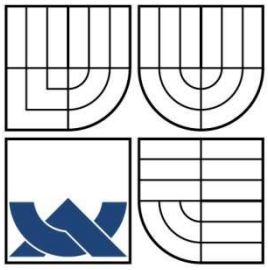
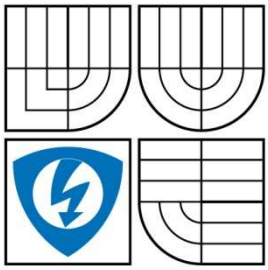


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
ÚSTAV TELEKOMUNIKACÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION
DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁNÍ EMOČNÍHO STAVU ČLOVĚKA Z ŘEČI

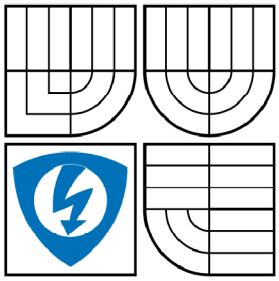
AUTOMATIC VOCAL-ORIENTED RECOGNITION OF HUMAN EMOTIONS

DIPLOMOVÁ PRÁCE
MASTER´S THESIS

AUTOR PRÁCE — BC. MIROSLAV HOUDEK
AUTHOR

VEDOUCÍ PRÁCE — ING. HICHAM ATASSI
SUPERVISOR

BRNO 2009



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav telekomunikací

Diplomová práce

magisterský navazující studijní obor
Telekomunikační a informační technika

Student: Houdek Miroslav, Bc. ***ID:*** 89089
Ročník: 2 ***Akademický rok:*** 20008/09

NÁZEV TÉMATU:

Rozpoznání emočního stavu člověka z řeči

POKYNY PRO VYPRACOVÁNÍ:

Navrhněte a zrealizujte v libovolném prostředí algoritmus, který bude schopen určit typ vstupního zvukového signálu, Porovnejte různé metody předzpracování pro výpočet příznaků a různé metody klasifikace. Získané výsledky vhodným grafickým a numerickým způsobem reprezentujte.

DOPORUČENÁ LITERATURA:

[1] Atassi H., Porovnání analýzy emočních stavů v závislosti na typu jazyka. Diplomová práce, Brno, VUT 2007.
[2] Psutka J., Komunikace s počítačem mluvenou řeči. Academia, Praha 1995.
[3] Psutka J., Muller L., Matoušek J., Radová V., Mluvíme s počítačem česky. Academia Praha 2006.
[4] R. Duda, P. Hart, D. Stork, Pattern Classification, druhé vydání. Wiley, 2003.

Termín zadání: 9.2.2009 ***Termín odevzdání:*** 26.5.2009

Vedoucí práce: Ing. Hicham Atassi

prof. Ing. Kamil Vrba, CSc.

Předseda oborové rady

UPOZORNENÍ:

ABSTRAKT

Tato diplomová práce pojednává o rozpoznání emočních stavů a určení pohlaví na základě analýzy řečového signálu. Pro popis řečového signálu jsme využili různých prozodických a kepstrálních příznaků. Součástí práce je popis neinvazivních metod pro odhad hlasivkových pulsů. Pro jednotlivé příznaky řeči jsme vytvořili funkce v programu MATLAB. Klasifikace byla provedena pomocí GMM klasifikátoru, který využívá Gaussova rozložení pravděpodobnosti pro modelování příznakového prostoru. Dále byl sestrojen systém pro rozpoznání emočních stavů mluvčího a systém pro rozpoznání pohlaví mluvčího z řeči. Úspěšnost vytvořených systémů jsme testovali s jednotlivými příznaky na různých délkách segmentů řečového signálu a výsledné procentuální úspěšnosti rozpoznávání porovnali. Závěrem jsme testovali vliv mluvčího a pohlaví na úspěšnost rozpoznání emočních stavů.

KLÍČOVÁ SLOVA

Emoce, GMM, hlasivkové pulsy, inverzní filtrace, LPC, MFCC, příznak, tempo řeči, základní tón.

ABSTRACT

This master thesis concerns with emotional states and gender recognition on the basis of speech signal analysis. We used various prosodic and cepstral features for the description of the speech signal. In the text we describe non-invasive methods for glottal pulses estimation. The described features of speech were implemented in MATLAB. For their classification we used the GMM classifier, which uses the Gaussian probability distribution for modeling a feature space. Furthermore, we constructed a system for recognition of emotional states of the speaker and a system for gender recognition from speech. We tested the success of created systems with several features on speech signal segments of various lengths and compared the results. In the last part we tested the influence of speaker and gender on the success of emotional states recognition.

KEYWORDS

Emotion, GMM, glottal pulses, inverse filtering, LPC, MFCC, feature, rate of speech, fundamental frequency.

HOUDEK, M. *Rozpoznání emočního stavu člověka z řeči.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2009. 62 s. Vedoucí diplomové práce Ing. Hicham Atassi.

PROHLÁŠENÍ

Prohlašuji, že svou diplomovou práci na téma „Rozpoznání emočního stavu člověka z řeči“ jsem vypracoval samostatně pod vedením vedoucího diplomové práce s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny uvedeny v seznamu literatury na konci práce.

Jako autor uvedené diplomové práce dále prohlašuji, že v souvislosti s vytvořením této diplomové práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961Sb.

V Brně dne ……. ……

(podpis autora)

PODĚKOVÁNÍ

Děkuji vedoucímu diplomové práce Ing. Hichamu Atassimu, za velmi užitečnou metodickou pomoc a cenné rady při zpracování práce.

V Brně dne ……. ……

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

Vztahy mezi lidmi a stroji se mění. Končí doba, kdy se lidé museli přizpůsobovat omezeným schopnostem strojů, ale stroje se začínají přizpůsobovat přirozenému prostředí lidí. Protože dochází k rozvoji složitého rozpoznávání informací, umělé inteligence, modelování emocí, tak se stále více rozšiřuje oblast konstrukce inteligentních systémů schopných se učit. Základním cílem je zkonstruovat novou generaci rozhraní mezi člověkem a počítačem nebo robotem, které by bylo schopno vnímat lidský hlas a jeho emocionální obsah. Možností využití je například konstrukce emocionálních autonomních agentů, kde emoce slouží jako signály, směrující pozornost agenta na ty podněty, které mají nějaký význam pro agentovy cíle, čímž těmto podnětům zajišťují přednostní zpracování. Vytvářejí tak sekundární rozhodovací systém, který ohodnocuje signály podle jejich významu pro hlavní úkoly agenta a patřičným způsobem pak upravuje jeho chování [1].

Cílem diplomové práce je analýza přiznaků řeči, které se využívají v systémech pro rozpoznání emočního stavu. V kapitole 1 je stručně popsána tvorba řeči. V kapitole 2 jsou popsány základní principy rozpoznávání a používané klasifikátory. V další části jsou popsány různé příznaky, které jsou využívány v systémech pro zpracování řeči. Kapitola 4 se zabývá analýzou hlasivkových pulsů a jednotlivými metodami jejich odhadu. Poslední kapitola je věnována testování kvality příznaků a vytvořených systémů pro rozpoznání emocí a rozpoznání pohlaví.

