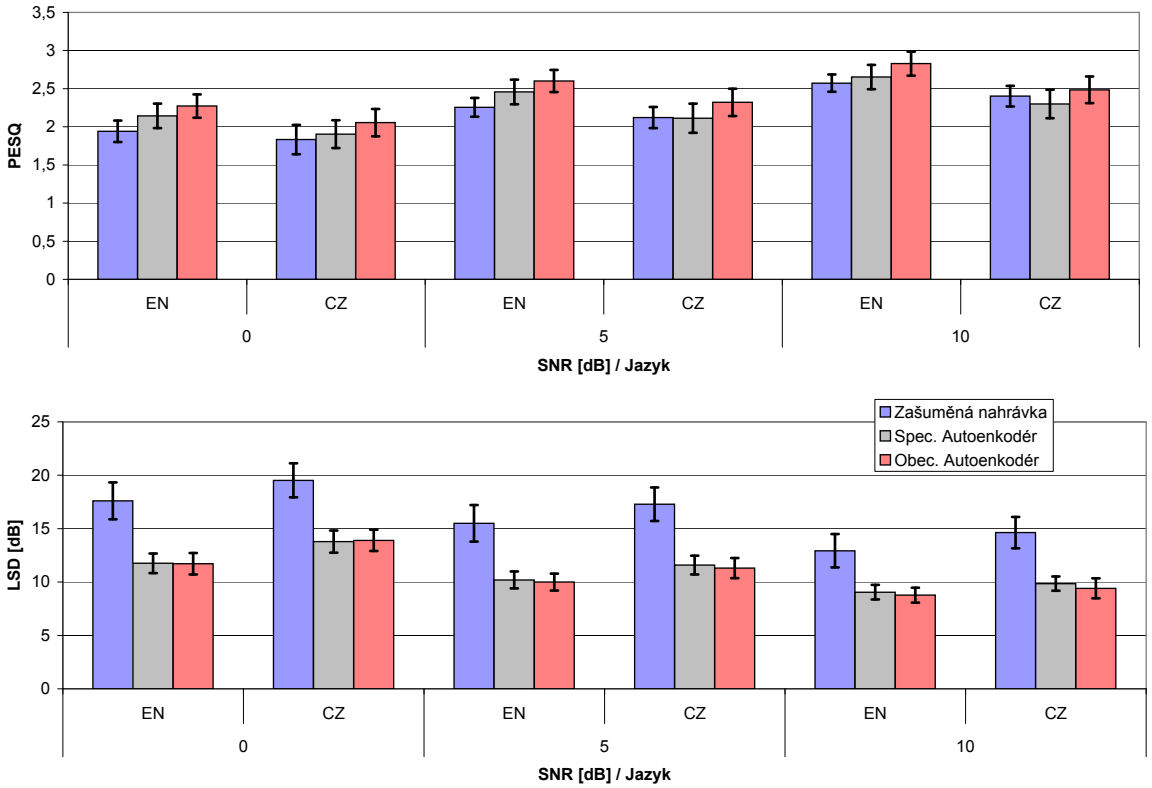
3.5 Vliv jazyka na výsledky

Databáze české řeči zmíněná v kapitole 2.2.6 byla využita k evaluaci robustnosti odstraňování šumu při výskytu neznámého jazyka.

Pro izolaci vlivu neznámého jazyka od ostatních vlivů (např. neznámý druh šumu) byly vybrány výsledky pro šum Kavárna, který je pro oba autoenkodéry známým prostředím. Výsledky jsou k vidění v grafu 13.

PESQ skóre, a tedy i poslechová kvalita a srozumitelnost, jsou nižší než při zpracovávání nahrávek v anglickém jazyce. To je nejspíše vlivem neznámé fonetické skladby jednotlivých slabik. Zde je nutno vzít v potaz délku signálu, se kterou autoenkodéry pracují - 11 rámců, každý rámec zahrnuje 32 ms nahrávky. I s polovičním překryvem rámců se tak jedná o 192 ms řeči, což je dostatečně dlouhý úsek pro obsažení celé slabiky. Klasické české „tvrdé“ fonémy (g, c, ř aj.) tak autoenkodéry nemohou z trénování znát a mohou je nesprávně interpretovat jako šum.

