Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

BAKALÁŘSKÁ PRÁCE

Plzeň, 2016

Martin Majer

Prohlášení

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne 13. května 2016 .....

Poděkování

Tímto bych rád poděkoval vedoucímu bakalářské práce, Ing. Luboši Šmídlovi, Ph.D., za cenné rady a připomínky.

Anotace

Tato práce se zabývá klasifikací izolovaných slov pomocí neuronových sítí, klasifikátoru SVM a klasifikátoru založeném na algoritmu Dynamic Time Warping s ohledem na nízkou výpočetní náročnost. V první části jsou představeny příznaky v časové a frekvenční oblasti a odvozeny využité klasifikační algoritmy. Ve druhé části jsou uvedeny zvolené parametrizace testovaných příznaků a struktura navržených klasifikačních algoritmů. V závěru práce je pak vyhodnocena přesnost klasifikace jednotlivých metod pro zvolené parametrizace příznaků.

Klíčová slova: zpracování akustického signálu, extrakce příznaků, detekce klíčových frází, dynamic time warping, support vector machine, neuronová síť

Abstract

This thesis focuses on low computational cost isolated word recognition using neural networks, SVM classifier and Dynamic Time Warping based classifier. First part of the thesis introduces features in time and frequency domain and used classification techniques are derived. Parameterizations of tested features and structure of proposed classification algorithms are described in the second part of the thesis. Classification accuracy results of proposed methods for feature parameterizations are presented at the end of the thesis.

Keywords: acoustic signal processing, feature extraction, keyword spotting, dynamic time warping, support vector machine, neural network

Obsah

1 Úvod

V posledních letech došlo k významnému vývoji v oblasti mobilních zařízení a s tím i k nárůstu popularity hlasového ovládání. Aplikace společnosti Google nabízejí možnost vyhledávání hlasem, Apple a Microsoft vytvořili inteligentní osobní asistentky Siri, resp. Cortana a stále častěji se hlasové ovládání začíná objevovat i v automobilovém průmyslu.

Jedním z řešených problémů hlasového ovládání je detekce klíčových frází v proudu řeči (KWS - Keyword spotting). Na pozadí zařízení nepřetržitě běží KWS systém, který naslouchá mluvenému slovu a při zaznění klíčové fráze provede určitou akci. V dnešní době se nejčastěji používají dva typy KWS systémů - systémy s předdefinovanými klíčovými frázemi a systémy využívající obsáhlý slovník řeči v textové podobě (LVCSR - Large vocabulary continuous speech recognition). LVCSR systémy jsou výpočetně velmi náročné, jelikož je potřeba nejprve převést mluvenou řeč do textové podoby a následně vyhledat klíčovou frázi v databázi. Vzhledem k tomu, že tyto systémy neumožňují vytvoření vlastní bezpečné klíčové fráze, propojují se většinou se systémem pro identifikaci řečníka [1].

Bylo by tedy vhodné vytvořit KWS systém, který by nebyl závislý na předem definovaných frázích a umožnil by uživateli volbu vlastních klíčových frází v libovolném jazyce. Také by měl být výpočetně nenáročný, aby byla zajištěna odezva v reálném čase na mobilních zařízeních. Tyto požadavky splňuje metoda Query-by-Example, kdy si uživatel vytvoří vzorovou klíčovou frázi a všechny budoucí promluvy jsou porovnávány vůči ní [2, 3].

Práce se věnuje zpracování akustického signálu a především problematice rozpoznání izolovaných slov. Uvedeme si nejčastěji využívané algoritmy strojového učení a porovnáme jejich rozpoznávací schopnosti a jejich vhodnost potenciálního využití v KWS systému.

2 Klasifikace

Klasifikace neboli rozpoznávání je úloha, kde je cílem správně zařadit objekt do jedné z k tříd. Aby bylo možné objekt klasifikovat, je třeba ho nejprve vhodně popsat pomocí tzv. příznakového vektoru (obrazu). Příznakový vektor se skládá z příznaků, které reprezentují jednotlivé měřitelné či pozorovatelné vlastnosti objektu. Příznaky je třeba volit v závislosti na řešeném problému tak, aby dostatečně popisovaly pozorovaný objekt. Může se zdát, že vytvořením příznakového vektoru ze všech měřitelných veličin objektu získáme dokonalý popis objektu, který přispěje k přesnosti klasifikace. Opak je však pravdou, jelikož velké množství příznaků s nedostatečnou informativní hodnotou může mít negativní vliv na přesnost klasifikace a vést k přetrénování modelu [4].

Algoritmus, který provádí klasifikaci, se nazývá klasifikátor. Tato práce se zabývá především klasifikátory založenými na učení s učitelem a klasifikátoru podle minimální vzdálenosti k vzorovým obrazům.

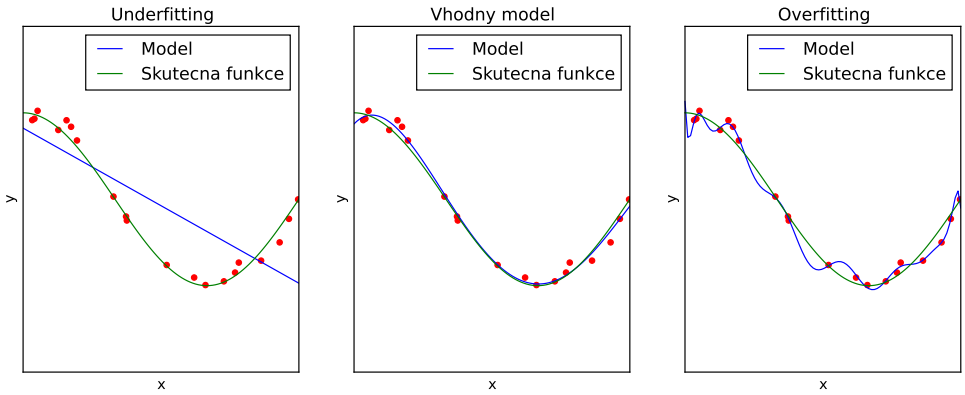
Učící se klasifikátor pracuje ve dvou fázích - trénovací fáze a klasifikační fáze:

1. **trénovací fáze** - v trénovací fází jsou klasifikátoru předkládány dvojice $\{x_i, y_i\}$, $i = 1, 2, \ldots, l$, kde x_i jsou příznakové vektory a y_i jsou cílové třídy jednotlivých obrazů z trénovací množiny. Klasifikátor vypočte svůj výstup pro příznakový vektor x_i (odhadovaná třída) a porovná jej s cílovou třídou. V případě neshody odhadované a cílové třídy upraví své parametry tak, aby minimalizoval danou chybovou funkci (jedná se tedy o optimalizační úlohu). Trénovací dvojice jsou klasifikátoru předkládány tak dlouho, dokud nedosáhne optimálního nastavení.

2. **klasifikační fáze** - klasifikátoru předkládáme neznámé obrazy a na základě natrénovaných parametrů klasifikujeme do jedné z k tříd [5, 6].

Při trénování dochází velmi často k tzv. přetrénování modelu (overfitting). Klasifikátor se naučí dokonale rozpoznávat pozorování z trénovací množiny, ale ztrácí schopnost generalizace a nová, neznámá pozorování klasifikuje chybně. Overfitting je způsoben volbou příliš komplexního modelu, nedostatkem trénovacích dat nebo vysokou dimenzí obrazů. Komplexitu modelu je tedy třeba volit s ohledem na povahu klasifikovaných dat, popř. lze využít jednu z metod, která overfitting omezí (regularizace, cross-validation, dropout vrstvy u neuronových sítí, atd.) [7]. Opačným případem overfittingu je tzv. underfitting,

kdy model na úkor generalizace ztrácí klasifikační schopnosti. Vliv volby modelu s různým počtem stupňů volnosti můžeme vidět na obrázku 1.



Obrázek 1: Příklad overfittingu a underfittingu.

3 Zpracování akustického signálu

Základním krokem pro klasifikaci řeči je samotné zpracování akustických signálů a jejich transformace na příznakové vektory. Při zpracování akustického signálu předpokládáme, že se jeho vlastnosti mění pomalu a na krátkých úsecích je téměř stacionární (v časové i frekvenční oblasti). Signál je tedy rozdělen na kratší úseky neboli mikrosegmenty, které jsou zpracovány samostatně, jako by se jednalo o různé signály. Pro každý mikrosegment pak vypočteme číslo (příznak), popřípadě vektor čísel na základě zvolené metriky. Délka mikrosegmentů se nejčastěji volí mezi 10 až 25ms a jednotlivé mikrosegmenty na sebe mohou navazovat nebo se i překrývat. Díky tomu lze celý signál popsat posloupností čísel (příznakový vektor), jejíž délka je přímo úměrná délce slova a počtu mikrosegmentů [4].

3.1 Okénková funkce

Jednotlivé mikrosegmenty jsou ze signálu vybírány pomocí okénka o určité délce, které se posouvá o daný počet vzorků signálu. Okénko také slouží k přidělení vah zpracovávaným vzorkům signálu. Mezi nejčastěji používané okénkové funkce při zpracování řeči patří pravoúhlé a Hammingovo okénko:

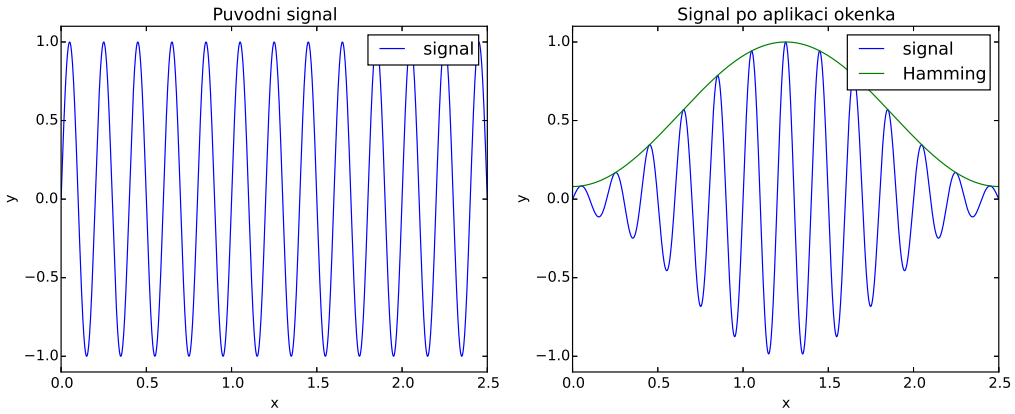
- **pravoúhlé okénko** - pravoúhlé okénko přidělí všem vzorkům mikrosegmentu stejnou váhu. Lze jej definovat vztahem

$$w(n) = \begin{cases} 1 & \text{pro } 0 \leq n \leq N-1 \\ 0 & \text{jinak,} \end{cases} \tag{3.1}$$

  kde N je počet vzorků mikrosegmentu.

- **Hammingovo okénko** - Hammingovo okénko je vhodné použít v případě, kdy je třeba potlačit vzorky na krajích zpracovávaného mikrosegmentu [4, 8]. Lze jej definovat vztahem

$$w(n) = \begin{cases} 0.54 - 0.46 \cos\left(\dfrac{2\pi n}{N-1}\right) & \text{pro } 0 \leq n \leq N-1 \\ 0 & \text{jinak.} \end{cases} \tag{3.2}$$



Obrázek 2: Aplikace Hammingova okénka na funkci $f(x) = \sin(10\pi x)$.

3.2 Příznaky v časové a frekvenční oblasti

3.2.1 Krátkodobá energie signálu

Funkce krátkodobé energie signálu je definována vztahem

$$E_n = \sum_{k=-\infty}^{\infty} \left[s(k) w(n-k) \right]^2, \tag{3.3}$$

kde $s(k)$ je vzorek signálu v čase k a $w(n)$ je daný typ okénka. Hodnoty této funkce poskytují informaci o průměrné hodnotě energie v každém mikrosegmentu signálu. Hlavním nedostatkem funkce krátkodobé energie signálu je její vysoká citlivost na velké výkyvy v amplitudě signálu.

3.2.2 Krátkodobá intenzita signálu

Funkce krátkodobé intenzity signálu je definována vztahem

$$I_n = \sum_{k=-\infty}^{\infty} |s(k)| w(n-k). \tag{3.4}$$

Na rozdíl od funkce krátkodobé energie signálu není tato funkce citlivá na velké změny amplitudy signálu.

3.2.3 Krátkodobé průchody nulou

Funkce krátkodobých průchodů nulou je definována vztahem

$$Z_n = \frac{1}{2} \sum_{k=-\infty}^{\infty} |sign\left[s(k)\right] - sign\left[s(k-1)\right]| w(n-k). \tag{3.5}$$

Funkce krátkodobých průchodů nulou nese informaci o frekvenci signálu - čím více průchodů nulou, tím vyšší je v daném úseku frekvence signálu a naopak. Frekvenci průchodů signálu nulovou úrovní můžeme tedy využít jako jednoduchou charakteristiku, která popisuje spektrální vlastnosti signálu. Hodnoty této funkce se nejčastěji využívají k detekci řeči v akustickém signálu (určení začátku a konce promluvy) [4, 9].

3.2.4 Mel-frekvenční kepstrální koeficienty

Nejčastěji používaným typem příznaků v rozpoznávání řeči jsou mel-frekvenční kepstrální koeficienty (MFCC). Jedná se o velice robustní typ příznaků ve frekvenční oblasti, který respektuje nelineární vlastnosti vnímání zvuků lidským uchem. Lineární frekvence f[Hz] je převedena na frekvenci f_m[mel] v nelineární melovské frekvenční škále

$$f_m = 2595 \log_{10}\left(\frac{f}{700}\right). \tag{3.6}$$

Při výpočtu MFCC se nejčastěji volí segmentace signálu na mikrosegmenty o délce 10 až 30ms, na které se aplikuje Hammingovo okénko s posunem o 10ms. Segmentovaný signál je následovně zpracován rychlou Fourierovou transformací (FFT), díky které získáme amplitudové spektrum $|S(f)|$ analyzovaného signálu. Vzhledem k použití FFT je vhodné volit počet vzorků okénka při dané frekvenci roven mocnině 2.