1 PROCES VYTVÁŘENÍ ŘEČI ČLOVĚKEM

Pro vytváření řeči existuje v lidském těle několik skupin orgánů, které se souhrnně nazývají řečové (artikulační) orgány, či jen mluvidla (artikulátory). Z hlediska tvorby řeči tyto řečové orgány tvoří hlasový trakt. Hlasový trakt lze rozdělit na tři základní ústrojí: dechové, hlasové a artikulační, viz obr. Obr. **1.1**[2].

1.1 Dechové ústrojí

Dechové ústrojí představuje fundamentální zdroj energie pro řeč. Je umístěno v hrudním koši a tvořeno přívodní dýchací cestou, plícemi a s nimi funkčně spjatými dýchacími svaly (bránicí). Při nádechu dochází k pohybu vzduchu, který tak poskytuje zdroj energie pro řeč. Při výdechu potom v plicích vzniká výdechový proud vzduchu, který je v zásadě základním materiálem pro tvorbu řeči. Výdechový proud vzduchu je z plic odváděn průdušnicí (tracheou), dále pak prochází hrtanem a nadhrtanovými dutinami, kde se modifikuje, a jako řečový signál je vyzařován rty do okolního prostoru. Síla výdechového proudu ovlivňuje způsob fungování hlasového ústrojí, a tím má vliv na sílu hlasu a částečné i na jeho výšku [2].

1.2 Hlasové ústrojí

Hlasové ústrojí je uloženo v hrtanu, který je s plícemi spojen průdušnicí. Z hlediska tvorby řeči nejdůležitější část hlasového ústrojí tvoří hlasivky, které se nacházejí v hrtanové dutině přímo za „ohryzkem". Jsou to dvě ostré slizniční řasy, které vedou napříč hrtanem v místě jeho nejužšího průchodu. Z jedné strany jsou napojeny na chrupavky hlasivkové a z druhé strany na chrupavku štítnou. Jsou pokryty sliznicí a jejich základ tvoří hlasový vaz a hlasivkový sval. Prostor mezi hlasivkami tvoří hlasivková štěrbina trojúhelníkového tvaru. Jestliže člověk mlčí, pak hlasivky drží hlasivkovou štěrbinu odkrytou (asi 8 mm na šířku), takže jí může bez odporu procházet vzduch k dýchání. Hlasové ústrojí tak využívá klidového postavení hlasivek. Při vytváření hlasu (fonaci) plní hlasivky jinou funkci - nacházejí se v tzv. hlasovém (fonačním) postavení [2]. Výdechový proud vzduchu postupuje bez odporu z plic průdušnicí až k hrtanu. Zde se mu do cesty postaví překážka vytvořená hmotou hlasivek, které cestu vzduchu úplně uzavřou. Stažené hlasivky se pod tlakem vzduchu stávají pružnými a začínají kmitat. Hlasivky se přitom střídavě postupně otevírají a

prudce uzavírají. V důsledku kmitání hlasivek se vzduchový proud (do té doby homogenní) „rozdrobí" tak, že se víceméně pravidelné střídá vždy kvantum hustšího a řidšího vzduchu - vzniká tzv. vzduchová vlna, kterou vnímáme jako zvuk. Tento periodický proud vzduchových pulsů tvoří základ lidského hlasu. Bývá označován termínem základní (hlasivkový) tón a představuje nosný zvuk řeči. Frekvence kmitání hlasivek se označuje F_0 a nazývá se fundamentální frekvence nebo frekvence základního hlasivkového tónu [2]. Převrácenou hodnotu $T_0 = 1/F_0$ nazýváme periodou základního hlasivkového tónu [2].

Klidové postavení hlasivek a postavení fonační představuji dvě základní protikladné varianty činnosti hlasového ústrojí. Obou těchto poloh (v drobných modifikacích) je využíváno při tvorbě řeči. Fonační postavení má za následek vznik hlasivkového tónu a používá se proto při vytvářeni znělých zvuků řeči (tj. samohlásek a znělých souhlásek) [2]. Při tvorbě samohlásek je hlasivková štěrbina téměř úplně uzavřena po celé délce a hlasové vazy jsou napjaté, kmitání hlasivek je tak pravidelné a vznikající zvuk má „čistě" tónovou strukturu. Na druhou stranu u znělých souhlásek může být napětí hlasivek menší, kmitáni je pak méně pravidelné a charakteristika výsledného zvuku již není čistě tónová. Neznělé zvuky jsou naopak tvořeny při klidovém postaveni hlasivek, neobsahují tedy základní hlasivkový tón a vznikají až modifikací výdechového proudu vzduchu v nadhrtanových dutinách [2].

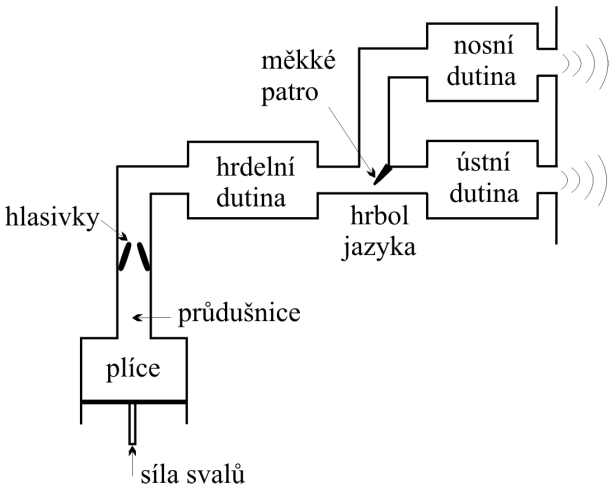
1.3 Artikulační ústrojí

Význam artikulačního ústrojí spočívá v tom, že umožňuje vytvářet velké množství různých zvuků, které charakterizují mluvený jazyk. Skládá se jednak z nadhrtanových dutin a jednak z artikulačních orgánů, které jsou v těchto dutinách uloženy nebo je obklopují [2].

Nadhrtanové dutiny se účastní procesu tvorby řeči pasivně (nepohybují se), artikulační orgány (artikulátory) se účastni tvorby řeči aktivně, neboť svým pohybem mění velikosti nadhrtanových dutin. Z hlediska vytváření řeči mezi nejvýznamnější artikulátory patří jazyk, rty a měkké patro, neboť se podílejí na vytváření největšího počtu různých zvuků. Dalšími artikulátory potom jsou zuby, tvrdé patro nebo čelisti [2].