Výsledky LSD značí, že změna jazyka řečníka neměla příliš velký vliv na schopnost autoenkodérů odstraňovat z nahrávky šum.



Graf 13: Výsledky PESQ a LSD na anglické a české testovací databázi, šum Kavárna

3.6 Porovnání výpočetní náročnosti

Další důležitou vlastností systému pro odstranění šumu je rychlost zpracování nahrávky. Pokud je systém dostatečně výpočetně nenáročný, pak jej lze potenciálně využít pro zpracování signálu v reálném čase, např. v mobilním telefonu. Pro otestování této vlastnosti bylo vytvořeno několik databází testovacích nahrávek různých délek, které byly zpracovány všemi metodami.

Výpočetní náročnost byla měřena v rámci času potřebného na zpracování jednotlivých nahrávek. Průměrné hodnoty jsou k vidění v tabulce 1.

Tabulka 1: Doba potřebná ke zpracování signálu testovanými metodami

	Průměrná délka zpracování nahrávky (s)			
Průměrná délka nahrávky (s)	Odečítání spektra	MMSE STSA	OMLSA	Obecný Autoenkodér
1,05	0,05	0,05	0,57	0,09
3,03	0,16	0,13	1,58	0,19
4,59	0,19	0,20	2,50	0,28
6,32	0,32	0,26	3,30	0,35
9,26	0,38	0,39	4,81	0,48
18,51	0,79	0,78	9,74	0,92
60,00	2,84	2,67	32,80	3,61

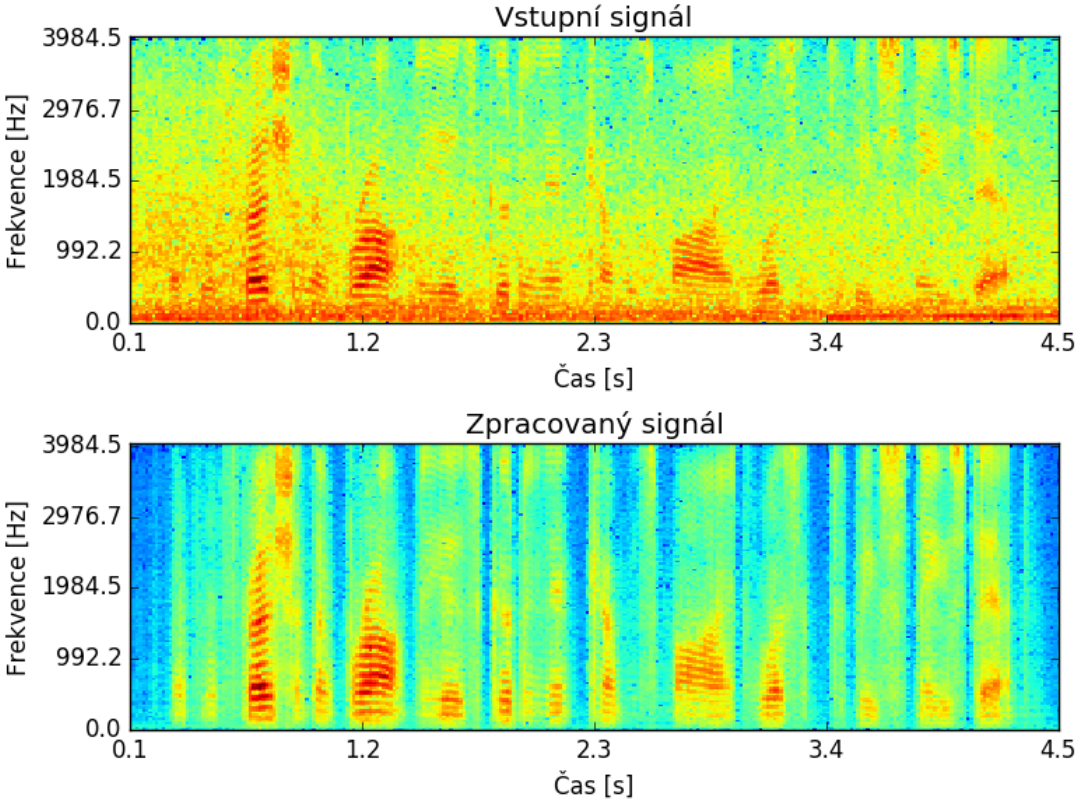
Zajímavé je, že i přes svou značně vyšší složitost dosahuje metoda MMSE STSA podobných a v některých případech i lepších výsledků než metoda odečítání spektra. Autoenkodér natrénovaný v této práci dosahuje srovnatelných výsledků, obvzláště při zpracovávání delších nahrávek. Algoritmus OMLSA se zdá být výpočetně nejnáročnější, doba zpracování jedné nahrávky se pohybuje o řád výše než doba, kterou potřebují ostatní metody.