Dalším krokem výpočtu je melovská filtrace, při níž využijeme banku trojúhelníkových pásmových filtrů. Jednotlivé trojúhelníkové filtry jsou rozloženy přes celé frekvenční pásmo on nuly až do Nyquistovy frekvence a jejich střední frekvence jsou rovnoměrně rozloženy podél frekvenční osy v melovské škále. Střední hodnoty filtrů b_m lze vyjádřit vztahem

$$b_{m,i} = b_{m,i-1} + \Delta_m, \tag{3.7}$$

kde $i = 1, 2, \ldots, M^*$ je počet trojúhelníkových filtrů v bance, $b_{m,0} = 0$ mel a $\Delta_m = B_{m,w}/(M^* + 1)$.

Dále je třeba vypočítat odezvy všech filtrů, čehož dosáhneme jejich vyjádřením ve frekvenční škále s měřítkem v herzích za využití původních koeficientů získaných FFT. Přepočteme tedy všechny střední frekvence v melovské škále $b_{m,i}, i = 1, \ldots, M^*+1$ pomocí inverzního vztahu $f = 700[exp(0.887 \cdot 10^{-3} f_m) - 1]$ na střední frekvence $b_i, i = 1, \ldots, M^*+1$. Nyní můžeme trojúhelníkové filtry vyjádřit vztahem

$$u(f,i) = \begin{cases} \dfrac{1}{b_i - b_{i-1}}(f - b_{i-1}) & \text{pro } b_{i-1} \leq f < b_i \\ \dfrac{1}{b_i - b_{i+1}}(f - b_{i+1}) & \text{pro } b_i \leq f < b_{i+1} \\ 0 & \text{jinak} \end{cases} \tag{3.8}$$

a odezvy jednotlivých filtrů $y_m(i)$ vztahem

$$y_m(i) = \sum_{f=b_{i-1}}^{b_{i+1}} |S(f)| u(f,i), \tag{3.9}$$

kde $i = 1, 2, \ldots, M^*$ a f jsou frekvence využité při výpočtu FFT. Při průchodu signálu filtrem je každý koeficient FFT násoben odpovídajícím ziskem filtru a výsledky pro jednotlivé filtry jsou akumulovány a následně zlogaritmovány. Tím dojde k dekorelaci energií filtrů a získané hodnoty lze použít jako plnohodnotné příznaky.

Nakonec provedeme zpětnou diskrétní kosinovou transformaci (DCT). Značnou výhodou DCT je vysoká nekorelovanost vzniklých koeficientů. Koeficienty jsou dány vztahem

$$c_m(j) = \sum_{i=1}^{M^*} \log y_m(i) \cos\left(\frac{\pi j}{M^*}(i - 0.5)\right) \quad \text{pro } j = 0, 1, \ldots, M, \tag{3.10}$$

kde M je počet mel-frekvenčních kepstrálních koeficientů [9, 10].

3.3 Delta a delta-delta koeficienty

Při výpočtu příznakových vektorů předpokládáme, že signál je na krátkém úseku stacionární. Je tedy zřejmé, že tyto příznakové vektory budou popisovat pouze statické vlastnosti signálu a dynamické vlastnosti se ztrácí. Tento nedostatek můžeme vyřešit rozšířením příznakových vektorů o dynamické koeficienty.

Využívají se delta (diferenční) koeficienty a delta-delta (akcelerační) koeficienty, které odpovídají první, respektive druhé derivaci příznakového vektoru. Delta koeficienty lze

definovat vztahem

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \tag{3.11}$$

kde d_t je delta koeficient v čase t a N je volitelný parametr. Tento parametr se většinou volí $N = 1$ nebo $N = 2$ a určuje, z kolika sousedních mikrosegmentů budou vypočítány delta koeficienty (jeden, resp. dva leví a praví sousedé mikrosegmentu). Delta-delta koeficienty lze spočítat využitím stejného vztahu s tím rozdílem, že se nepočítají ze statických koeficientů, ale z delta koeficientů [9, 10].

3.4 Normalizace příznaků

Jednotlivé příznaky se mohou pohybovat na různých definičních oborech hodnot a při klasifikačních úlohách se mohou projevit s různou váhou. Proto je vhodné příznaky transformovat na stejný definiční obor hodnot, popř. do rozdělení o stejných parametrech, a tím zaručit, že všechny příznaky budou mít stejný vliv. Tuto transformaci nazýváme normalizací dat a v praxi se nejčastěji používají následující dvě metody:

- **Z-score normalizace** - tato metoda se používá v případě, kdy data odpovídají normálnímu rozložení. Z-score normalizací přetransformujeme vstupní data x, na data z, které odpovídají normálnímu rozdělení o střední hodnotě $\mu = 0$ s rozptylem $\sigma = 1$.

$$z = \frac{x - \mu}{\sigma} \tag{3.12}$$

- **Min-Max normalizace** - Min-Max normalizace přeškáluje data do definovaného intervalu (nejčastěji $[0, 1]$ nebo $[-1, 1]$, popř. $[0, 255]$ při zpracování obrazu). Tím dosáhneme menšího rozptylu a eliminujeme vliv odlehlých bodů (tzv. outliers).

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.13}$$

Některé klasifikační algoritmy přímo vyžadují, aby data byla normalizována. Příkladem mohou být algoritmy založené na gradientních metodách (logistická regrese, SVM, neuronové sítě, atd.), kdy jsou změny parametrů modelu při trénování přímo závislé na vstupním příznakovém vektoru. Při velkých rozdílech definičních oborů jednotlivých příznaků se tedy některé parametry mohou měnit výrazně rychleji než ostatní. Normalizací dat zajistíme rychlejší a stabilnější konvergenci parametrů. Dalším příkladem mohou být shlu-

kovací algoritmy pracující s Euklidovou vzdáleností, kde je vhodné, aby všechny příznaky měly na shlukování stejný vliv [11].

4 Support Vector Machine

Předpokládejme, že máme trénovací množinu $\{x_i, y_i\}$, $i = 1, 2, \ldots, l$, $y_i \in \{-1, 1\}$, která je složena z příznakových vektorů x_i a odpovídajících cílových tříd y_i (binární klasifikace). Pro tuto množinu si odvodíme lineární klasifikátor založený na podpůrných vektorech pro separabilní a neseparabilní případ a nelineární klasifikátor. Nakonec si uvedeme metody, jak lze klasifikovat do více tříd [14, 15].

4.1 Nadrovina

Pro odvození klasifikátoru SVM (a později i pro odvození neuronové sítě) je vhodné nejprve zavést pojem nadrovina. Předpokládejme tedy, že máme p-dimenzionální prostor. Nadrovinou pak rozumíme libovolný afinní podprostor o dimenzi $p - 1$. Například ve dvoudimenzionálním prostoru je nadrovinou jednodimenzionální podprostor - přímka, ve třídimenzionálním prostoru je nadrovinou plocha. Nadrovinu o dimenzi p lze definovat následovně

$$b + w_1 x_1 + w_2 x_2 + \ldots w_p x_p = 0, \tag{4.1}$$

kde $x = (x_1, x_2, \ldots, x_p)^T$ je bod v p-dimenzionálním prostoru vyhovující rovnici. Bod x tedy leží na nadrovině.

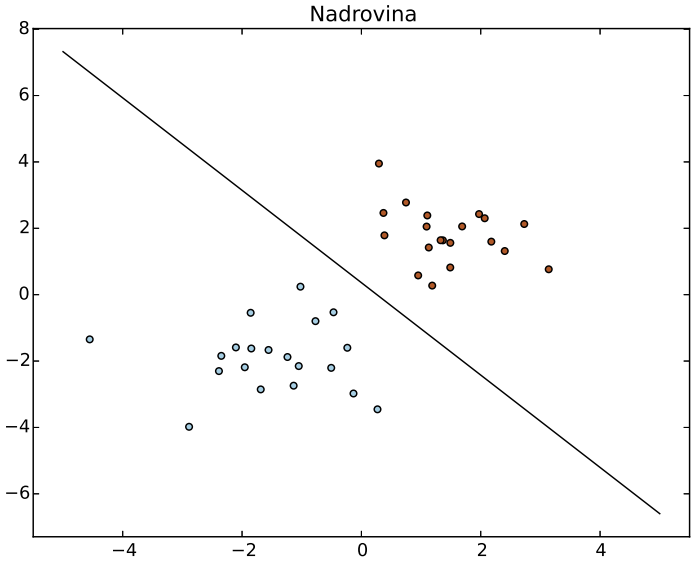
V případě, že bod x nevyhovuje rovnici 4.1, tedy

$$b + w_1 x_1 + w_2 x_2 + \ldots w_p x_p > 0, \tag{4.2}$$

bod leží na jedné straně poloroviny, resp. pokud

$$b + w_1 x_1 + w_2 x_2 + \ldots w_p x_p < 0, \tag{4.3}$$

bod leží na druhé straně poloroviny. Nadrovina tedy dělí p-dimenzionální prostor na dvě poloviny, tzv. poloprostory. Náležitost bodu do těchto poloprostorů pak můžeme zjistit jednoduchým výpočtem znaménka levé strany rovnice 4.1 [12].



Obrázek 3: Nadrovina oddělující body ve dvoudimenzionálním prostoru [13].

4.2 Lineárně separabilní případ

Předpokládejme, že existuje nadrovina, která oddělí body trénovací množiny jednotlivých tříd od sebe (oddělující nadrovina). Body x, které leží na této rovině splňují rovnici $w \cdot x + b = 0$, kde w je normála nadroviny. Dále si definujme kolmou vzdálenost oddělující nadroviny k počátku jako $\frac{|b|}{||w||}$, kde $||w||$ je Euklidovská norma w.

Dále si zaveďme nejkratší vzdálenost d_+ (d_-) mezi oddělující rovinou a pozitivním (negativním) příkladem a tzv. margin (odstup) jako $d_+ + d_-$. Pro lineárně separabilní případ hledá klasifikátor takovou nadrovinu, pro kterou nabývá margin nejvyšší hodnoty. Všechny body trénovací množiny tedy splňují tyto podmínky

$$x_i \cdot w + b \geq +1 \quad \text{pro } y_i = +1, \tag{4.4}$$

$$x_i \cdot w + b \leq -1 \quad \text{pro } y_i = -1, \tag{4.5}$$

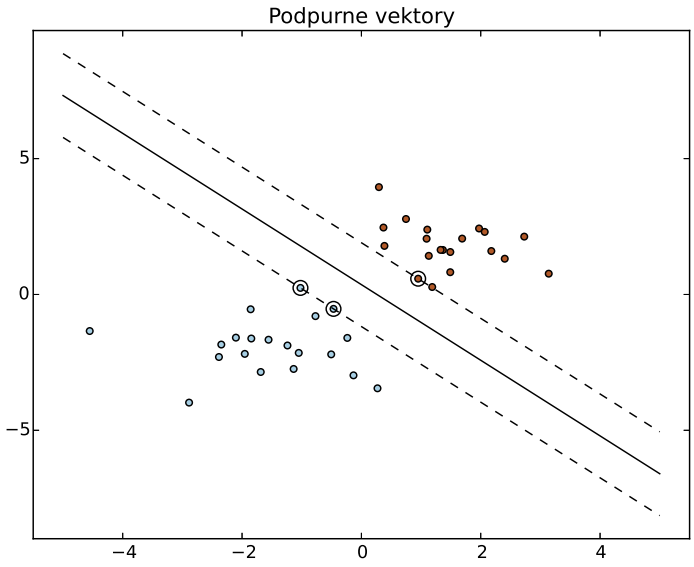
které lze sloučit do jedné množiny nerovností

$$y_i(x_i \cdot w + b) - 1 \geq 0 \quad \forall i. \tag{4.6}$$

Body, pro které v nerovnici 4.4 platí rovnost, leží na nadrovině $H_1 : x_i \cdot w + b = +1$ a jejich kolmou vzdálenost vůči počátku lze vyjádřit jako $\frac{|1-b|}{||w||}$. Obdobně body, pro které

v nerovnici 4.5 platí rovnost, leží na nadrovině $H_2 : x_i \cdot w + b = -1$ v kolmé vzdálenosti od počátku $\frac{|-1-b|}{||w||}$. Z toho plyne, že $d_+ = d_- = \frac{1}{||w||}$ a margin je roven $\frac{2}{||w||}$. Nadroviny H_1 a H_2 mají stejnou normálu w (jsou rovnoběžné) a nenachází se mezi nimi žádný bod z trénovací množiny. Nalezení takových nadrovin, které maximalizují margin, provedeme minimalizací $||w||^2$ za respektování omezujících podmínek.

Body ležící na nadrovinách H_1 a H_2 se nazývají podpůrné vektory (support vectors, zvýrazněné body na obrázku 4) a oddělující rovina je na nich přímo závislá. Pokud dojde k jejich pohybu nebo odstranění, změní se i výsledné řešení.



Obrázek 4: Vizualizace podpůrných vektorů [13].

Formulujme si nyní tento problém pomocí Lagrangeových multiplikátorů α_i, $i = 1, 2, \ldots, l$ - jeden pro každou nerovnost 4.6, neboli jeden pro každý prvek trénovací množiny $\{x_i, y_i\}$. Dostáváme Lagrangeovu funkci

$$L_P \equiv \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i y_i (x_i \cdot w + b) + \sum_{i=1}^{l} \alpha_i. \tag{4.7}$$

Nyní stačí minimalizovat L_P podle w a b a zároveň požadujeme, aby $\frac{\partial L_P}{\partial \alpha_i} = 0$, $\alpha_i \geq 0$ (označme si tuto množinu omezujících podmínek $\mathcal{C}_1$). Vzhledem k tomu, že se jedná o úlohu konvexního kvadratického programování, můžeme ekvivalentně řešit duální problém: maximalizace L_P za omezujících podmínek $\frac{\partial L_P}{\partial w} = 0$ a $\frac{\partial L_P}{\partial b} = 0$ a zároveň $\alpha_i \geq 0$ (označme tuto množinu omezujících podmínek jako $\mathcal{C}_2$). Tato duální formulace problému má tu

vlastnost, že maximum L_P za podmínek $\mathcal{C}_2$ nastává při stejných hodnotách w, b a α jako minimum L_P za podmínek $\mathcal{C}_1$.