Výrazných změn ve svém složení doznává zvuk v nadhrtanových dutinách, kam jako výdechový proud vzduchu (v podobě periodického proudu vzduchových pulsů nebo jako prostý proud vzduchu) postupuje z hlasového ústrojí. Tyto změny jsou v zásadě dvojího druhu [2]:

- **Vytváření tónové struktury.** Při průchodu základního hlasivkového tónu nadhrtanovými dutinami dochází vlivem rezonance uvnitř hrdelní, ústní a popř. i nosní dutiny ke změně rozloženi akustické energie ve vznikajícím řečovém signálu. Akustická energie se soustředí kolem určitých frekvencí, kterým říkáme formantové frekvence. Oblasti koncentrace (zesílení) akustické energie se nazývají formanty a označuji se čísly, počínaje formantem s nejnižší frekvencí, $F_1, F_2, \ldots, F_n$. Pokud se do procesu vytváření řeči zapojí i nosní dutina, dochází navíc vlivem jejích antirezonančních vlastností k potlačeni některých frekvenčních oblastí, tzv. antiformantů (značí se $A_1, A_2, \ldots, A_n$). Vzniká složený zvuk tónového charakteru, kterému říkáme hlas. Tvoří podstatu znělých částí řeči, především samohlásek. Různé zvuky (tj. zvuky s různou spektrální strukturou) přitom vznikají tak, že pohyblivé artikulátory mění tvar nadhrtanových dutin, a tím i frekvenční vlastnosti vytvářených zvuků [2].
- **Vytváření šumové složky.** Artikulátory ovlivňují svým pohybem průchod vzduchu nadhrtanovými dutinami. Výdechový proud vzduchu se pak prodírá přes vytvořené překážky a podle charakteru překážky vzniká šum různého druhu. Šum tvoří základ těch částí řeči, kterým při popisu jazyka říkáme souhlásky [2].



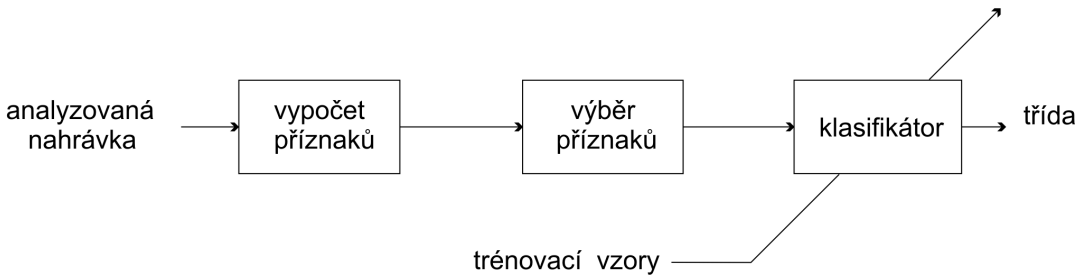
Obr. 1.1: Model hlasového traktu.

2 PROCES ROZPOZNÁNÍ

Základní blokové schéma systému pro rozpoznání je zobrazeno na obrázku Obr. 2**.1**. Proces rozpoznání se využívá pro zpracování řeči, zpracování lékařských výsledků atd. Nejčastěji používané metody rozpoznání patří např. rozpoznání mluvčího, pohlaví, emočního stavu, ale také logopedických vad malých dětí atd.

Prvním krokem systému pro rozpoznání je výpočet příznaků, neboli parametrizace řeči, kdy se snažíme řeč reprezentovat takovými příznaky (parametry), které co nejvíce popisují námi sledované vlastnosti řeči. V dalším kroku provedeme transformaci nebo selekci příznaků. Tento krok chápat jako výběr pouze podmnožiny z množiny získaných příznaků, které budou dále použity pro další zpracování a klasifikaci. Výběr vhodných příznaků snižuje dimenzi vstupního prostoru (výpočetní náročnost) a často tím umožňuje snadnější nalezení přesnějšího klasifikátoru. Snížením počtu příznaků tedy docílíme zvýšení úspěšnosti klasifikace [17].

Klíčovou část systému tvoří klasifikátor. Základním principem klasifikátoru je zařazení analyzovaných dat do již definovaného prostoru dat a to na základě pravděpodobnosti, kdy se data přiřazují do skupiny s nejvyšší pravděpodobností podobnosti.



Obr. 2.1: Blokové schéma systému pro rozpoznávání.

3 PŘÍZNAKY

Při určování příznaků se snažíme řeč reprezentovat takovými parametry, které co nejvíce popisují námi sledované vlastnosti řeči. Základním požadavky na příznaky je, aby byla separace příznaků mezi rozpoznávanými třídami maximální a aby měli příznaky co nejmenší rozptyl v rámci vlastní třídy.

Mezi sledované příznaky pro rozpoznání emocí patří např. tempo řeči, harmonicita, základní tón, krátkodobá energie, formantové frekvence, počet průchodů nulovou rovinou, koeficienty lineární predikce (LPC), percepční lineární koeficienty (PLP), mel-frekvenční kepstrální koeficienty (MFCC).

3.1 Krátkodobá energie

Hodnoty funkce krátkodobé energie poskytuji pro každý segment informaci o průměrné hodnotě energie v segmentu a je definována vztahem

$$E = \sum_{n=0}^{N-1} s^2[n], \tag{3.1}$$

kde $s[n]$ značí vzorek signálu v čase n a N znační počet vzorků segmentu reprezentuje příslušný typ okénka [2]. Jedním z nedostatků teto charakteristiky je značná citlivost na velké změny úrovně signálu. Již tak vysoká dynamika řečového signálu je vlivem kvadrátu v rovnici (3.1) ještě zvýšena [2]. Z toho důvodu se velmi často využívá krátkodobá intenzita, která zmíněný nedostatek odstraňuje

$$M = \sum_{n=0}^{N-1} |s[n]| \quad . \tag{3.2}$$

Hodnoty krátkodobé energie i krátkodobé intenzity mohou být využity například při automatickém oddělování segmentů ticha od segmentů řeči, lze jich též využít při oddělování znělých a neznělých částí promluvy [2].

3.2 Tempo

Mluvená řeč obsahuje tzv. prozodickou informaci. Termínem prozodie se označuje melodická stránka řeči. Takové vlastnosti řečového signálu, které souvisí s melodickou stránkou řeči, jsou např. frekvence základního hlasivkového tónu, energie a časování. Mezi vlastnosti řeči spojené s časováním patří počet, umisťování a délka pauz, rytmus a tempo řeči [2].

Tempo řeči se váže na celkovou rychlost vyslovení promluvy a obvykle vyjadřuje počet slov nebo slabik za jednotku času [2]. Řečovou jednotkou, která nese informaci o trvání, je slabika. Ta se skládá z dalších segmentů (samohláskových, souhláskových). Změnu trvání slabik totiž ovlivní trvání samohláskových segmentů mnohem více než segmentů souhláskových [2]. Trvání některých krátkodobých zvuků (např. výbuchy ploziv nebo přechodové segmenty typu samohláska-souhláska, souhláska-samohláska) se přitom mění jen minimálně. Příklad rozložení slabik ve slově: [mo-tor], [to-vár-na], [bra-tr].

Jednotlivé jazyky mají vlastní průměrné tempo, rychlá je např. italština, ruština je rychlejší než čeština atd. [8]. Tempo řeči závisí na spoustě různých faktorů, např. stylu mluvení (čtená nebo spontánní řeč), emocionálním stavu řečníka, členění řeči a umísťování pauz, způsobu artikulace (pečlivá nebo ledabylá) nebo rytmu řeči (slovní přízvuk či větný přízvuk, popř. bez přízvuku). Trvání slabiky může být také ovlivněno umístěním slabiky ve slově, frázi nebo v celé promluvě [2].