Další zajímavostí je, že doba zpracování jedné nahrávky obecným autoenkodérem roste v závislosti na délce vstupního signálu pomaleji než u konvenčních metod. To je dáno tím, že zpracování signálu autoenkodérem probíhá pomocí operací násobení matic (vážení signálů), sčítání matic (sumační funkce neuronu) a aplikaci aktivační funkce. Všechny tyto operace je možné zrychlit pomocí dedikovaných knihoven, přičemž zrychlení se projevuje více s větší velikostí operandů. To však souvisí s jednou nevýhodou - pro rychlé zpracování signálu je optimální volit pro zpracování co nejdelší úseky vstupních dat, to však znamená vyšší paměťovou náročnost, která je už ze základu u neuronových sítí nezanedbatelná (je nutné uchovávat miliony parametrů ve formátu desetinných čísel s plovoucí řádovou čárkou).

Nutno podotknout, že zde uvedené výsledky vycházejí z výchozí implementace od autorů algoritmů a tento experiment se nezabývá možnými implementačními optimalizacemi. Také je nutno podotknout, že i přes relativní výpočetní nenáročnost při použití neuronových sítí je třeba brát v úvahu čas potřebný k natrénování takové sítě.

4 Aplikace

Obecný autoenkodér byl po natrénování využit v rámci aplikace pro odstranění šumu z řečových nahrávek (ve formátu „wav“, se vzorkovací frekvencí 8 kHz). Aplikace byla implementována v jazyce Python 2.7. Aplikace je schopna zpracovávat nahrávky jednotlivě a dávkově. Pokud je pracovávána pouze jedna nahrávka, pak je možné porovnat původní a zpracovaný signál pomocí vykreslených spektrogramů, viz obrázek 11.



Obrázek 11: Grafický výstup z aplikace

Také je možné specifikovat jméno výstupního souboru. V dávkovém módu umožňuje aplikace zpracovávat soubory hromadně, jako vstupní parametr pak přijímá cestu k adresáři s nahrávkami ke zpracování. V dávkovém módu aplikace nezobrazuje spektrogramy a názvy výstupních souborů jsou automaticky odlišeny příponou „_denoised“, např. „soubor_1.wav“ je zpracován a výsledná nahrávka je uložena jako „soubor_1_denoised.wav“ ve stejném adresáři.

5 Závěr

V rámci diplomové práce byl navržen systém pro odstraňování šumu z řečových nahrávek založený na trénování hlubokých neuronových sítí (autoenkodér pro odstraňování šumu). K tomuto účelu byly vytvořeny dvě trénovací sady, jedna zaměřená na jeden konkrétní druh šumu a druhá obsahující širší spektrum druhů šumu. Pomocí těchto trénovacích sad byly natrénovany dvě hluboké neuronové sítě a jejich schopnosti byly pomocí kritérií PESQ a LSD otestovány ve známých i neznámých prostředích. Pro otestování robustnosti při odstraňování šumu v nahrávkách s neznámým jazykem řečníka byla autorem práce sestavena a nahrána malá databáze české řeči.

Výsledky testování ukazují, že autoenkodér trénovaný na širším spektru šumů (obecný autoenkodér) dosahoval lepších výsledků než autoenkodér trénovaný na jednom druhu šumu.