Omezeními gradientu $\frac{\partial L_P}{\partial w} = 0$ a $\frac{\partial L_P}{\partial b} = 0$ získáme rovnice

$$w = \sum_i \alpha_i y_i x_i, \tag{4.8}$$

$$\sum_i \alpha_i y_i = 0. \tag{4.9}$$

Dosazením rovnic 4.8 a 4.9 do 4.7 dostaneme

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j. \tag{4.10}$$

Po vyřešení optimalizační úlohy (trénovací fáze) můžeme klasifikovat libovolný vektor x na základě toho, na které straně oddělující nadroviny leží. Třída vektoru x tedy bude $\operatorname{sgn}(w \cdot x + b)$ [14].

4.3 Lineárně neseparabilní případ

Pokud použijeme výše zmíněný algoritmus na lineárně neseparabilní data, nenalezneme žádné přijatelné řešení, jelikož duální Langrangeova funkce L_D poroste nade všechny meze. Chceme-li algoritmus rozšířit i pro neseparabilní data, musíme uvolnit podmínky 4.4 a 4.5. Toho dosáhneme zavedením volných (slack) proměnných ε_i, $i = 1, 2, \ldots, l$

$$x_i \cdot w + b \geq +1 - \varepsilon_i \quad \text{pro } y_i = +1 \tag{4.11}$$

$$x_i \cdot w + b \leq -1 + \varepsilon_i \quad \text{pro } y_i = -1 \tag{4.12}$$

$$\varepsilon_i \geq 0 \quad \forall i. \tag{4.13}$$

Chyba klasifikace tedy nastane v případě, že $\varepsilon > 1$. Horní mez počtu chyb klasifikace prvků trénovací množiny je tedy $\sum_i \varepsilon_i$. Vhodným způsobem, jak navýšit hodnotu kriteriální funkce za každou chybu, je minimalizovat $\frac{||w||^2}{2} + C(\sum_i \varepsilon_i)^k$ místo původního kritéria $\frac{||w||^2}{2}$. Parametr C se volí a odpovídá penalizaci za chybnou klasifikaci - čím vyšší hodnota C, tím větší penalizace. Pro $k > 0$, $k \in \mathbb{Z}$ se jedná o úlohu konvexního programování, pro $k = 2$ a $k = 1$ se jedná přímo o úlohu kvadratického programování. Volbou $k = 1$ navíc

zajistíme, že ε_i ani jejich multiplikátory se neobjeví v definici duálního problému

$$L_D \equiv \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \tag{4.14}$$

za podmínek

$$0 \leq \alpha_i \leq C, \tag{4.15}$$

$$\sum_i \alpha_i y_i = 0. \tag{4.16}$$

Řešením je pak

$$w = \sum_{i=1}^{N_S} \alpha_i y_i x_i, \tag{4.17}$$

kde N_S je počet podpůrných vektorů.

Primární úloha je dána Lagrangeovou funkcí

$$L_P = \frac{1}{2}||w||^2 + C \sum_i \varepsilon_i - \sum_i \alpha_i \{y_i(x_i \cdot w + b) - 1 + \varepsilon_i\} - \sum_i \mu_i \epsilon_i, \tag{4.18}$$

kde μ_i jsou Lagrangeovské multiplikátory zajištující kladnost ε_i. Řešitelnost 4.18 je dána Karush-Kuhn-Tucker podmínkami (více v [14]). Pro výpočet prahu b postačí pouze dvě z Karush-Kuhn-Tucker podmínek

$$\alpha_i \{y_i(x_i \cdot w + b) - 1 + \varepsilon_i\} = 0, \tag{4.19}$$

$$\mu_i \epsilon_i = 0. \tag{4.20}$$

Ačkoliv k výpočtu prahu b stačí znát pouze jeden prvek trénovací množiny splňující podmínku $0 < \alpha_i < C$ a zároveň $\varepsilon_i = 0$, z numerického hlediska je rozumnější určit výsledný práh jako průměr prahů přes všechny prvky trénovací množiny [14].

4.4 Nelineárně separabilní případ

Uvažujme nyní případ, kdy oddělující nadrovina není lineární funkcí trénovacích dat - potřebujeme tedy zobecnit rovnice 4.14, 4.15 a 4.16 pro nelineární oddělující nadrovinu. Všimněme si, že v těchto rovnicích se data trénovací množiny vyskytují vždy jako skalární součin $x_i \cdot x_j$. Předpokládejme, že existuje zobrazení Φ z n-dimenzionálního prostoru do

jiného Euklidovského prostoru $\mathcal{H}$ o vyšší dimenzi (až nekonečnědimenzionálního)

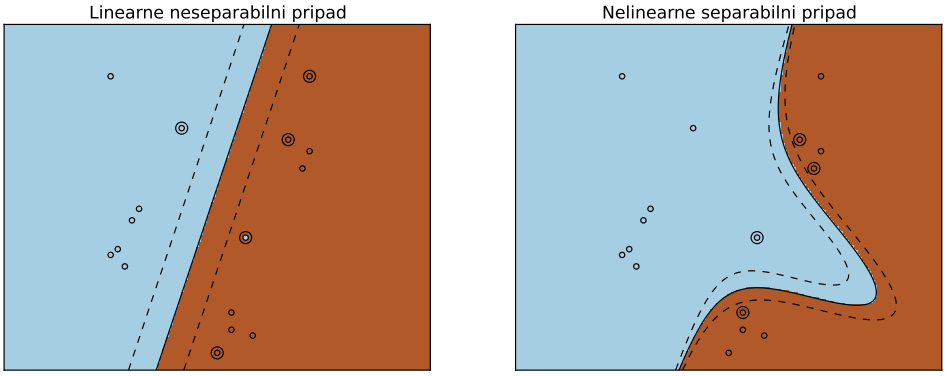
$$\Phi : \mathbb{R}^n \mapsto \mathcal{H}. \tag{4.21}$$

Potom by trénovací algoritmus závisel pouze na skalárních součinech $\Phi(x_i) \cdot \Phi(x_j)$ v prostoru $\mathcal{H}$. Definujme si tedy jádrovou funkci (kernel function) $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, kterou můžeme nahradit skalární součin $x_i \cdot x_j$ v trénovacím algoritmu 4.14 a opět dopočítat normálu w oddělující nadroviny. Nemusíme tedy explicitně znát zobrazení Φ, pouze jádrovou funkci.

Klasifikaci libovolného bodu x provedeme výpočtem znaménka funkce $f(x)$

$$f(x) = \sum_{i=1}^{N_S} \alpha_i y_i \Phi(s_i) \cdot \Phi(x) + b = \sum_{i=1}^{N_S} \alpha_i y_i K(s_i, x) + b, \tag{4.22}$$

kde s_i jsou podpůrné vektory [14, 16].



Obrázek 5: SVM s lineární a polynomiální jádrovou funkcí [13].

4.5 Klasifikace do více tříd

Zatím jsme se zabývali pouze klasifikaci do dvou tříd neboli binární klasifikací. Nyní uvažujme trénovací množinu $\{x_i, y_i\}$, $i = 1, 2, \ldots, l$, $y_i \in \{1, 2, \ldots, k\}$, kde $k > 2$ je počet cílových tříd. Uveďme si dva základní přístupy, jak do těchto tříd klasifikovat:

- **One-Versus-One** - natrénujeme $\frac{k(k-1)}{2}$ binárních klasifikátorů, kde každý z nich porovnává dvě různé třídy navzájem. Klasifikace testovacího vektoru je pak založena na hlasování, kdy každý klasifikátor hlasuje pro jednu třídu a testovací vektor je zařazen do té třídy, která dostala nejvíce hlasů.

- **One-Versus-All** - natrénujeme k binárních klasifikátorů, kde každý z nich porovnává jednu z k tříd oproti zbylým $k-1$ třídám. Testovací vektor je zaklasifikován do té třídy, pro kterou nabývá oddělující nadrovina 4.22 nejvyšší hodnoty (tj. bod leží nejdále od oddělující nadroviny a byl zaklasifikován s největší "jistotou") [15,16].

5 Neuronové sítě

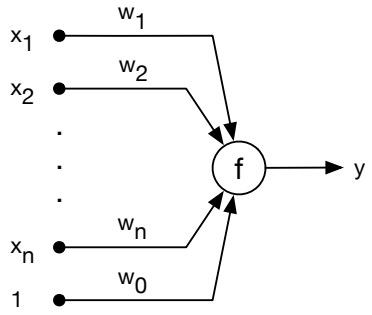
Neuronová síť je algoritmus inspirovaný funkcí neuronů a jejich propojením v lidském mozku. Skládá se z mnoha výpočetních jednotek (neurony) propojených pomocí numerických parametrů (synapse) a úpravou těchto parametrů je schopna se učit. Tato práce je omezena pouze na dopředné neuronové sítě a jejich trénování pomocí učení s učitelem.

5.1 Perceptron a aktivační funkce

Základní výpočetní jednotkou neuronových sítí je tzv. perceptron. Výstupní hodnota perceptronu je dána vztahem

$$y = f\left(\sum_{i=1}^{n} w_i x_i + w_0\right) = f(w^T x + w_0), \tag{5.1}$$

kde $w = [w_1, w_2, \ldots, w_n]^T$ je váhový vektor, $x = [x_1, x_2, \ldots, x_n]^T$ je vstupní vektor, w_0 je práh a $f(\cdot)$ je aktivační funkce. Práh reprezentuje váhu vedoucí od jednotkového vstupního bodu, který se zavádí z důvodu generalizace sítě.



Obrázek 6: Perceptron.

Pro zjednodušení následujících odvození si rozšiřme váhový vektor w o práh w_0 a vstupní vektor o jednotkový vstupní bod

$$w = [w_0, w_1, \ldots, w_n]^T, \tag{5.2}$$

$$x = [1, x_1, \ldots, x_n]^T. \tag{5.3}$$

Uveďme si také příklady nejznámějších aktivačních funkcí:

- **Sigmoidální aktivační funkce** - sigmoidální aktivační funkce transformuje reálné

číslo do intervalu $(0, 1)$. Nevýhodou tohoto rozsahu je, že velmi malá čísla jsou transformována na hodnoty blízké nule, což má za následek i velice nízkou hodnotu lokálního gradientu a neuronem tak projde minimum signálu (více u algoritmu back-propagation). Při inicializaci vah vysokými hodnotami naopak může dojít k saturaci signálu a síť nebude schopná se učit. Výstupy neuronu s touto aktivační funkcí zároveň nemají střední hodnotu v nule, což má za následek, že pro kladný vstup budou mít všechny váhy vedoucí k neuronu stejné znaménko.

$$f(\xi) = \frac{1}{1 + e^{-\xi}} \tag{5.4}$$

- **Tanh** - aktivační funkce tanh transformuje reálně číslo do intervalu $(-1, 1)$. Výstupní interval má střední hodnotu v nule a řeší nedostatky sigmoidální aktivační funkce.

$$f(\xi) = \tanh(\xi) = \frac{2}{1 + e^{-2\xi}} - 1 \tag{5.5}$$

- **ReLU (Rectified Linear Unit)** - aktivační funkce ReLU je lineární aktivační funkce s prahem v hodnotě nula. Oproti výše zmíněným aktivačním funkcím výrazně urychluje konvergenci gradientu a není tak výpočetně náročná. Při nevhodně zvolené konstantě učení však muže dojít k "deaktivaci"neuronů, kdy jejich gradient poklesne na nulu a tyto neurony již nikdy nebudou aktivovány.

$$f(\xi) = \max(0, \xi) \tag{5.6}$$

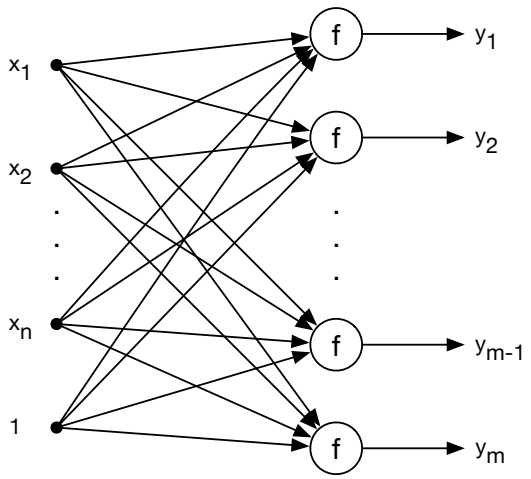
- **Maxout** - Maxout je generalizací aktivační funkce ReLU. Na rozdíl od výše zmíněných aktivačních funkcí nemá stejný funkcionální tvar $f(w^T x + b)$, ale počítá hodnotu funkce $\max(w_1^T x + b_1, w_2^T x + b_2)$. Maxout řeší nedostatky ReLU, ovšem za cenu zdvojnásobení počtu parametrů pro každý neuron [17].

Můžeme si všimnout, že argument aktivační funkce definuje nadrovinu v n-dimenzionálním prostoru. Volbou aktivační funkce

$$f(\xi) = \text{sign}(\xi) = \begin{cases} 1 & \text{pro } \xi \geq 0 \\ 0 & \text{jinak} \end{cases} \tag{5.7}$$

získáme jednoduché rozhodovací pravidlo, které přiřazuje body do poloroviny na základě znaménka argumentu aktivační funkce. Jedná se tedy o jednoduchý klasifikátor, který dokáže klasifikovat do dvou tříd (jedná se o analogii lineárně separabilního případu u klasifikátoru SVM).

Paralelním spojením více perceptronů získáme nejjednodušší typ dopředné neuronové sítě, tzv. jednovrstvou neuronovou síť [5, 18].