3.3 Základní tón

Jak je uvedeno v kapitole 3.2, frekvence základního tónu je prozodickým příznakem a vypovídá nám o melodické stránce řeči. Pomocí průměrné hodnoty základního tónu se dá odhadnout pohlaví mluvčího.

Udává se, že základní kmitočet je v rozsahu 60-400Hz. Tato hodnota je rozdílná u dětí a dospělých a také u mužů a žen [9]. Při neutrální promluvě se hodnota pohubuje v rozmezí 80-160Hz u mužů, 150-300Hz u žen a 200-600Hz u dětí. Základní tón je značně závislý na emočním stavu mluvčího, pokud tedy muž začne být vzteklí, tak základní tón řeči vzroste a může být zaměněn se základním tónem u neutrální promluvy od ženy [3]. Na obrázku 3.1a je zobrazen průběh segmentu znělé promluvy na kterém je zobrazena perioda základního tónu řeči T_0.

3.3.1 Metody detekce

Metody detekce základního tónu lze podle oblasti výpočtu rozdělit mezi 3 hlavní skupiny:

- detekce v časové oblasti
- detekce v kmitočtové oblasti
- detekce v kepstru

Mezi metody detekce základního tónu v časové oblasti patří např. metoda centrálního klipování, metoda založená na inverzní filtraci. V kmitočtové oblasti slouží pro detekci základního tónu spektrální metoda, v kepstru pak kepstrální metoda [9].

3.3.2 Centrální klipování

Metoda vychází z výpočtu autokorelační funkce (ACF). Výpočet pomocí ACF určuje míru podobnosti v segmentu promluvy. Metoda vychází z toho, že k výpočtu ACF stačí znát pouze jednotlivé špičky v průběhu promluvy [9]. Příklad výpočtu základního tónu metodou centrálního klipování je zobrazen na obrázku 3.1.

Postup výpočtu metody centrálního klipování se dá shrnout do následujících bodů:

1) Segmentace promluvy.
2) Pro jednotlivé segmenty vypočítáme práh signálu P. Prahy signálu jsou voleny symetricky k nulové rovině, tzn. $P = P_+ = -P_-$. Pro výpočet prahové hodnoty *i*-tého segmentu P_i, jsou nejdříve určena maxima obou okolních segmentů MAX_{i-1} a MAX_{i+1}, a P_i je pak vypočten jako

$$P_i = k \min\left(MAX_{i-1}, MAX_{i+1}\right), \tag{3.3}$$

   kde k značí redukční faktor s obvyklou hodnotou 0,8 [9].
3) Signál se po prahování (obr. 3.1b) normalizuje na jednotkovou velikost a tím získáme signál, který nabývá pouze tří hodnot 1, 0 a −1 (obr. 3.1c).
4) Spočteme ACF normovaného segmentu signálu dle rovnice

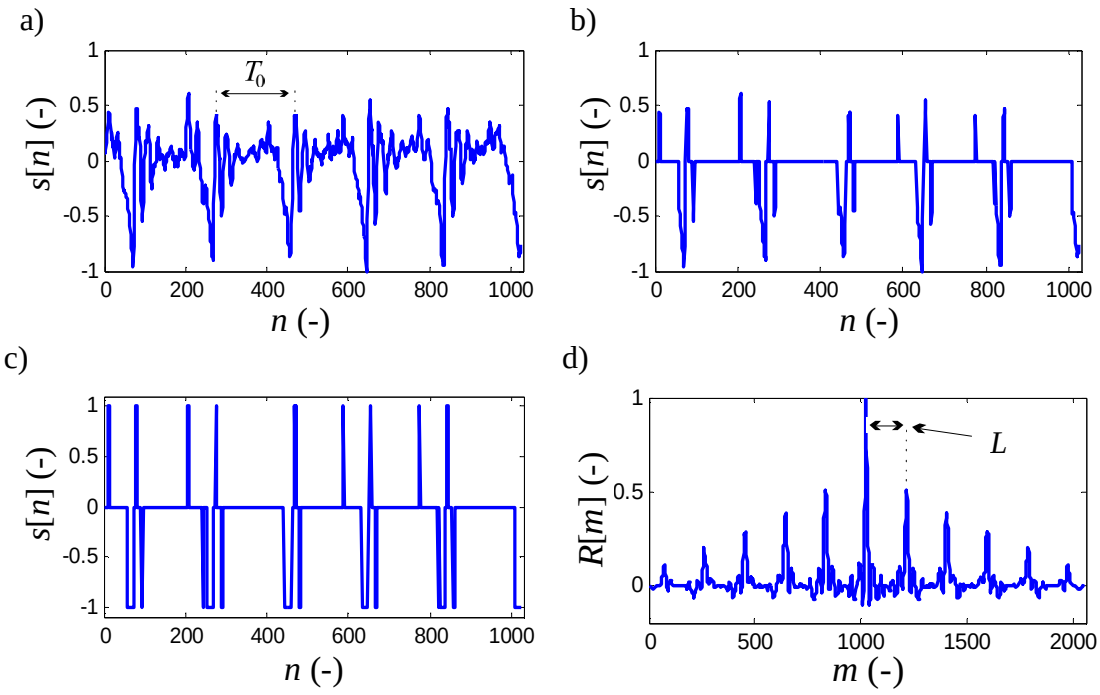
$$R[m] = \sum_{n=0}^{N-n-1} s[n]s[n+m], \tag{3.4}$$

kde N značí délku segmentu signálu a $s[n]$ vzorky signálu. V případě neznělého segmentu je její hodnota zanedbatelná kromě hodnoty $R[0]$. V případě znělého segmentu s periodou F_0 jsou patrné opakující se vrcholy v okamžicích $F_0, 2 \cdot F_0, 3 \cdot F_0, \ldots$, viz obr. 3.1d.

5) Na znělém segmentu promluvy detekujeme první vrchol $R[k]$ následující po maximální hodnotě $R[0]$ a vypočítáme kmitočet základního tónu řeči dle rovnice

$$F_0 = \frac{f_{vz}}{L} \ , \tag{3.5}$$

kde f_{vz} je vzorkovací kmitočet a L je tzv. „lag“, který udává základní periodu ve vzorcích.



Obr. 3.1: Postup výpočtu: a) segment promluvy, b) segment po prahování, c) segment po klipování, d) ACF klipovaného signálu.