Dále byly výsledky obecného autoenkodéru porovnány s výsledky tří konvenčních metod (odečítání spektra, MMSE STSA, OMLSA). Ukázalo se, že obecný autoenkodér natrénovaný v této práci dosahuje konzistentně lepších výsledků LSD (menší logaritmické vzdálenosti mezi výkonovými spektry signálů) než všechny testované konvenční metody. Dále byla měřena poslechová kvalita a srozumitelnost zpracovaných signálů pomocí PESQ. Výsledky ukázaly, že obecný autoenkodér dosahuje vyššího zlepšení PESQ skóre pro všechny testované varianty silně nestacionárních druhů šumu (kavárna, lidé) než konvenční metody. Pro stacionární a lehce nestacionární druhy testovaných šumů (autobus, ulice) dosahuje obecný autoenkodér lepších PESQ skóre než konvenční metody pouze pro nízká počáteční SNR (< 5 dB).

Robustnost při odstraňování šumu obecným autoenkodérem byla otestována pomocí neznámých úrovní SNR, neznámého šumu a neznámého jazyka řečníka. Výsledky ukázaly, že obecný autoenkodér je schopný zpracovávat řečové nahrávky s neznámými šumy a v neznámých úrovních SNR bez znatelné ztráty funkcionality. Při zpracovávání nahrávek s neznámým jazykem řečníka byl zaznamenán úbytek srozumitelnosti a poslechové kvality výsledků vlivem výskytu neznámé fonetické skladby, i přesto ale obecný autoenkodér konkuruje schopnostem konvenčních metod a překonává je pro nízké počáteční úrovně SNR.

Ve všech testovaných případech obsahovaly řečové nahrávky zpracované obecným autoenkodérem nejméně šumových residuí ze všech testovaných metod. Navíc nebyly zaznamenány žádné hudební artefakty, které se často vyskytují při použití konvenčních metod.

Naproti tomu bylo pomocí testování na čistých řečových nahrávkách ukázáno, že obecný autoenkodér natrénovaný v této práci není plně schopen vyjádřit slabé řečové komponenty,

čímž degraduje kvalitu řeči. Z tohoto důvodu může být použití konvenčních metod výhodnější pro odstraňování slabého a stacionárního šumu.

Dále byla porovnána výpočetní náročnost jednotlivých metod pomocí empirických testů. Bylo ukázáno, že obecný autoenkodér natrénovaný v této práci konkuruje v době potřebné pro zpracování nahrávky konvenčním metodám odečítání spektra a MMSE STSA a překonává metodu OMLSA.

Váhy a biasy obecného autoenkodéru byly využity k implementaci aplikace, která umožňuje jednotlivě a dávkově zlepšovat řečové nahrávky odstraněním šumu. Aplikace též umožňuje vizuálně pomocí spektrogramů porovnávat výsledky tohoto zlepšení. Tímto byly splněny všechny body zadání.

Výsledky obecného autoenkodéru by nejspíše mohly být zlepšeny několika způsoby, například:

- Větším počtem neuronů ve skrytých vrstvách pro lepší schopnosti modelování vztahů mezi vstupním a výstupním spektrem
- Rozšířením trénovací sady o nahrávky obsahující pouze čistou řeč pro lepší vystižení charakteristických vlastností řeči, popřípadě o nahrávky obsahující vyšší úrovně SNR
- Rozšířením trénovací sady o jiné druhy šumu a jiné jazyky řečníků pro zajištění větší robustnosti při odstraňování šumu v neznámých podmínkách
- Delší dobou trénování, neboť optimalizační kritérium během trénování nevykazovalo známky tendence divergovat a trénování bylo zastaveno z časových důvodů
- Změnou topologie neuronové sítě, například na konvoluční
- Dodatečným zpracováním výstupních signálů metodou, která by se pokusila odstranit zkreslení řečového signálu vzniklé při odstraňování šumu a tím zvýšit poslechovou kvalitu nahrávek

Použitá literatura

[1] DAVIES, Mike E. a JAMES, Christopher J. Source separation using single channel ICA. *Signal Processing*. 2007, 87, s. 1819-1832. ISSN 0165-1684.