Obrázek 7: Jednovrstvá neuronová síť s n vstupy, aktivační funkcí $f(\cdot)$ a m výstupy.

5.2 Trénování jednovrstvé neuronové sítě

Jedním ze způsobů trénování jednovrstvé neuronové sítě je tzv. perceptronové pravidlo. Nejprve síti přidělíme náhodné váhy (většinou se volí malá čísla v okolí nuly) a poté přivádíme na vstup sítě jednotlivé příznakové vektory z trénovací množiny. V případě, že síť zaklasifikuje bod chybně, dojde k úpravě hodnot vah. Tento proces probíhá tak dlouho, dokud nejsou všechny body zaklasifikovány správně.

Jednotlivé váhy w_i vedoucí od i-tého vstupu x_i jsou modifikovány pomocí perceptronového pravidla

$$w_i(k+1) = w_i(k) + \Delta w_i(k) = w_i(k) + \alpha(t-y)x_i, \tag{5.8}$$

kde k je iterace učícího algoritmu, t je očekávaný výstup neuronu, y je skutečný výstup neuronu a $\alpha \in \mathbb{R}^+$ je konstanta učení. V případě, že je vhodně zvolena konstanta učení α a data jsou lineárně separabilní, algoritmus učení dokonverguje k optimálnímu nastavení

vah.

Druhým způsobem učení je tzv. delta pravidlo, které zajišťuje konvergenci i pro lineárně neseparabilní data. Pro zavedení delta pravidla si nejprve definujme trénovací chybu

$$E(w) = \frac{1}{2}\sum_{i=0}^{n}(t_i - y_i)^2, \tag{5.9}$$

kde n je počet prvků trénovací množiny.

Cílem učení je tuto chybu minimalizovat, čehož dosáhneme pomocí gradientního sestupu (gradient descent). Jedná se o iterativní algoritmus, který opět začíná s náhodně inicializovanými hodnotami vah a v každém kroku je upraví ve směru největšího sestupu gradientu. Tento proces probíhá tak dlouho, dokud není nalezeno globální minimum chybové funkce. Modifikace vah je tedy závislá na gradientu $E(w)$ podle w

$$\nabla E = \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n}\right]. \tag{5.10}$$

Gradient $\nabla E(w)$ udává směr největšího růstu chybové funkce - jeho zápornou hodnotou tedy získáme směr největšího poklesu a trénovací pravidlo v maticovém tvaru bude

$$w(k+1) = w(k) + \Delta w(k) = w(k) - \alpha \nabla E(w) = w(k) - \alpha(t - y)x^T. \tag{5.11}$$

Vzhledem k tomu, že daná chybová funkce má pouze jedno globální minimum, gradientní sestup při vhodně zvolené konstantě učení vždy dokonverguje k takovému váhovému vektoru, který zajišťuje minimální chybu [18].

5.3 Vícevrstvá dopředná neuronová síť

Jak již bylo zmíněno, jednovrstvé neuronové sítě umožňují vyjádřit pouze lineární rozhodovací hranici. Nelineární hranici můžeme vyjádřit pomocí vícevrstvé sítě, která se skládá z jedné vstupní vrstvy, jedné nebo více skrytých vrstev a výstupní vrstvy. Každá skrytá vrstva se skládá z libovolného počtu neuronů a prahové jednotky a jednotlivé skryté vrstvy mohou mít různé aktivační funkce.

Pro zjednodušení následujících odvození předpokládejme síť se vstupní vrstvou s I vstupními jednotkami, skrytou vrstvu s J neurony a výstupní vrstvu s K výstupními jednotkami. Neurony ve skryté vrstvě mají aktivační funkci $f(\cdot)$ a výstupní vrstva má ak-

tivační funkci $g(\cdot)$ (znázorněno na obrázku 8). Opět rozšíříme váhovou matici w o práh w_0 vedoucí k jednotkovému vstupnímu bodu x_0, kterým rozšíříme vstupní vektor x. Výstup neuronu j ve skryté vrstvě tedy můžeme zapsat jako

$$net_j = \sum_{i=1}^{I} x_i w_{ji} + x_0 w_{j0} = \sum_{i=0}^{I} x_i w_{ji} = w_j^T x, \tag{5.12}$$

$$z_j = f(net_j), \tag{5.13}$$

kde w_{ji} značí váhu mezi j-tým neuronem skryté vrstvy a i-tou vstupní jednotkou. Obdobně lze vypočítat výstup k-té výstupní jednotky

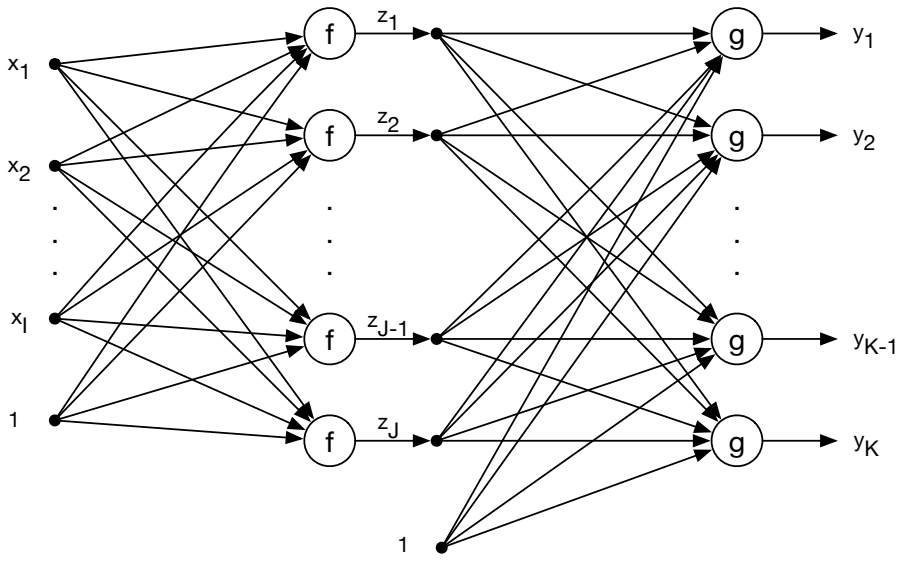
$$net_k = \sum_{j=1}^{J} z_j w_{kj} + z_0 w_{k0} = \sum_{j=0}^{J} z_j w_{kj} = w_k^T z, \tag{5.14}$$

$$y_k = g\left(net_k\right). \tag{5.15}$$

Úpravou lze k-tý výstup zapsat jako

$$y_k = g\left(\sum_{j=1}^{J} w_{kj} f\left(\sum_{i=1}^{I} x_i w_{ji} + x_0 w_{j0}\right) + w_{k0}\right). \tag{5.16}$$

Obdobným způsobem lze vyjádřit výstup pro neuronovou síť s libovolným počtem skrytých vrstev.



Obrázek 8: Dvouvrstvá neuronová síť.

5.4 Trénování vícevrstvé neuronové sítě

Pro trénování vícevrstvých neuronových sítí se využívá algoritmus backpropagation neboli algoritmus zpětného šíření chyby, který je stejně jako delta pravidlo založen na minimalizaci chybové funkce pomocí gradientu. Definujme si tedy trénovací chybu E jako kvadrát sumy rozdílů mezi skutečným výstupem k-té výstupní jednotky y_k a očekávaným výstupem t_k

$$E(w) = \frac{1}{2}\sum_{k=1}^{K}(t_k - y_t)^2 = \frac{1}{2}(t-y)^2. \tag{5.17}$$

Algoritmus backpropagation, stejně jako delta pravidlo, vychází z algoritmu gradientního sestupu. Jednotlivé váhy jsou inicializovány náhodnými malými čísly a jsou modifikovány ve směru největšího poklesu chybové funkce

$$\Delta w = -\alpha\frac{\partial E}{\partial w}, \tag{5.18}$$

čímž opět získáme pravidlo pro modifikaci vah

$$w(k+1) = w(k) + \Delta w(k). \tag{5.19}$$

Nejprve si odvoďme pravidlo pro modifikaci vah mezi skrytou a výstupní vrstvou. Vzhledem k tomu, že chybová funkce není přímo závislá na w_{jk}, musíme použít řetězové pravidlo.

$$\frac{\partial E}{\partial w_{jk}} = \frac{\partial E}{\partial net_k}\frac{\partial net_k}{\partial w_{jk}} = \delta_k\frac{\partial net_k}{\partial w_{jk}} \quad \Rightarrow \quad \delta_k = -\frac{\partial E}{\partial net_k} \tag{5.20}$$

Hodnotu δ_k nazýváme citlivost neuronu a popisuje, jak se změní celková chyba při jeho aktivaci. Derivací rovnice 5.17 získáme hledané pravidlo

$$\Delta w_{jk} = \alpha\delta_k z_j = \alpha(t_k - y_k)f'(net_k)z_j. \tag{5.21}$$

Obdobným způsobem vyjádříme pravidlo pro modifikaci vah mezi vstupní a skrytou vrstvou

$$\frac{\partial E}{\partial w_{ji}} = \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ji}} \quad \Rightarrow \quad \delta_j = f'(net_j)\sum_{k=1}^{K} w_{kj}\delta_k, \tag{5.22}$$

$$\Delta w_{ji} = \alpha x_i\delta_j = \alpha x_i f'(net_j\sum_{k=1}^{K} w_{kj}\delta_k). \tag{5.23}$$

Podrobnější odvození algoritmu backpropagation lze nalézt v [6, 18].

5.4.1 Momentum

Chybová funkce vícevrstvé sítě může mít více lokálních minim a na rozdíl od chybové funkce jednovrstvé sítě (parabolická funkce s jedním globálním minimem) gradientní sestup nezaručuje nalezení globálního minima, ale pouze lokálního minima. Z tohoto důvodu se zavádí tzv. momentum (setrvačnost), které zabraňuje uváznutí učícího algoritmu v mělkém lokálním minimu a urychluje konvergenci na plochých částech povrchu chybové funkce. Přidáním momentového členu získáme následující pravidlo pro úpravu vah

$$\Delta w_{ji}(k) = \alpha x_i \delta_j + \mu \Delta w_{ji}(k-1), \tag{5.24}$$

kde $0 < \mu < 1$ je momentum a k je iterace učícího algoritmu. Změna vah tedy závisí i na změně vah v minulé iteraci [6, 18].

5.4.2 Sekvenční a dávkové učení

Jak již bylo zmíněno, při trénování s učitelem jsou neuronové síti předkládány obrazy z trénovací množiny, která se pak snaží minimalizovat celkovou chybu mezi výstupy sítě a očekávanými hodnotami. V každém trénovacím cyklu algoritmu jsou síti předloženy všechny obrazy z trénovací množiny. V praxi se nejčastěji používají dva typy trénování neuronových sítí:

- **sekvenční učení** - neuronové síti jsou postupně předkládány jednotlivé obrazy z trénovací množiny (většinou v náhodném pořadí), pro každý obraz je spočtena chyba klasifikace a následně jsou upraveny parametry sítě.
- **dávkové učení** - neuronové síti jsou postupně předkládány jednotlivé obrazy z trénovací množiny, jednotlivé chyby klasifikace jsou akumulovány a k úpravě parametrů sítě dojde až na konci trénovacího cyklu s ohledem na celkovou chybu klasifikace [6].

6 Dynamic Time Warping

Rozpoznávání řeči je velmi obtížná úloha, jelikož žádné dvě promluvy nejsou stejné. Hlasy různých osob se liší a stejně tak se liší i jejich artikulace nebo tempo a barva řeči. Ani promluvy jedné osoby nejsou stejné - jedna promluva může být pronesena potichu, druhá nahlas nebo šeptem, mohou být proneseny různě rychle nebo např. pod vlivem nachlazení. Na akustickém signálu se dále projevuje přítomnost šumu a rušení na pozadí [9].

Jedním z řešení tohoto problému je využití klasifikátoru podle minimální vzdálenosti k vzorovým obrazům. Při klasifikaci se slovo zpracovává jako celek a je zařazeno do té třídy, k jejímuž vzorovému obrazu má nejmenší vzdálenost. Základním problémem je určení této vzdálenosti, jelikož obrazy mají různé délky v závislosti na délce signálu. Odlišnosti mezi podobnými signály tedy nejsou ve spektrální oblasti, ale v časové oblasti. K určení vzdálenosti mezi dvěma signály se tedy využívá algoritmus Dynamic Time Warping (DTW), neboli nelineární "borcení"časové osy, který je založen na metodě dynamického programování. "Borcením"časové osy obrazu jedné z nahrávek dojde k maximalizaci shody mezi nahrávkami.

6.1 Základní algoritmus

Předpokládejme, že máme dvě nahrávky, které jsou reprezentovány svými obrazy. Označme obraz testovaného slova

$$A = \{a_1, a_2, \ldots, a_n, \ldots, a_I\} \tag{6.1}$$

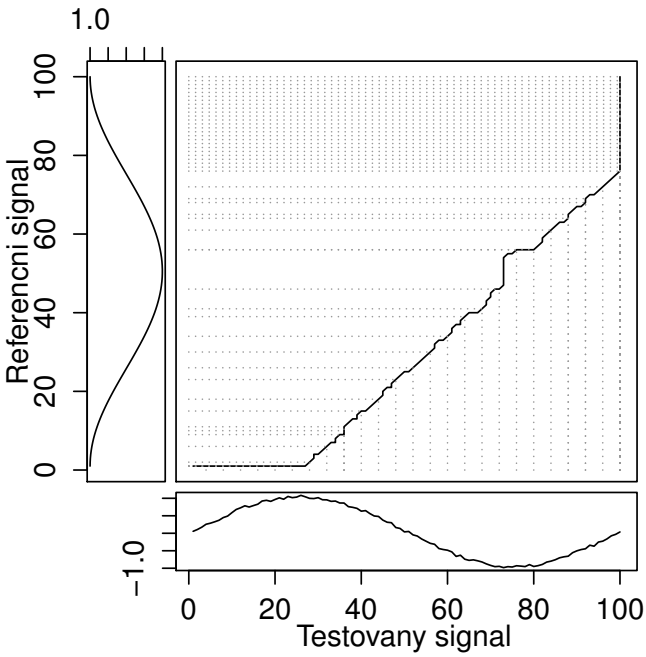
a obraz referenčního slova

$$B = \{b_1, b_2, \ldots, b_m, \ldots, b_J\}, \tag{6.2}$$

kde a_n je n-tý příznak testovaného slova a b_m je m-tý příznak referenčního slova. Algoritmus DTW pak hledá v rovině (n,m) optimální cestu $m = \Psi(n)$, která minimalizuje vzdálenost mezi obrazy A a B

$$D(A, B) = \sum_{n=1}^{I} \hat{d}\left(a_n, b_{\Psi(n)}\right), \tag{6.3}$$

kde $\hat{d}(a_n, b_{\Psi(n)})$ je vzdálenost mezi n-tým prvkem testovaného obrazu a m-tým prvkem referenčního obrazu.