3.4 Mel-frekvenční kepstrální koeficienty

Je známo, že vnímání tonů lidským uchem není lineárně závislé na frekvenci poslouchaného tónu, proto s měnící se frekvencí se mění i vnímáni zvuku. Zpracování pomocí mel-frekvenčních kepstrálních koeficientů (MFCC) je navrženo tak, aby do jisté míry respektovalo nelineární vnímání frekvencí lidským uchem a to využitím banky trojúhelníkových pásmových filtrů s lineárním rozložením frekvencí v tzv. melovské frekvenční škále, jež je definována vztahem

$$\hat{f}_{\text{mel}} = 2595 \log_{10}\left(1 + \frac{f_{\text{lin}}}{700}\right), \tag{3.6}$$

kde f_{lin} [Hz] je frekvence v lineární škále a $\hat{f}_{\text{mel}}$ [mel] je odpovídající frekvence v nelineární melovské škále. Převod zpět na lineární škálu je definován vztahem

$$f_{\text{lin}} = 700 \cdot \left(e^{\frac{f_{\text{mel}}}{2595}} - 1\right). \tag{3.7}$$

Trojúhelníkové filtry jsou standardně rozloženy přes celé frekvenční pásmo od nuly až do Nyquistovy frekvence [2]. Pro střední frekvence jednotlivých filtrů f_{b_i} platí vztah

$$f_{\text{b}_i} = \left(\frac{N}{f_{\text{vz}}}\right) \cdot \hat{f}_{\text{mel}}^{-1}\left(\hat{f}_{\text{mel}}(f_{\text{L}}) + i \cdot \frac{\hat{f}_{\text{mel}}(f_{\text{H}}) - \hat{f}_{\text{mel}}(f_{\text{L}})}{M_{\text{b}} + 1}\right), \tag{3.8}$$

kde funkce $f_{\text{mel}}(.)$ značí převod dle rovnice (3.6), $\hat{f}_{\text{mel}}^{-1}(.)$ značí převod dle rovnice (3.7), M_{b} je počet filtrů, f_{vz} značí vzorkovací frekvenci, f_{L} značí nejnižší a f_{H} nejvyšší hranici celé banky filtrů [10]. Každý filtr je definovaný jako

$$H_i(k) = \begin{cases} 0 & \text{pro} \quad k < f_{\text{b}_{i-1}} \\ \dfrac{k - f_{\text{b}_{i-1}}}{f_{\text{b}_i} - f_{\text{b}_{i-1}}} & \text{pro} \quad f_{\text{b}_{i-1}} \le k \le f_{\text{b}_i} \\ \dfrac{f_{\text{b}_{i+1}} - k}{f_{\text{b}_{i+1}} - f_{\text{b}_i}} & \text{pro} \quad f_{\text{b}_i} \le k \le f_{\text{b}_{i+1}} \\ 0 & \text{pro} \quad k > f_{\text{b}_{i+1}} \end{cases}, \tag{3.9}$$

kde i značí index filtru, $k = 1,2,...,N$ odpovídá k-tému koeficientu z N-bodové FFT [10].

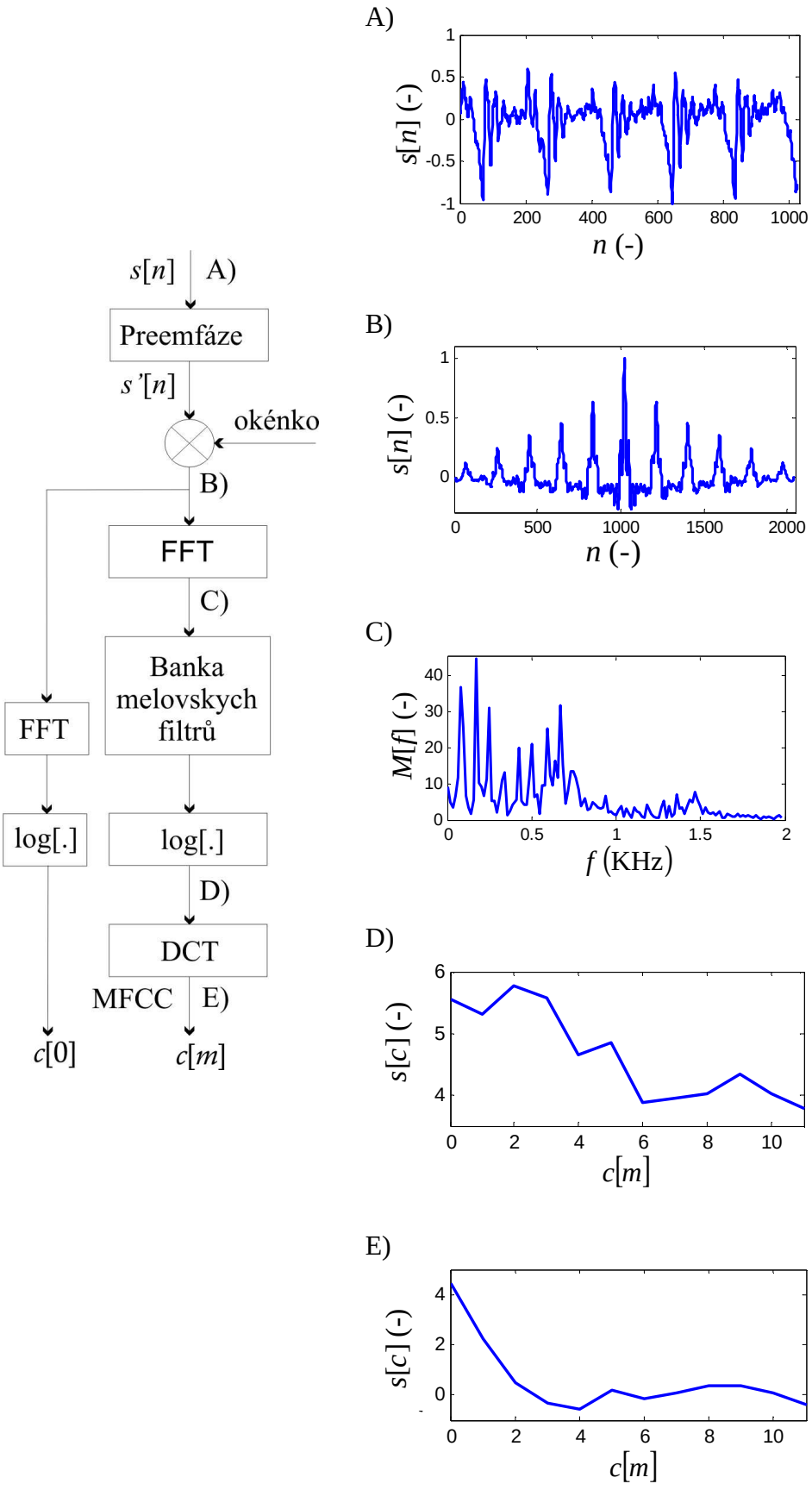
Frekvenční pásmové odezvy jednotlivých filtrů mají v melovské frekvenční škále tvar rovnoramenných trojúhelníků a jsou rovnoměrně rozloženy ve frekvenci, viz obr. 3.3a. Rozmístění banky filtrů v lineární škále je zobrazen na obrázku 3.3b. Proces určení MFCC koeficientů je zobrazen na obrázku 3.2 a lze popsat následujícím postupem [2].