[2] BOLL, Steven F. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*. 1979, 27(2), s. 113-120. ISSN 0096-3518.

[3] EPHRAIM, Yariv a MALAH, David. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 1984, 32(6), s. 1109-1121. ISSN 0096-3518.

[4] KAWAMURA, Arata, THANHIKAM, Weerawut a IIGUNI, Youji. Single Channel Speech Enhancement Techniques in Spectral Domain. *ISRN Mechanical Engineering*. 2012 [cit. 18.12.2015]. ISSN 2090-5130. Dostupné z: doi:10.5402/2012/919234

[5] COHEN, Israel a BERDUGO, Baruch. Speech enhancement for non-stationary noise environments. *Signal processing*. 2001, 81(11), s. 2403-2418. ISSN 0165-1684.

[6] WAN, Eric. A. a NELSON, Alex T. Networks for speech enhancement. In: *Handbook of neural networks for speech processing*. Boston, USA: Artech House, 1999, kapitola 16. ISBN 978-0890069547.

[7] BIANCHINI, Monica a SCARSELLI, Franco. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*. 2014, 25(8), s. 1553-1556. ISSN 2162-237X.

[8] HINTON, Geoffrey E. Learning multiple layers of representation. *Trends in Cognitive Sciences*. 2007, 11, s. 428-434. ISSN 1364-6613.

[9] XU, Yong, DU, Jun, DAI, Li-Rong a LEE, Chin-Hui. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*. 2014, 21(1), s. 65-68. ISSN 1070-9908.

[10] XU, Yong, DU, Jun, DAI, Li-Rong a LEE, Chin-Hui. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2015, 23(1), s. 7-19. ISSN 2329-9290.

[11] BOAZ, Porat. A course in digital signal processing. New York: Wiley, 1997, 1. ISBN 978-0471149613.

[12] ITU-R Recommendation BS.468-4. *Measurement of Audio-Frequency Noise Voltage Level in Sound Broadcasting*. ITU, 1986.

[13] HU, Yi a LOIZOU, Philipos C. Subjective comparison of speech enhancement algorithms. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing.* 2006, 1, s. 153-156. ISSN 1520-6149.

[14] MCCULLOCH, Warren S. a PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics.* 1943, 5(4), s. 115-133.

[15] WERBOS, Paul. Beyond regression: New tools for prediction and analysis in the behavioral sciences. 1974.

[16] HINTON, Geoffrey E., OSINDERO, Simon a TEH, Yee-Whye. A fast learning algorithm for deep belief nets. *Neural computation.* 2006, 18(7), s. 1527-1554. ISSN 0899-7667.

[17] NGUYEN, Derrick a WIDROW, Bernard. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In: *Proceedings of the International Joint Conference on Neural Networks.* 1990, 3, s. 21-26.

[18] BENGIO, Yoshua. Practical recommendations for gradient-based training of deep architectures. In: *Neural Networks: Tricks of the Trade.* Springer Berlin Heidelberg, 2012, s. 437-478. ISBN 978-3-642-35289-8.

[19] BURNEY, Syed Muhammad Aqil, JILANI, Tahseen Ahmed a ARDIL, Cemal. A Comparison of First and Second Order Training Algorithms for Artificial Neural Networks. In: *International Conference on Computational Intelligence.* Istanbul, Turecko, 2004, s. 12-18. ISBN 975-98458-1-4.

[20] GAROFOLO, John S., et al. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N 93.* 1993.

[21] BARKER, John, MARXER, Ricard, VINCENT, Emmanuel a WATANABE, Shinji. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. *IEEE 2015 Automatic Speech Recognition and Understanding Workshop.* 2015.

[22] RIX, Antony W., et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing.* 2001, 2, s. 749-752.

[23] DU, Jun a HUO, Qiang. A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions. In: *INTERSPEECH.* 2008, s. 569-572. ISBN 9781615673780.

Obsah přiloženého CD

- text diplomové práce
- zdrojový kód aplikace popsané v kapitole 4, návod k použití
  - DAE.py
  - README.txt
- Databáze české řeči popsaná v kapitole 2.2.6