Obrázek 9: Průběh funkce DTW pro $sin(x)$ se šumem a $cos(x)$.

6.2 Omezení pohybu funkce

Optimální cesta by měla zachovávat základní vlastnosti časových os obou obrazů (souvislost, monotónnost, atd.). Z toho důvodu se zavádí omezení na pohyb funkce DTW. Zavedeme obecnou časovou proměnnou k a časové proměnné m a n vyjádříme jako funkce k

$$n = i(k), \tag{6.4}$$

$$m = j(k), \tag{6.5}$$

kde $k = 1, 2, \ldots, K$ a K je délka obecné časové osy.

6.2.1 Omezení na hraniční body

Hraniční body funkce DTW jsou dány podmínkami

$$i(1) = 1 \quad i(K) = I, \tag{6.6}$$

$$j(1) = 1 \quad j(K) = J. \tag{6.7}$$

6.2.2 Omezení na lokální souvislost

Při průchodu funkce DTW může dojít k nadměrné expanzi či kompresi časové osy. Proto je vhodné omezit lokální monotónnost a souvislost DTW funkce

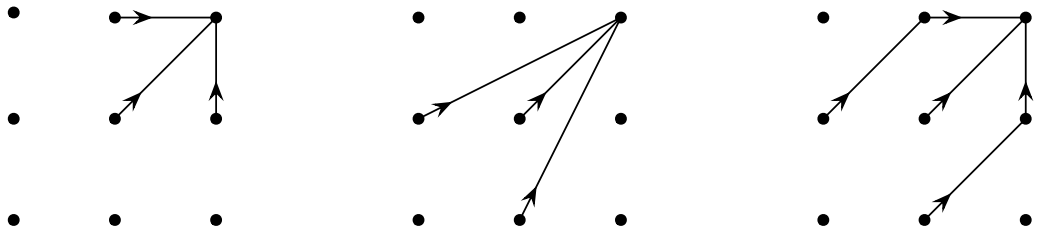
$$0 \leq i(k) - i(k-1) \leq \bar{I}, \tag{6.8}$$

$$0 \leq j(k) - j(k-1) \leq \bar{J}, \tag{6.9}$$

přičemž $\bar{I}$ a $\bar{J}$ jsou volitelné konstanty. Nejčastěji volíme hodnoty $\bar{I}, \bar{J} = 1, 2, 3$ s tím, že při hodnotě větší než 1 může funkce DTW při porovnávání některé mikrosegmenty přeskočit.

6.2.3 Omezení na lokální strmost

Pro funkci není vhodný příliš velký, ani příliš malý přírůstek, a tak se zavádí omezení na lokální strmost. Pokud se zastupující bod $[i(k), j(k)]$ pohybuje ve směru jedné osy $\bar{n}$-krát po sobě při rostoucím k, pak se v tomto směru nesmí nadále pohybovat, dokud neudělá $\bar{m}$ kroků v jiném směru.



Obrázek 10: Nejčastěji používaná lokální omezení (více v [4, 19]).

6.2.4 Globální vymezení oblasti pohybu funkce

Splněním podmínek pro omezení na hraniční body a zobecněním podmínek omezení na lokální strmost na celou rovinu (n, m) lze vymezit přípustnou oblast průchodu funkce DTW

$$1 + \alpha\left[i(k) - 1\right] \leq j(k) \leq 1 + \beta\left[i(k) - 1\right], \tag{6.10}$$

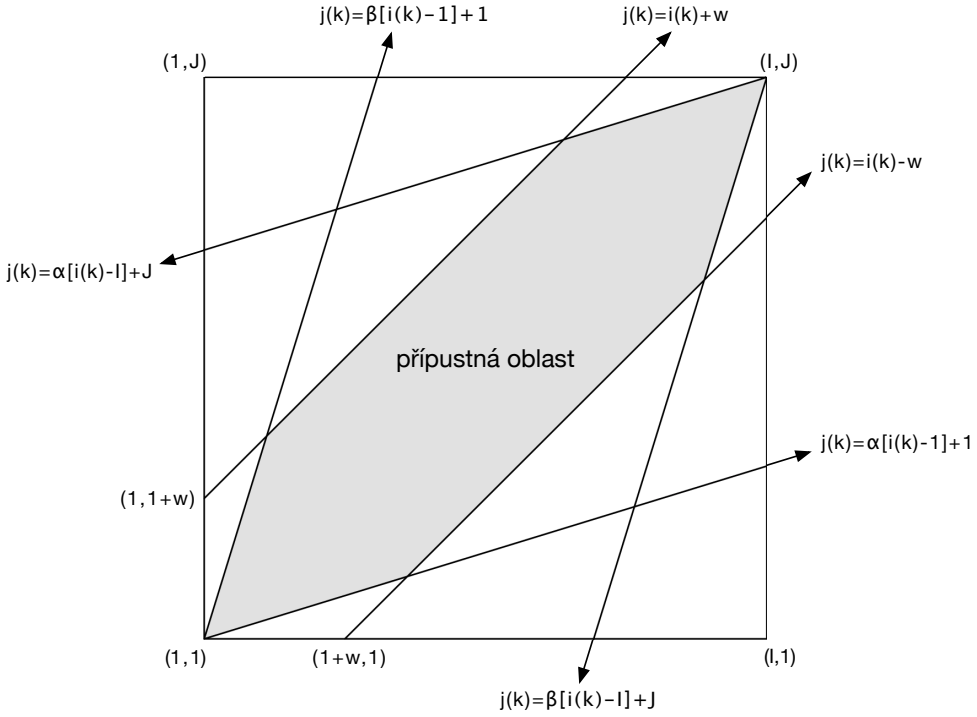
$$J + \beta\left[i(k) - I\right] \leq j(k) \leq J + \alpha\left[i(k) - I\right], \tag{6.11}$$

kde α je minimální směrnice a β maximální směrnice přímky vymezující přípustnou oblast.

Předpokládejme, že při porovnání testovaného a referenčního obrazu, které reprezentují stejné slovo, nemůže dojít k zásadním časovým rozdílům mezi příslušnými úseky stejných obrazů zapříčiněných kolísáním tempa řeči. S ohledem na tento předpoklad lze tedy stanovit podmínku pro druhé globální vymezení oblasti pohybu funkce DTW

$$|i(k) - j(k)| \leq w, \tag{6.12}$$

kde w je celé číslo, které určuje šířku okénka. Šířka okénka musí být menší než $|J - I|$, aby do přípustné oblasti funkce DTW bylo možné zahrnout i koncový bod (I, J).



Obrázek 11: Globální vymezení oblasti pohybu funkce.

6.3 Minimální vzdálenost

Celkovou minimální vzdálenost mezi testovacím obrazem A a referenčním obrazem B lze vyjádřit vztahem

$$D(A,B) = \min_{\{i(k),j(k),K\}} \left[\frac{\sum_{k=1}^{K} d\left[i(k),j(k)\right] \hat{W}(k)}{N(\hat{W})} \right], \tag{6.13}$$

kde $d[i(k),j(k)]$ je lokální vzdálenost mezi n-tým úsekem testovaného obrazu A a m-tým úsekem referenčního obrazu B, $\hat{W}(k)$ je hodnota váhové funkce pro k-tý úsek a $N(\hat{W})$ je normalizační faktor, jenž je funkcí váhové funkce. Váhová funkce je závislá pouze na lokální cestě funkce DTW.

Implementace váhové funkce se volí na základě zvolených lokálních omezení funkce DTW. Nejčastěji se využívá jeden z těchto čtyř typů váhových funkcí:

a) symetrická váhová funkce

$$\hat{W}(k) = \left[i(k) - i(k-1)\right] + \left[j(k) - j(k-1)\right], \tag{6.14}$$

b) asymetrická váhová funkce

b1)

$$\hat{W}(k) = i(k) - i(k-1), \tag{6.15}$$

b2)

$$\hat{W}(k) = j(k) - j(k-1), \tag{6.16}$$

c)

$$\hat{W}(k) = \min\left[i(k) - i(k-1), j(k) - j(k-1)\right], \tag{6.17}$$

d)

$$\hat{W}(k) = \max\left[i(k) - i(k-1), j(k) - j(k-1)\right], \tag{6.18}$$

přičemž $i(0) = j(0) = 0$.

6.4 Normalizační faktor

Normalizační faktor $N(\hat{W})$ kompenzuje počet kroků funkce DTW, který se pro různě dlouhé testovací a referenční sekvence může výrazně lišit. Lze jej definovat jako

$$N(\hat{W}) = \sum_{k=1}^{K} \hat{W}(k). \tag{6.19}$$

Dosazením vztahů 6.14, 6.15, 6.16 pro váhové funkce typu a) a b) získáme normalizační faktory

$$N(\hat{W}_a) = \sum_{k=1}^{K} [i(k) - i(k-1) + j(k) - j(k-1)] = \tag{6.20}$$

$$= i(K) - i(0) + j(K) - j(0) = I + J, \tag{6.21}$$

$$N(\hat{W}_{b1}) = \sum_{k=1}^{K} [i(k) - i(k-1)] = i(K) - i(0) = I, \tag{6.22}$$

$$N(\hat{W}_{b2}) = \sum_{k=1}^{K} [j(k) - j(k-1)] = j(K) - j(0) = J. \tag{6.23}$$

Ze vztahů je patrné, že pro váhové funkce typu a) a b) je hodnota normalizačního faktoru nezávislá na konkrétním průběhu funkce DTW. Pro váhové funkce typu c) a d) je hodnota normalizačního faktoru silně závislá na průběhu funkce DTW a nelze ji určit pomocí metod dynamického programování. Pro tyto případy se hodnota normalizačního faktoru volí nezávisle na průběhu funkce DTW, aby bylo možné použít rekurzivní algoritmus.

$$N(\hat{W}_c) = N(\hat{W}_d) = I \tag{6.24}$$

6.5 Rekurzivní algoritmus

Díky nezávislosti normalizačního faktoru na průběhu funkce DTW lze vztah 6.13 pro výpočet celkové minimální vzdálenosti mezi dvěma obrazy A a B zjednodušit do tvaru

$$D(A,B) = \frac{1}{N(\hat{W})} \{ \min_{\{i(k),j(k),K\}} \sum_{k=1}^{K} d\,[i(k), j(k)]\, \hat{W}(k) \} \tag{6.25}$$

Výslednou hodnotu vztahu 6.25 lze určit rekurzivně algoritmem dynamického programování, kdy zavedeme funkci g částečné akumulované vzdálenosti:

1. Inicializace

$$g\left[i(1), j(1)\right] = d\left[i(1), j(1)\right] \hat{W}(1) \tag{6.26}$$

2. Rekurze

$$g\left[i(k), j(k)\right] = \min_{\{i(k), j(k)\}} \{g\left[i(k-1), j(k-1)\right] + d\left[i(k), j(k)\right] \hat{W}(k)\} \tag{6.27}$$

3. Konečná normalizovaná vzdálenost

$$D(A, B) = \frac{1}{N(\hat{W})} g\left[i(K), j(K)\right] = \frac{1}{N(\hat{W})} g\left[I, J\right] \tag{6.28}$$

Rekurzivní vztahy pro různé typy lokálních omezení lze odvodit dosazením za $\hat{W}(k)$ [4,19].

7 Klasifikace izolovaných slov

Pro porovnání jednotlivých metod byla vytvořena datová sada obsahující 240 nahrávek od šesti řečníků. Mezi řečníky byli čtyři muži (v tabulkách s výsledky značeni čísly 1, 2, 3 a 6) a dvě ženy (značeny čísly 4, 5). Každý řečník pronesl čtyřikrát po sobě číslovky nula až devět. První dvě nahrávání proběhla v tichém prostředí, druhá dvě za mírného okolního šumu, díky čemuž lze lépe porovnat robustnost zkoumaných metod. Tato datová sada byla následně rozdělena na referenční a testovací sadu. Jako referenční nahrávky byly zvoleny číslovky nula až devět z prvního nahrávání každého řečníka, ostatní nahrávky tvoří testovací sadu.

Nahrávky z referenční sady pak byly využity jako trénovací data pro klasifikátor SVM a neuronovou síť. Pro příznaky generované neuronovou sítí byla trénovací sada rozšířena o nahrávky vytvořené při vývoji hlasového rozhraní pro Škoda Auto, kdy každý řečník třikrát pronesl povely hlasového ovládání navigace, rádia, telefonu a poté promluvy města, ulice a číslovek nula až devět.

Porovnávané metody byly implementovány v programovacím jazyce Python s využitím knihoven pro vědecké výpočty NumPy a SciPy [8], knihovny pro strojové učení scikit-learn [13] a knihovny pro neuronové sítě Lasagne [20].

7.1 Zpracování akustického signálu

Prvním krokem pro klasifikaci izolovaných slov je extrakce příznaků z akustického signálu. Pro tyto účely byl vytvořen samostatný modul obsahující metody pro zpracování akustického signálu a transformaci příznakových vektorů. Modul umožňuje segmentaci signálu s využitím pravoúhlého, Hammingova nebo Hanningova okénka s volitelnou délkou v milisekundách a volitelným posunem.