Na vstup systému jsou přiváděny vzorky řečového signálu $s[n]$. Je provedena preemfáze signálu a na segmenty signálu je aplikováno nejčastěji Hammingovo okénko. V dalším bloku se pomocí FFT vypočte amplitudové spektrum $|S(f)|$ analyzovaného signálu. Poté přichází na řadu filtrace pomocí banky melovských filtrů. Každý koeficient FFT je násoben odpovídajícím ziskem filtru a výsledky jsou pro příslušné filtry akumulovány. Dalším krokem je výpočet logaritmu výstupů jednotlivých filtrů. Logaritmováním akumulovaných koeficientů se vkládá do procesu zpracování důležitý příznak kepstrální analýzy, který jednak příznivě omezí dynamiku signálu (podobně jako to dělá i lidské ucho), a dále přinese i užitečné předpoklady pro případné další zpracování signálu. Posledním krokem při výpočtu MFCC koeficientů je provedení zpětné diskrétní Fourierovy transformace (IFFT). Vzhledem k tomu, že výkonové spektrum je reálné a symetrické, bude se IFFT redukovat na diskrétní kosinovou transformaci (DCT) [2].

První koeficient $c[0]$ je roven

$$c[0] = \log E \ , \tag{3.10}$$

kde E je krátkodobá energie signálu počítaná ze vzorků signálu daného segmentu dle rovnice (3.1).



Obr. 3.2: Postupu výpočtu mel-frekvenčních kepstrálních koeficientů.

3.5 Kepstrální koeficienty typu HF

Kepstrální koeficienty typu HF (Human Factor Cepstral Coefficients - HFCC) představili v roce 2004 Skowronski a Harris, kteří navrhli banku složenou z 29 filtrů. HFCC koeficienty vychází z MFCC koeficientů a mají pouze aktualizovanou banku filtrů [10]. Rozdíl mezi HFCC a MFCC bankou filtrů je v šířce frekvenčních pásem jednotlivých filtrů a jejich rozestupu. Na obrázku 3.3c je zobrazeno rozmístění banky filtrů v lineární škále. Šířka pásma filtrů je v HFCC odvozena z ekvivalentní pravoúhlé šířky pásma ERB představené Moorem a Glasbergem

$$ERB = 6{,}23 \cdot 10^{-6} \cdot f_{\mathrm{c}}^{2} + 93{,}39 \cdot 10^{-3} \cdot f_{\mathrm{c}} + 28{,}52, \tag{3.11}$$

kde f_{c} je střední frekvence filtru v Hz [10]. Šířka pásma filtrů je dále násobena konstantou, kterou Skowronski a Harris označil jako E-faktor.

Při návrhu HFCC banky filtrů se nejprve určí nejnižší f_{L} a nejvyšší f_{H} hranice celé banky filtrů a počet filtrů M_{b}. Hodnoty středních frekvencí prvního a posledního filtru se vypočítají dle

$$f_{\mathrm{c}_i} = \frac{1}{2} \cdot \left(-\bar{b} + \sqrt{\bar{b}^{2} - 4 \cdot \bar{c}} \right), \tag{3.12}$$

kde index i je buď 1 nebo M a $\bar{b}, \bar{c}$ jsou definovány jako

$$\bar{b} = \frac{b - \hat{b}}{a - \hat{a}} \quad \text{a} \quad \bar{c} = \frac{c - \hat{c}}{a - \hat{a}}. \tag{3.13}$$

Hodnoty a, b, c jsou určeny z rovnice (3.11) a platí pro ně

$$a = 6{,}23 \cdot 10^{-6},\ b = 93{,}39 \cdot 10^{-3},\ c = 28{,}52. \tag{3.14}$$

Pro první filtr se hodnoty koeficientů $\hat{a}, \hat{b}, \hat{c}$ vypočítají takto

$$\hat{a} = \frac{1}{2} \cdot \frac{1}{700 + f_{\mathrm{L}}},\ \hat{b} = \frac{1}{700 + f_{\mathrm{L}}},\ \hat{c} = -\frac{f_{\mathrm{L}}}{2} \cdot \left(1 + \frac{700}{700 + f_{\mathrm{L}}} \right). \tag{3.15}$$

Pro poslední filtr takto

$$\hat{a} = -\frac{1}{2} \cdot \frac{1}{700 + f_{\mathrm{H}}},\ \hat{b} = -\frac{1}{700 + f_{\mathrm{H}}},\ \hat{c} = \frac{f_{H}}{2} \cdot \left(1 + \frac{700}{700 + f_{\mathrm{H}}} \right). \tag{3.16}$$

Jakmile vypočítáme hodnoty středních kmitočtů prvního a posledního filtru střední kmitočty filtrů umístněnými mezi nimi snadno dopočítáme, protože jsou od sebe lineárně vzdáleny na melovské stupnici. Přírůstek $\Delta\hat{f}$ mezi středními hodnotami ostatních filtrů spočteme jako

$$\Delta\hat{f} = \frac{\hat{f}_{c_M} - \hat{f}_{c_1}}{M-1}, \tag{3.17}$$

kde jednotky všech frekvencí jsou v melech. Konverze f_{c_1} na $\hat{f}_{c_1}$ a f_{c_M} na $\hat{f}_{c_M}$ je dána dle vztahu (3.6). Pokud máme $\Delta\hat{f}$, pak střední frekvence jednotlivých filtrů $\hat{f}_{c_i}$ spočítáme jako