Z důvodu úspornějšího zápisu výsledků si zaveďme značení pro jednotlivé typy příznaků, které modul implementuje:

značení	typ příznaků
ste	krátkodobá energie
sti	krátkodobá intenzita
stzcr	krátkodobé průchody nulou
ste_sti_stzcr	kombinace příznaků krátkodobé energie, intenzity a průchodů nulou (tvoří matici, kde sloupce odpovídají jednotlivým typům příznaků v čase)
log_fb_en	logaritmované energie banky filtrů
mfcc	mel-frekvenční kepstrální koeficienty (implementace byla převzata z modulu [22])

Tabulka 1: Značení typů příznaků.

Po vygenerování příznakového vektoru je možné aplikovat Z-score nebo Min-Max normalizaci a dopočítat delta a delta-delta koeficienty (implementace byla převzata z modulu LibROSA [21]).

Vygenerované parametrizace příznaků jsou uvedeny v tabulce 2. První číslo v názvu příznaků značí délku okénka v milisekundách, druhé posun okénka v milisekundách. Využití Hammingova okénka je značeno zkratkou *ham* (pokud není uvedeno, příznak byl vygenerován za využití pravoúhlého okénka), delta a delta-delta koeficienty zkratkou *deltas* a normalizované příznaky *norm*, např.:

- log_fb_en_25_10_ham_deltas - logaritmované energie banky filtrů, segmentováno pomocí Hammingova okénka o délce 25ms s posunem 10ms, vypočteny delta a delta-delta koeficienty,
- stzcr_10_10_norm - krátkodobé průchody nulou, segmentováno pomocí pravoúhlého okénka o délce 10ms s posunem 10ms, normalizovány

ste_10_10	ste_10_10_norm
sti_10_10	sti_10_10_norm
stzcr_10_10	stzcr_10_10_norm
ste_sti_stzcr_10_10	ste_sti_stzcr_10_10_norm
log_fb_en_25_10_ham	log_fb_en_25_10_ham_norm
log_fb_en_25_10_ham_deltas	log_fb_en_25_10_ham_deltas_norm
mfcc_25_10_ham	mfcc_25_10_ham_norm
mfcc_25_10_ham_deltas	mfcc_25_10_ham_deltas_norm

Tabulka 2: Vygenerované parametrizace příznaků.

Z tabulky 2 je patrné, že všechny příznaky v časové oblasti byly vygenerovány s využitím pravoúhlého okénka o délce 10ms, které se posouvá o 10ms. Nedochází tedy k přesahu mezi jednotlivými mikrosegmenty. Příznaky ve frekvenční oblasti byly vygenerovány s využitím Hammingova okénka o délce 25ms s posunem 10ms (jednotlivé mikrosegmenty se částečně překrývají) a 512 bodové FFT. Pro výpočet logaritmovaných energií banky filtrů bylo využito 40 filtrů, pro výpočet MFCC 26 filtrů.

7.2 Dynamic Time Warping

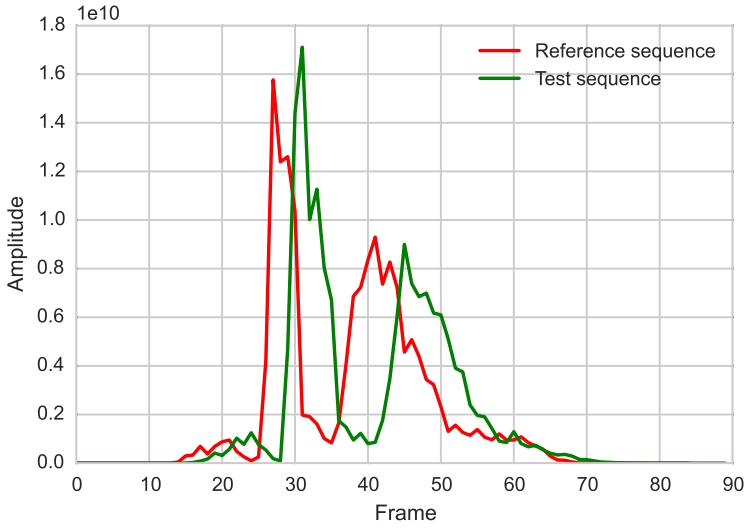
Pro výpočet DTW vzdálenosti byl vytvořen modul, který obsahuje základní DTW algoritmus bez globálního vymezení pohybu funkce a využívá symetrické omezení na lokální strmost (první zleva na obrázku 10). Jako vzdálenostní metrika pro všechny typy příznaků kromě *ste_sti_stzcr* byla zvolena Euklidova vzdálenost.

7.2.1 Optimalizace parametrů vzdálenostní metriky

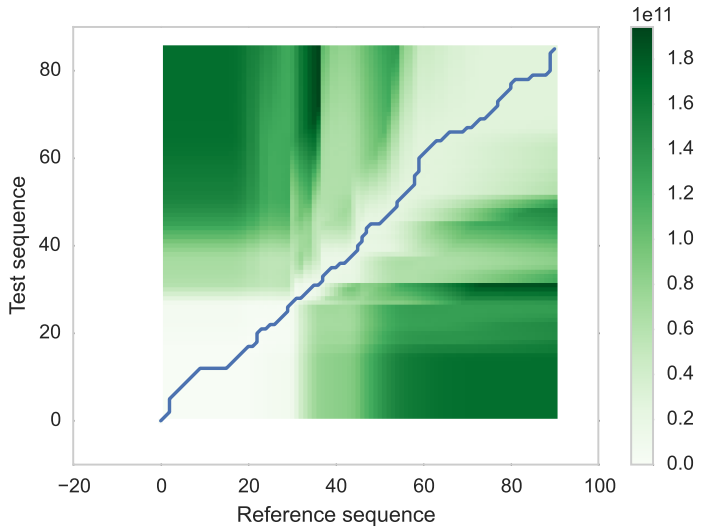
Pro kombinaci příznaků složenou z krátkodobé energie, intenzity a průchodů nulou byla využita vlastní vzdálenostní metrika

$$d(a_i, b_j) = \alpha|a_{ste,i} - b_{ste,j}| + \beta|a_{sti,i} - b_{sti,j}| + \gamma|a_{stzcr,i} - b_{stzcr,j}|, \tag{7.1}$$

kde a_i je i-tý příznak testovaného obrazu A, b_j je j-tý příznak referenčního obrazu B a α, β, γ jsou volitelné parametry.



Obrázek 12: Krátkodobé energie pro dvě různé nahrávky číslovky 7.



Obrázek 13: Průběh funkce DTW pro dva různé obrazy reprezentující číslovku 7.

Pro nenormalizovanou verzi byly zvoleny parametry $\alpha = \beta = \gamma = 1$, pro normalizovanou verzi byla provedena optimalizace parametrů, aby bylo možné určit, který z příznaků v časové oblasti má největší informativní hodnotu pro problém klasifikace izolovaných slov. Pro optimalizaci byla zvolena brute-force metoda, kdy byl procházen seznam možných parametrů s krokem 0.1 za podmínky $\alpha + \beta + \gamma = 1$. Pro každou trojici parametrů pak byla vyhodnocena přesnost klasifikace.

Nejvyšší přesnosti klasifikace v rámci jednoho řečníka bylo dosaženo s parametry $\alpha = 0.3$, $\beta = 0.3$, $\gamma = 0.4$ - každý typ příznaků se tedy projeví přibližně stejnou mírou. Optimalizací mezi všemi řečníky pak byly získány parametry $\alpha = 0$, $\beta = 0.6$, $\gamma = 0.4$ - je tedy zřejmé, že krátkodobá energie v kombinaci s krátkodobou intenzitou a průchody

nulou negativně ovlivňuje přesnost klasifikace (pravděpodobně z důvodu závislosti na řečníkovi, viz. kapitola Vyhodnocení).

Na tyto dvě parametrizace se dále budeme odkazovat jako *ste_sti_stzcr_10_10 _norm_single*, resp. *ste_sti_stzcr_10_10_norm_all*.

7.3 Support Vector Machine

Ke klasifikaci normalizovaných příznaků byl využit také SVM klasifikátor s předpočítanou jádrovou funkcí ve tvaru

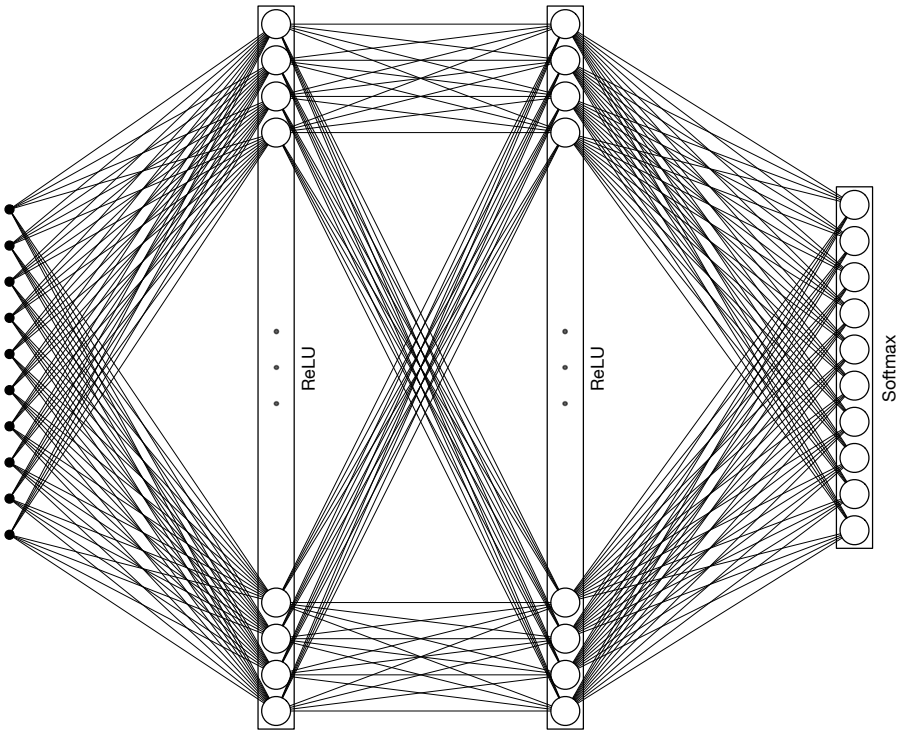
$$K(A, B) = 1 - DTW(A, B). \tag{7.2}$$

Jádrová funkce byla vypočtena ze všech referenčních obrazů navzájem a funkcionálně odpovídá Gramově matici $G = X^T X$.

7.4 Neuronová síť

Pro každého řečníka byla natrénována třívrstvá neuronová síť (znázorněna na obrázku 14). Vstupem této sítě je DTW vzdálenost testované nahrávky vůči referenčním nahrávkám, výstupem pak vektor obsahující pravděpodobnosti náležitosti do daných tříd (zajištěno aktivační funkcí Softmax). Po otestovaní několika různých parametrizací neuronové sítě byly zvoleny dvě skryté vrstvy o 200 neuronech s aktivační funkcí ReLU.

Síť byla trénována 1500 trénovacích epoch s využitím dávkového učení s dávkami o velikosti 10. Váhy sítě byly modifikovány stochastickým gradientním sestupem rozšířeným o Nesterovo momentum $\mu = 0.9$. Konstanta učení byla zvolena $\alpha = 0.01$. Kvůli předpokladům tohoto klasifikačního algoritmu byla neuronová síť natrénována pouze pro normalizovaná data.

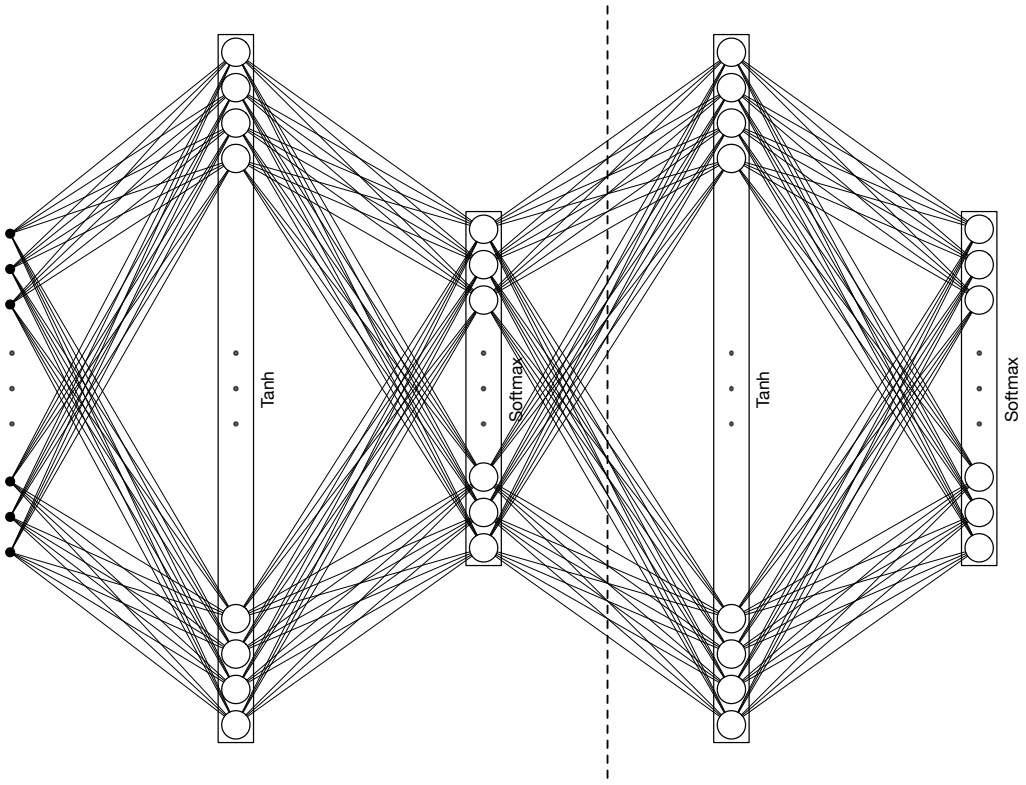


Obrázek 14: Třívrstvá neuronová síť.