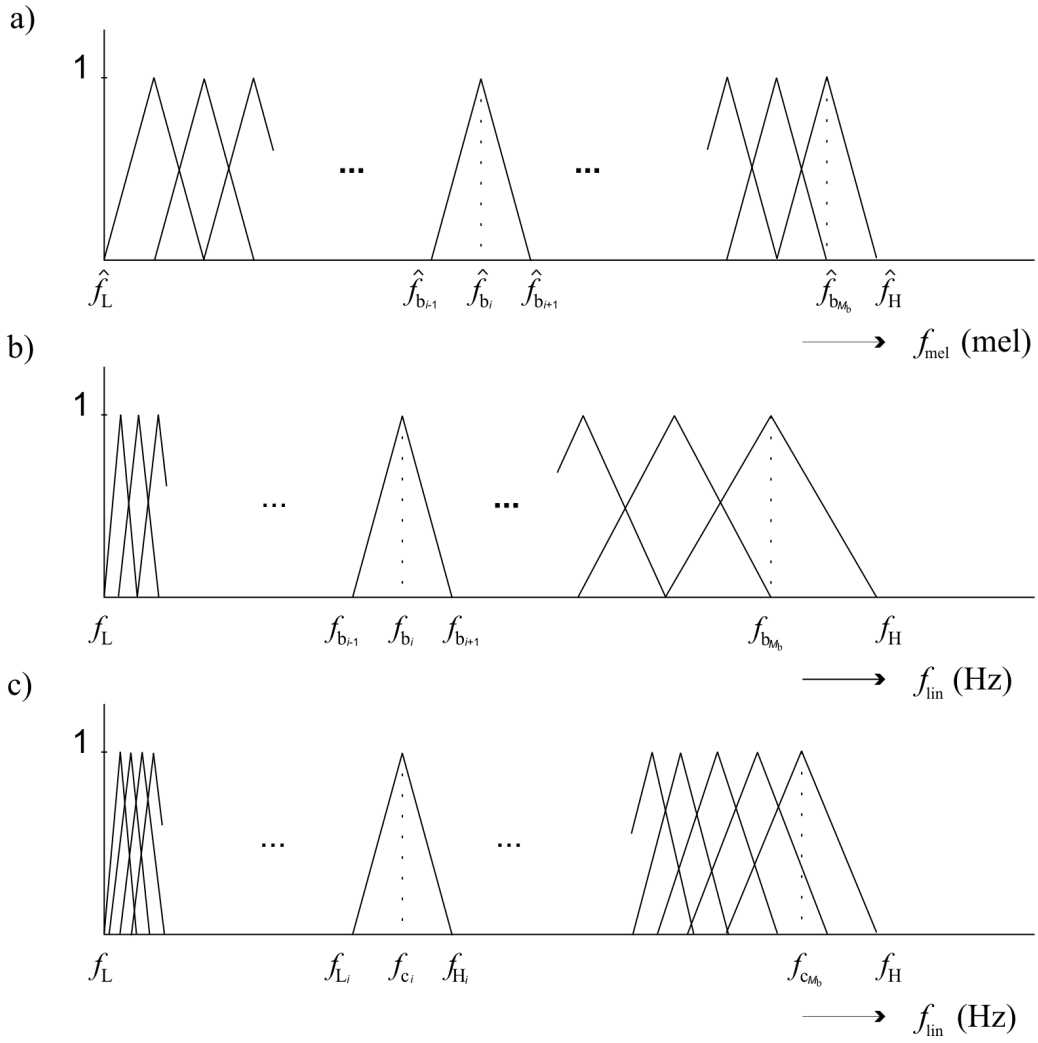
$$\hat{f}_{c_i} = \hat{f}_{c_1} + (i-1)\cdot\Delta\hat{f} \quad , \qquad \text{pro } i = 2,....,M-1. \tag{3.18}$$

Poté provedeme zpětnou konverzí $\hat{f}_{c_i}$ na f_{c_i} dle rovnice (3.7) a spočteme šířku filtru ERB_i dle rovnice (3.11) pro každý střední kmitočet f_{c_i}. Nakonec spočteme spodní hranice f_{L_i} a horní hranice f_{H_i} pro každý filtr banky jako

$$f_{L_i} = -(700 + ERB_i) + \sqrt{(700 + ERB_i)^2 + f_{c_i}\left(f_{c_i} + 1400\right)}\ , \tag{3.19}$$

$$f_{H_i} = f_{L_i} + 2\cdot ERB_i. \tag{3.20}$$

Nulový koeficient $c_H[k]$ je stejně jako u MFCC roven logaritmu krátkodobé energie počítané přímo z časových vzorků signálu dle rovnice (3.10).



Obr. 3.3: Banka trojúhelníkových filtrů a) MFCC v melovské škále, b) MFCC v lineární škále, c) HFCC v lineární škále.

3.6 Lineární predikční analýza

Lineární prediktivní kódováni (LPC) je jednou z nejefektivnějších metod analýzy akustického signálu. Jeto metoda, která se snaží na krátkodobém základu odhadnout přímo z řečového signálu parametry modelu vytvářeni řeči [2]. Mezi největší výhody metody patří její přijatelná výpočetní náročnost [2].

Model vytváření řeči se skládá ze systému s časově proměnným přenosem a generátoru budicích funkcí, který budí systém při vytváření znělých zvuků posloupností impulsů a při vytváření neznělých zvuků náhodným šumem [2]. Proces modelování řeči je znázorněn na obrázku 3.4.

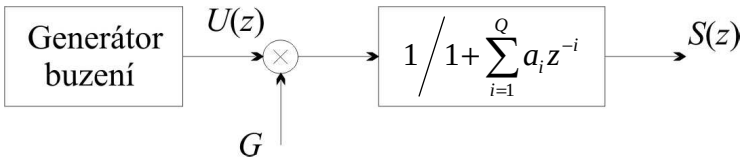
Princip metody LPC je založen na předpokladu, že *n*-tý vzorek signálu $s[n]$ lze popsat lineární kombinací Q předchozích vzorků a buzeni $u[n]$.

$$s[n] = -\sum_{i=1}^{Q} a_i s[n-1] + Gu[n], \tag{3.21}$$

kde G je koeficient zesíleni, a_i jsou koeficienty číslicového filtru a Q je řád modelu. Přenosovou funkci modelu $H(z)$ lze pak zapsat ve tvaru

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{A(z)} = \frac{G}{1+\sum_{i=1}^{Q} a_i z^{-i}} \ . \tag{3.22}$$

Sledovanými parametry jsou zde koeficienty číslicového filtru a_i a koeficient zesílení G [2].



Obr. 3.4: Model vytváření řeči s lineárním číslicovým filtrem.

Koeficienty a_i se počítají autokorelační metodou dle rovnice (3.3). Vztah mezi autokorelačními koeficienty $R[m]$ a koeficienty a_i lze vyjádřit jako soustavu Q rovnic, jenž lze přepsat do maticového tvaru

$$\begin{bmatrix} R[0] & R[1] & R[2] & \cdots & R[Q-1] \\ R[1] & R[0] & R[1] & \cdots & R[Q-2] \\ \vdots & & & & \\ R[Q-1] & R[Q-2] & R[Q-3] & \cdots & R[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_Q \end{bmatrix} = \begin{bmatrix} -R[1] \\ -R[2] \\ \vdots \\ -R[Q] \end{bmatrix}, \tag{3.23}$$

protože je matice symetrická a Toeplitzově tvaru, lze k výpočtu koeficientů a_i použít vysoce účinného iterativního Levinson-Durbinova algoritmu kde

$$\begin{aligned} & E^{(0)} = R[0] \,, \\ & k_i = -\frac{R[i] + \sum_{j=1}^{i-1} a_j^{(i-1)} R[i-j]}{E^{(i-1)}} \,, \\ & a_i^{(i)} = k_i \,, \\ & a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \,, \qquad \text{pro} \quad j = 1,2,\ldots,i-1, \\ & E^{(i)} = \left(1 - k_i^2\right) E^{(i-1)} \,, \end{aligned} \tag{3.24}$$

kde $a_j^{(i)}$ značí *j*-tý koeficienty predátoru řádu *i*. V praxi se často místo koeficientů a_i využívá mezivýsledků Levinson-Durbinova algoritmu tzv. koeficientů odrazu k_i kde $i = 1,2,\ldots,Q$ [2]. Tyto koeficienty jsou známy také pod označením PARCOR [2].

Rovnice pro výpočet zesílení G je dán vztahem

$$G^2 = R[0] + \sum_{i=1}^{Q} a_i R[i]. \tag{3.25}$$

Koeficientů LPC se využívá při určování spektrálních charakteristik modelu hlasového ústrojí a lze s jejich pomocí dobře identifikovat formantovou strukturu řeči. Protože koeficienty a_i nesou informaci o spektrálních vlastnostech řečového signálu, lze je využít jednak pro syntézu řeči, ale též přímo jako příznaků v systémech rozpoznávání řeči [2].