7.5 Bottleneck

Pro generování příznaků pomocí neuronové sítě byl použit tzv. bottleneck. Bottleneckem je nazývána taková struktura neuronové sítě, kdy jedna ze skrytých vrstev (bottleneck vrstva) obsahuje výrazně nižší počet neuronů než její sousední vrstvy. Po natrénování neuronové sítě je bottleneck vrstva využita jako výstupní vrstva a všechny následující vrstvy jsou odstraněny. Tím dochází ke kompresi vstupní informace do příznakového vektoru o konstantní délce.

Bottleneck byl vytvořen natrénováním čtyřvrstvé neuronové sítě a odstraněním výstupní a poslední skryté vrstvy (znázorněno na obrázku 15). První a třetí skrytá vrstva se skládá z 1024 neuronů s aktivační funkcí Tanh, druhá (bottleneck) vrstva a výstupní vrstva mají aktivační funkci Softmax.



Obrázek 15: Čtyřvrstvá neuronová síť.

Neuronová síť byla natrénována pro dvě různé datové sady - původní datovou sadu šesti řečníků a původní datovou sadu rozšířenou o nahrávky pro Škoda Auto. Vstupem neuronové sítě je vektor o velikosti 440 složený z logaritmované energie banky filtru pro jeden foném (vektor o velikosti 40) a jeho kontextu zleva a zprava (vektory o velikosti $5 \cdot 40 = 200$).

Síť byla trénována pro klasifikaci 20 fonému pro původní datovou sadu s 8 a 16 neurony v bottleneck vrstvě a pro klasifikaci 40 fonému s 8, 16 a 32 neurony pro rozšířenou sadu. Pro přehlednější vyhodnocení si opět zavedeme značení pro vygenerované příznaky:

- *bn_X* - neuronová síť natrénovaná pro původní datovou sadu,
- *bn_SA_X* - neuronová síť natrénovaná pro rozšířenou datovou sadu,

kde X značí počet neuronů bottleneck vrstvy.

8 Vyhodnocení

Při vyhodnocení procentuální přesnosti klasifikace jednotlivých metod byly uvažovány tři různé varianty:

1. přesnost klasifikace v rámci jednoho řečníka (testovací nahrávky jednoho řečníka vůči svým referenčním nahrávkám)

2. přesnost klasifikace v rámci jednoho řečníka vůči ostatním (testovací nahrávky všech řečníků vůči referenčním nahrávkám jednoho řečníka)

3. přesnost klasifikace mezi všemi řečníky (testovací nahrávky všech řečníků vůči referenčním nahrávkám všech řečníků)

```
for speaker in speakers do
    reference_features, reference_targets = get_references(speaker)
    test_features, test_targets = get_test(speaker)
    model = train(reference_features, reference_targets)
    prediction = predict(test_features)
    accuracy = calculate_accuracy(prediction, test_targets)
end
```

Algoritmus 1: Vyhodnocení přesnosti klasifikace v rámci jednoho řečníka.

```
for speaker in speakers do
    /* combine test features/targets of all speakers                    */
    test_features, test_targets += get_test(speaker)
end
for speaker in speakers do
    reference_features, reference_targets = get_references(speaker)
    model = train(reference_features, reference_targets)
    prediction = predict(test_features)
    accuracy = calculate_accuracy(prediction, test_targets)
end
```

Algoritmus 2: Vyhodnocení přesnosti klasifikace v rámci jednoho řečníka vůči ostatním.

```
for speaker in speakers do
    /* combine reference and test features/targets of all speakers  */
    reference_features, reference_targets += get_references(speaker)
    test_features, test_targets += get_test(speaker)
end
model = train(reference_features, reference_targets)
prediction = predict(test_features)
accuracy = calculate_accuracy(prediction, test_targets)
```

Algoritmus 3: Vyhodnocení přesnosti klasifikace mezi všemi řečníky.

8.1 Přesnost klasifikace v rámci jednoho řečníka

Z tabulky 3 je zřejmé, že nejvyšší přesnosti klasifikace v rámci jednoho řečníka je dosaženo využitím příznaků generovaných neuronovou sítí. Pro příznaky generované neuronovou sítí natrénovanou nad původní sadou s bottleneck vrstvou o 8 neuronech je dokonce dosaženo nejvyšší přesnosti ze všech zkoumaných metod.

Velice vysoké přesnosti také dosahují příznaky ve frekvenční oblasti s tím, že Z-score normalizace jejich přesnost mírně snižuje. Ačkoliv by přidáním delta a delta-delta koeficientů mělo dojít k nárůstu přesnosti klasifikace, z neznámých důvodů došlo k jejímu poklesu. V případě normalizovaných příznaků se tento pokles pohybuje dokonce mezi 25-30%.

Jednotlivé příznaky v časové oblasti dle očekávání nedosahují vysokých přesností. Pro krátkodobou energii a intenzitu ovšem velmi pomáhá normalizace dat. Dále si můžeme povšimnout, že přesnost nenormalizované krátkodobé energie je stejná jako přesnost nenormalizované kombinace příznaků krátkodobé energie, intenzity a průchodů nulou. To je způsobeno tím, že nenormalizované hodnoty krátkodobé energie se pohybují v řádech desetitisíců, zatímco hodnoty krátkodobé intenzity a průchodů nulou v řádech desítek a oproti krátkodobé energii se projeví jen minimálně. Normalizací a kombinací těchto tří typů příznaků pak dostáváme typ příznaku s poměrně vysokou informační hodnotou. Optimalizací parametrů vzdálenostní metriky skutečně došlo k navýšení přesnosti a to o 1.1%.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	90.0	93.3	96.7	96.7	93.3	100.0	95.0
bn_16	100.0	90.0	93.3	96.7	86.7	90.0	92.8
bn_SA_8	90.0	83.3	86.7	83.3	83.3	93.3	86.7
bn_SA_16	96.7	93.3	93.3	96.7	86.7	80.0	91.1
bn_SA_32	93.3	93.3	96.7	93.3	76.7	93.3	91.1
log_fb_en_25_10_ham	80.0	83.3	100.0	93.3	83.3	93.3	88.9
log_fb_en_25_10_ham_norm	86.7	86.7	93.3	90.0	76.7	93.3	87.8
log_fb_en_25_10_ham_deltas	70.0	83.3	96.7	86.7	80.0	93.3	85.0
log_fb_en_25_10_ham_deltas_norm	63.3	60.0	73.3	76.7	66.7	76.7	69.4
mfcc_25_10_ham	86.7	83.3	100.0	93.3	90.0	90.0	90.6
mfcc_25_10_ham_norm	90.0	90.0	96.7	86.7	86.7	90.0	90.0
mfcc_25_10_ham_deltas	86.7	83.3	100.0	86.7	80.0	90.0	87.8
mfcc_25_10_ham_deltas_norm	70.0	60.0	76.7	70.0	66.7	76.7	70.0
ste_10_10	60.0	20.0	33.3	46.7	43.3	26.7	38.3
ste_10_10_norm	56.7	40.0	43.3	50.0	46.7	56.7	48.9
ste_sti_stzcr_10_10	60.0	20.0	33.3	46.7	43.3	26.7	38.3
ste_sti_stzcr_10_10_norm	96.7	86.7	80.0	80.0	66.7	76.7	81.1
ste_sti_stzcr_10_10_norm_single	96.7	90.0	83.3	83.3	66.7	73.3	82.2
sti_10_10	63.3	36.7	40.0	66.7	70.0	66.7	57.2
sti_10_10_norm	76.7	43.3	56.7	70.0	56.7	70.0	62.2
stzcr_10_10	56.7	70.0	60.0	63.3	60.0	60.0	61.7
stzcr_10_10_norm	60.0	66.7	53.3	60.0	43.3	53.3	56.1

Tabulka 3: Přesnost klasifikace v rámci jednoho řečníka pro DTW.

Výsledky dosažené klasifikátorem SVM (tabulka 4) jak pro příznaky ve frekvenční oblasti, tak pro bottleneck příznaky, jsou nižší než výsledky dosažené metodou DTW. Přidáním dynamických koeficientů opět došlo k výraznému poklesu přesnosti - 48.9% pro logaritmované energie banky filtrů a 32.3% pro MFCC. Přesnosti dosažené u příznaků v časové oblasti jsou srovnatelné s přesnostmi metody DTW.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	83.3	86.7	93.3	90.0	90.0	93.3	89.4
bn_16	90.0	83.3	90.0	93.3	86.7	86.7	88.3
bn_SA_8	86.7	73.3	80.0	70.0	80.0	90.0	80.0
bn_SA_16	83.3	80.0	76.7	90.0	76.7	76.7	80.6
bn_SA_32	80.0	70.0	86.7	80.0	70.0	80.0	77.8
log_fb_en_25_10_ham_norm	80.0	90.0	80.0	83.3	76.7	86.7	82.8
log_fb_en_25_10_ham_deltas_norm	60.0	46.7	20.0	20.0	26.7	30.0	33.9
mfcc_25_10_ham_norm	70.0	56.7	76.7	76.7	70.0	73.3	70.6
mfcc_25_10_ham_deltas_norm	60.0	46.7	30.0	30.0	26.7	36.7	38.3
ste_10_10_norm	60.0	33.3	46.7	43.3	50.0	53.3	47.8
ste_sti_stzcr_10_10_norm	90.0	76.7	73.3	60.0	66.7	73.3	73.3
sti_10_10_norm	73.3	40.0	43.3	60.0	53.3	66.7	56.1
stzcr_10_10_norm	60.0	63.3	53.3	50.0	40.0	50.0	52.8

Tabulka 4: Přesnost klasifikace v rámci jednoho řečníka pro SVM.

Přesnost klasifikace pomocí neuronové sítě (tabulka 5) pro příznaky ve frekvenční oblasti je srovnatelná s přesností příznaků generovaných neuronovou sítí a klasifikovaných pomocí DTW. Přidání dynamických koeficientů tuto přesnost opět snižuje. Vysokých přesností také dosahují bottleneck příznaky. Přesnost příznaků v časové oblasti je mírně horší než u SVM.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	83.3	90.0	93.3	96.7	90.0	96.7	91.7
bn_16	96.7	90.0	86.7	83.3	83.3	83.3	87.2
bn_SA_8	86.7	90.0	93.3	80.0	93.3	83.3	87.8
bn_SA_16	93.3	93.3	96.7	96.7	83.3	80.0	90.6
bn_SA_32	90.0	86.7	100.0	96.7	83.3	76.7	88.9
log_fb_en_25_10_ham_norm	86.7	93.3	93.3	100.0	93.3	90.0	92.8
log_fb_en_25_10_ham_deltas_norm	86.7	73.3	80.0	66.7	66.7	76.7	75.0
mfcc_25_10_ham_norm	93.3	96.7	93.3	96.7	86.7	90.0	92.8
mfcc_25_10_ham_deltas_norm	93.3	73.3	76.7	86.7	83.3	86.7	83.3
ste_10_10_norm	53.3	20.0	50.0	36.7	36.7	53.3	41.7
ste_sti_stzcr_10_10_norm	96.7	66.7	76.7	86.7	66.7	73.3	77.8
sti_10_10_norm	53.3	50.0	56.7	53.3	46.7	56.7	52.8
stzcr_10_10_norm	53.3	50.0	50.0	46.7	30.0	56.7	47.8

Tabulka 5: Přesnost klasifikace v rámci jednoho řečníka pro neuronovou síť.

8.2 Přesnost klasifikace v rámci jednoho řečníka vůči ostatním

Stejně jako při klasifikaci v rámci jednoho řečníka dosahuje nejlepší výsledků metoda DTW (tabulka 6) s využitím příznaků vygenerovaných neuronovou sítí. Nejvyšší přesnosti dosahuje bottleneck o 8 a 16 neuronech vytvořený natrénováním neuronové sítě nad původní datovou sadou.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	86.1	85.0	88.9	89.4	83.3	95.6	88.1
bn_16	90.6	87.8	86.1	90.0	86.1	92.2	88.8
bn_SA_8	85.6	71.7	80.0	67.8	80.0	81.7	77.8
bn_SA_16	86.7	81.1	85.6	80.0	87.2	81.1	83.6
bn_SA_32	77.8	76.7	78.9	80.0	83.9	87.2	80.7
log_fb_en_25_10_ham	68.9	70.6	82.8	70.6	77.8	82.8	75.6
log_fb_en_25_10_ham_norm	67.8	70.6	74.4	72.2	76.1	78.9	73.3
log_fb_en_25_10_ham_deltas	61.1	66.7	77.8	65.0	71.1	72.2	69.0
log_fb_en_25_10_ham_deltas_norm	49.4	52.8	58.3	53.9	57.2	55.0	54.4
mfcc_25_10_ham	70.6	72.8	81.7	72.8	75.6	87.2	76.8
mfcc_25_10_ham_norm	59.4	63.3	66.7	59.4	56.1	73.9	63.1
mfcc_25_10_ham_deltas	67.8	68.9	80.0	70.0	74.4	83.3	74.1
mfcc_25_10_ham_deltas_norm	47.8	46.7	52.2	42.8	48.3	47.2	47.5
ste_10_10	18.3	16.1	22.8	22.8	18.9	23.3	20.4
ste_10_10_norm	28.3	22.8	25.6	29.4	22.2	27.2	25.9
ste_sti_stzcr_10_10	18.3	16.1	22.8	22.8	18.9	23.3	20.4
ste_sti_stzcr_10_10_norm	70.0	60.6	65.6	60.6	59.4	66.7	63.8
ste_sti_stzcr_10_10_norm_all	70.0	60.6	68.3	65.6	61.7	64.4	65.1
sti_10_10	20.6	30.6	26.1	38.9	30.0	39.4	30.9
sti_10_10_norm	34.4	33.3	40.6	41.1	36.1	47.2	38.8
stzcr_10_10	52.2	49.4	48.3	41.7	54.4	56.1	50.4
stzcr_10_10_norm	50.6	47.8	49.4	44.4	47.8	51.1	48.5

Tabulka 6: Přesnost klasifikace v rámci jednoho řečníka vůči ostatním pro DTW.

Porovnáním tabulek 6, 7 a 8 je patrné, že příznaky ve frekvenční oblasti nejhůře klasifikuje SVM a příznaky v časové oblasti neuronová síť. Normalizace a aplikace dynamických koeficientů má na přesnost klasifikace obdobný vliv jako v případě klasifikace v rámci jednoho řečníka.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	81.1	80.0	83.9	82.2	82.8	93.9	84.0
bn_16	86.1	78.3	81.1	82.2	82.2	89.9	83.2
bn_SA_8	80.6	60.0	75.6	60.0	70.0	77.8	70.7
bn_SA_16	75.0	60.6	71.1	73.3	75.0	77.8	72.1
bn_SA_32	68.9	58.3	69.4	69.4	62.8	70.0	66.5
log_fb_en_25_10_ham_norm	62.2	65.0	64.4	62.2	65.6	66.7	64.4
log_fb_en_25_10_ham_deltas_norm	39.4	37.2	13.3	18.3	20.0	27.8	26.0
mfcc_25_10_ham_norm	50.6	46.7	43.3	41.1	35.0	53.3	45.0
mfcc_25_10_ham_deltas_norm	37.2	35.0	15.0	21.1	18.9	24.4	25.3
ste_10_10_norm	28.3	23.9	28.9	27.2	21.1	27.2	26.1
ste_sti_stzcr_10_10_norm	66.7	56.7	59.4	48.3	55.6	60.6	57.9
sti_10_10_norm	36.7	33.3	35.6	37.2	33.3	43.9	36.7
stzcr_10_10_norm	50.6	47.2	44.4	42.8	46.1	48.9	46.7

Tabulka 7: Přesnost klasifikace v rámci jednoho řečníka vůči ostatním pro SVM.

Typ příznaků	Přesnost klasifikace [%]						
	Řečník						Průměr
	1	2	3	4	5	6	
bn_8	77.2	80.6	83.9	88.3	78.9	88.3	82.9
bn_16	85.0	88.9	78.9	89.4	82.8	83.9	84.8
bn_SA_8	76.1	70.6	81.7	72.2	82.2	77.8	76.8
bn_SA_16	74.4	80.6	78.9	75.6	87.8	76.1	78.9
bn_SA_32	70.0	77.2	78.3	87.8	82.8	73.9	78.3
log_fb_en_25_10_ham_norm	70.0	74.4	80.0	74.4	80.0	85.6	77.4
log_fb_en_25_10_ham_deltas_norm	57.8	49.4	53.3	57.2	42.2	48.9	51.5
mfcc_25_10_ham_norm	68.3	72.8	68.3	66.7	68.9	78.3	70.6
mfcc_25_10_ham_deltas_norm	63.9	58.9	63.9	54.4	62.8	67.2	61.9
ste_10_10_norm	23.3	19.4	26.1	28.3	22.2	25.6	24.2
ste_sti_stzcr_10_10_norm	56.7	46.7	64.4	58.3	58.3	66.1	58.4
sti_10_10_norm	27.2	33.9	30.0	33.9	27.8	35.6	31.4
stzcr_10_10_norm	45.6	40.6	45.6	45.6	40.6	44.4	43.7

Tabulka 8: Přesnost klasifikace v rámci jednoho řečníka vůči ostatním pro neuronovou síť.

8.3 Přesnost klasifikace mezi všemi řečníky

Při klasifikaci mezi všemi řečníky (tabulka 9) navzájem dosahují opět nejlepších výsledků bottleneck příznaky a to zejména 8 neuronový natrénovaný nad původní datovou sadou a 16 neuronový natrénovaný nad rozšířenou sadou. Velmi vysokých přesností také dosahují příznaky ve frekvenční oblasti a kombinace příznaků v časové oblasti.

Jak již bylo zmíněno v teoretické části, krátkodobá energie je silně ovlivňována výkyvy v amplitudě akustického signálu, zatímco krátkodobá intenzita tento problém nemá. Všimněme si tedy, že krátkodobá intenzita dosahuje téměř dvakrát větší přesnosti než krátkodobá energie.

	Průměrná přesnost [%]		
Typ příznaků	DTW	SVM	NN
bn_8	98.9	94.4	98.9
bn_16	96.7	92.2	95.6
bn_SA_8	90.6	76.7	91.1
bn_SA_16	98.9	73.9	92.8
bn_SA_32	94.4	75.0	91.1
log_fb_en_25_10_ham	90.6	-	-
log_fb_en_25_10_ham_norm	92.8	74.4	91.1
log_fb_en_25_10_ham_deltas	86.1	-	-
log_fb_en_25_10_ham_deltas_norm	73.3	30.0	10.0
mfcc_25_10_ham	91.7	-	-
mfcc_25_10_ham_norm	90.6	52.8	65.0
mfcc_25_10_ham_deltas	89.4	-	-
mfcc_25_10_ham_deltas_norm	74.4	33.3	10.0
ste_10_10	32.8	-	-
ste_10_10_norm	36.1	29.4	52.2
ste_sti_stzcr_10_10	32.8	-	-
ste_sti_stzcr_10_10_norm	82.2	76.1	77.2
ste_sti_stzcr_10_10_norm_all	85.6	-	-
sti_10_10	60.6	-	-
sti_10_10_norm	62.2	46.1	57.2
stzcr_10_10	63.9	-	-
stzcr_10_10_norm	56.7	53.3	58.9

Tabulka 9: Přesnost klasifikace mezi všemi řečníky.

Přesnost klasifikátoru pro příznaky v časové a frekvenční oblasti SVM je výrazně horší než přesnost DTW. Neuronová síť se zdá být velice robustní pro normalizované logaritmované energie banky filtrů, nicméně pro příznaky s dynamickými koeficienty se nepodařilo síť s danou strukturou úspěšně natrénovat a přesnost je pouhých 10%. Příznaky generované pomocí neuronové sítě s bottleneck vrstvou o 8 neuronech nad původní datovou sadou se zdají být nezávislé na klasifikační metodě a dosahují velmi vysokých přesností. Z hlediska výpočetních nároků je ovšem vhodnější volit pouze metodu DTW, jelikož klasifikátor SVM i neuronová síť využívají předpočítané DTW vzdálenosti.

9 Závěr

Cílem této práce bylo porovnat různé typy příznaků pro úlohu klasifikace izolovaných slov. V první částí práce byly představeny metody zpracování akustického signálu včetně jednotlivých typů příznaků a byly odvozeny klasifikační algoritmy. Ve druhé části pak byly představeny testované parametrizace příznaků a navržené algoritmy využité ke klasifikaci.

Typ příznaků	Průměrná přesnost klasifikace [%]								
	DTW			SVM			NN		
	1to1	AlltoAll	Rozdíl	1to1	AlltoAll	Rozdíl	1to1	AlltoAll	Rozdíl
bn_8	95.0	98.9	3.9	89.4	94.4	5	91.7	98.9	7.2
bn_16	92.8	96.7	3.9	88.3	92.2	3.9	87.2	95.6	8.4
bn_SA_8	86.7	90.6	3.9	80.0	76.7	-3.3	87.8	91.1	3.3
bn_SA_16	91.1	98.9	7.8	80.6	73.9	-6.7	90.6	92.8	2.2
bn_SA_32	91.1	94.4	3.3	77.8	75.0	-2.8	88.9	91.1	2.2
log_fb_en_25_10_ham	88.9	90.6	1.7	-	-	-	-	-	-
log_fb_en_25_10_ham_norm	87.8	92.8	5.0	82.8	74.4	-8.4	92.8	91.1	-1.7
log_fb_en_25_10_ham_deltas	85.0	86.1	1.1	-	-	-	-	-	-
log_fb_en_25_10_ham_deltas_norm	69.4	73.3	3.9	33.9	30.0	-3.9	75.0	10.0	-65.0
mfcc_25_10_ham	90.6	91.7	1.1	-	-	-	-	-	-
mfcc_25_10_ham_norm	90.0	90.6	0.6	70.6	52.8	-17.8	92.8	65.0	-27.8
mfcc_25_10_ham_deltas	87.8	89.4	1.6	-	-	-	-	-	-
mfcc_25_10_ham_deltas_norm	70.0	74.4	4.4	38.3	33.3	-5	83.3	10.0	-73.3
ste_10_10	38.3	32.8	-5.5	-	-	-	-	-	-
ste_10_10_norm	48.9	36.1	-12.8	47.8	29.4	-18.4	41.7	52.2	10.5
ste_sti_stzcr_10_10	38.3	32.8	-5.5	-	-	-	-	-	-
ste_sti_stzcr_10_10_norm	81.1	82.2	1.1	73.3	76.1	2.8	77.8	77.2	-0.6
sti_10_10	57.2	60.6	3.4	-	-	-	-	-	-
sti_10_10_norm	62.2	62.2	0.0	56.1	46.1	-10.0	52.8	57.2	4.4
stzcr_10_10	61.7	63.9	2.2	-	-	-	-	-	-
stzcr_10_10_norm	56.1	56.7	0.6	52.8	53.3	0.5	47.8	58.9	11.1

Tabulka 10: Porovnání průměrné přesnosti klasifikace v rámci jednoho řečníka (*1to1*) a mezi všemi řečníky (*AlltoAll*).

Porovnáním průměrných přesností klasifikace (tabulka 10) se ukázalo, že nejpřesnějším a nejrobustnějším příznakem (nezávislý na řečníkovi) je příznak vygenerovaný neuronovou sítí s bottleneck vrstvou. Velice dobrých výsledků také dosahovaly příznaky založené na logaritmované energii banky filtrů klasifikovaných jak pomocí metody DTW, tak pomocí neuronové sítě.

Mezi nejméně přesné typy příznaků pak patřily příznaky v časové oblasti. Ukázalo

se ovšem, že zkombinováním jednotlivých příznaků v časové oblasti lze vytvořit příznak s poměrně vysokou informační hodnotou. Ačkoliv se čekalo, že příznaky v časové oblasti budou silně závislé na řečníkovi a při klasifikaci mezi všemi řečníky dojde k poklesu přesnosti, došlo k této situaci pouze pro krátkodobou energii u metod DTW a SVM a pro krátkodobou intenzitu u SVM. U ostatních příznaků došlo naopak k navýšení přesnosti.

Pro další zlepšení dosažených výsledků by bylo vhodné zaměřit se na příznaky generované neuronovou sítí a pokusit se optimalizovat její strukturu. Dalším krokem by pak bylo otestování rekurentních neuronových sítí (zejména typu LSTM), které v dnešní době pro podobné úlohy dosahují velice dobrých výsledků. Pro využití v praxi by pak bylo potřeba tento systém propojit se systémem pro detekci hlasové aktivity (voice activity detection).

Seznam obrázků

Seznam tabulek

Reference

[1] Ami Moyal, Vered Aharonson, Ella Tetariy, , and Michal Gishri. *Phonetic Search Methods for Large Speech Databases*. Springer-Verlag New York, 2013. ISBN: 978-1-4614-6488-4.

[2] G. Chen, C. Parada, and T. N. Sainath. Query-by-example keyword spotting using long short-term memory networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[3] T. J. Hazen, W. Shen, and C. White. Query-by-example spoken term detection using phonetic posteriorgram templates. In *Automatic Speech Recognition Understanding, 2009. ASRU 2009. IEEE Workshop on*, 2009.

[4] Josef Psutka. *Komunikace s počítačem mluvenou řečí*. Academia, 1995. ISBN: 978-8-020-00203-7.

[5] Ethem Alpaydin. *Introduction to Machine Learning*. The MIT Press, 2009. ISBN: 978-0-262-01243-0.

[6] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2000. ISBN: 978-0-471-05669-0.

[7] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN: 978-1-600-49006-4.

[8] SciPy developers. Numpy and scipy documentation, 2016. [Online], Dostupné z: http://docs.scipy.org/doc/.

[9] Josef Psutka, Jindřich Matoušek, Luděk Muller, and Vlasta Radová. *Mluvíme s počítačem česky*. Academia, 2006. ISBN: 978-8-020-01309-5.

[10] M. A. Hossan, S. Memon, and M. A. Gregory. A novel approach for mfcc feature extraction. In *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, pages 1–5, 2010.

[11] Sebastian Raschka. About feature scaling and normalization, 2014. [Online], Dostupné z: http://sebastianraschka.com/Articles/2014_about_feature_scaling.html.

[12] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2015. ISBN: 978-1-461-47137-0.

[13] scikit-learn developers. scikit-learn examples, 2016. [Online], Dostupné z: http://scikit-learn.org/stable/auto_examples/index.html.

[14] Chris J.C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition*. Kluwer Academic Publishers, 1998.

[15] Jan Švec. *Diskriminativní model pro porozumeňí mluvené řeči*. PhD thesis, Západočeská univerzita v Plzni, 2013.

[16] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. *The Elements of Statistical Learning*. Springer, 2009. ISBN: 978-0-387-84857-0.

[17] Andrej Karpathy. Cs231n convolutional neural networks for visual recognition, 2016. [Online], Dostupné z: http://cs231n.github.io/neural-networks-1/.

[18] Tom M. Mitchell. *Machine Learning*. McGraw-Hill Education, 1997. ISBN: 978-0-070-42807-2.

[19] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1978.

[20] Lasagne contributors. Lasagne documentation, 2016. [Online], Dostupné z: https://lasagne.readthedocs.org/en/latest/.

[21] LibROSA development team. Librosa documentation, 2016. [Online], Dostupné z: https://bmcfee.github.io/librosa/.

[22] James Lyons. Python speech features, 2016. [Online], Dostupné z: https://github.com/jameslyons/python_speech_features.