3.7 Kepstrální koeficienty lineární predikce

Lineární systém, kterým je modelován hlasový trakt, je také možné popsat pomoci kepstrálních koeficientů lineární predikce (LPCC). Kepstrální koeficienty lineární predikce $c_{\mathrm{L}}[k]$ jsou vztaženy ke spektrální obálce a lze je odvodit přímo z LPC koeficientů a_i dle rovnice

$$c_{\mathrm{L}}[1] = a_1 \ , \tag{3.26}$$

$$c_{\mathrm{L}}[k] = -a_k - \sum_{i=1}^{k-1} \left(\frac{i}{k} \right) c[i] a_{k-1} \ , \qquad \text{pro} \quad k = 2,3,\ldots,Q \, .$$

Výhodou LPCC je to, že jsou obecně dekorelované, což vyhovuje systémům rozpoznání řeči založených na skrytých Markovových modelech s diagonálními kovariančními maticemi [2].

4 HLASIVKOVÉ PULSY

Tvar hlasivkových pulsů nebyl doposud zcela přesně změřen. Odhad tvaru hlasivkových pulsů z řečového signálu je problematický, protože v řečovém signálu je jejich tvar značně modifikován artikulačními orgány [11].

4.1 Metody analýzy

Analýza kmitání hlasivek v průběhu promluvy není vůbec jednoduchá, protože samotné hlasivky nejsou snadno dostupné [11]. Přesto bylo vyvinuto několik technik na zkoumání hlasivkové aktivity, ale jde i tak jde pouze o odhad nikoli o přesný průběh. V této práci představím různé metody pro měření hlasivkových pulsů.

Metody analýzy kmitání by měli být spolehlivé a přesné. Důležitou podmínkou je, aby byly metody neinvazivní, což znamená, že není zapotřebí žádné klinické operace k tomu, abychom připojili měřící zařízení. Mezi další důležité podmínky patří, aby měřící technika co nejméně zasahovala a měnila souvislou řeč. Dnešní metody se liší právě dle jednotlivých kritérií a každá z metod má své výhody i nevýhody [11].

4.1.1 Elektroglottogram

Jednou z možnosti měření činnosti hlasivek je pomocí elektroglottogramu (EGG). Je to metoda pro neinvazivní měření hlasivkových pulsů a je nyní využívána pro klinické a výzkumné účely. Základem pro EGG je měření impedance napříč krkem mluvčího. Když jsou hlasivky zavřeny tak skrz ně prochází elektrický proud. Naopak když jsou hlasivky otevřené, vzduchová mezera mezi hlasivkami způsobuje daleko vyšší impedanci napříč hrdlem. Změna impedance napříč hrdlem tedy signalizuje pohyb hlasivek [11].

Elektrody jsou umístěny na pokožku z každé strany hrdla a je jimi pouštěn vysokofrekvenční střídavý proud, který měří impedanci mezi elektrodami. Kmitočet signálu je v řádu megahertzů a proud je omezený na několik miliampérů což zajistí, že je proud nepostřehnutelný a neškodný pro mluvčího [11].

4.1.2 Videostroboskop a Videokymogram

Jedná se o obrazové nahrávání hrdla, které nám poskytuje cennou informaci o pohybech hlasivek během řeči. Pevný endoskop je vložen do úst mluvčího, světelný zdroj je použit k osvětlení hrtanu a kamera zaznamenává sled obrazů. Typická frekvence kmitání hlasivek je pro muže 120 Hz a pro ženy 200 Hz, ale mezinárodní standarty pro video (PAL a NTSC) poskytují pouze 50 a 60 snímků za vteřinu. Video techniky jsou nedostačující k pozorování hlasivek, protože snímací kmitočet je moc malý [11].

Videostoboskop je běžně užívaná technika k získání videosekvence hlasivek. Blikající světelný zdroje je využit k osvětlení hrtanu. Frekvence blikání je nastavena na nižší frekvenci než je kmitání hlasivek a to kvůli tomu aby záblesk snímal vždy pozdější fázi z periody kmitu hlasivek. Metoda má nevýhodu v tom, že z jedné periody kmitu dostaneme maximálně jeden obraz, není tedy možné získat informace o přechodných jevech uvnitř jednotlivého pulsu. Pro vysokou kvalitu pořízeného videa je zapotřebí stabilní a periodické chvění hlasových orgánů [11].

Videokymogram slouží k tomu, abychom získali vizuální informaci chvění hlasových orgánů v mnohem vyšším časovém rozlišení než od videostroboskopu. Snímací frekvence se pohybuje v kolem 8000 Hz, což nám dá dostatečné časové rozlišení [11].

4.1.3 Další metody

Například photoglottogram (PGG) je metoda, která měří míru otevření hlasivkové štěrbiny jejím osvětlováním shora nebo zdola a snímáním světelné intenzity na druhé straně za účelem dostat odhad otevření hlasivek [11].

I.R. Titze (2000) popsal relativně novou techniku pro náhled na hlasivkovou štěrbinu, tzv. elektromagnetický glottogram (EMGG). Metoda používá krátkovlnné elektromagnetické vlny, k tomu aby měřili pohyb tkáně v hrtanu. Metoda nevyžaduje kontakt s kůží mluvčího. EMGG je alternativa k tradičnímu elektroglottogamu [11].

Metody pro měření hlasivkových pulsů popsané výše jsou také využívané k hodnocení výsledků mezi sebou [11].

4.2 Inverzní filtrace

Teorie zdroje a filtru poskytuje teoretický základ pro techniku inverzní filtrace. Jestliže známe přenosovou funkci hlasového traktu, pak můžeme rekonstruovat inverzní filtr. Hlasivkový signál pak získáme filtrováním hovorového signálu inverzním filtrem [11].

Inverzní filtraci představil Miller v roce 1959, který jako první představil analogové elektronické filtry k tomu, aby odstranil vliv dvou nejnižších formantů a vyzařování rtů na řečovém signálu zachyceném s mikrofonem. Analogové techniky nahradily číslicové filtry, které poprvé aplikoval na řeč jako inverzní filtr Holmes (1962). Dnes veškeré metody inverzní filtrace používají číslicové filtry kvůli flexibilitě, opakovatelnosti a snadnosti implementace číslicových technik oproti analogovým [11].

Číslicové metody inverzní filtrace můžeme rozdělit na ruční a automatické techniky. Ruční metody požadují zásah člověka k nastavení filtrů řečového signálu. Automatické metody nastavují model hlasového traktu automaticky a to nejčastěji přes LPC analýzu. Jsou zde také poloautomatické metody, např. metoda IAIF od Paavo Alku, kdy je model hlasového traktu nastavován automaticky, ale uživatel ještě ovládá několik parametrů, které mohou ovlivnit výsledný hlasivkový signál. I tak výsledný signál lze brát pouze jako odhad, protože není známa metoda, která by zcela přesně určila podobu hlasivkových pulsů [11].

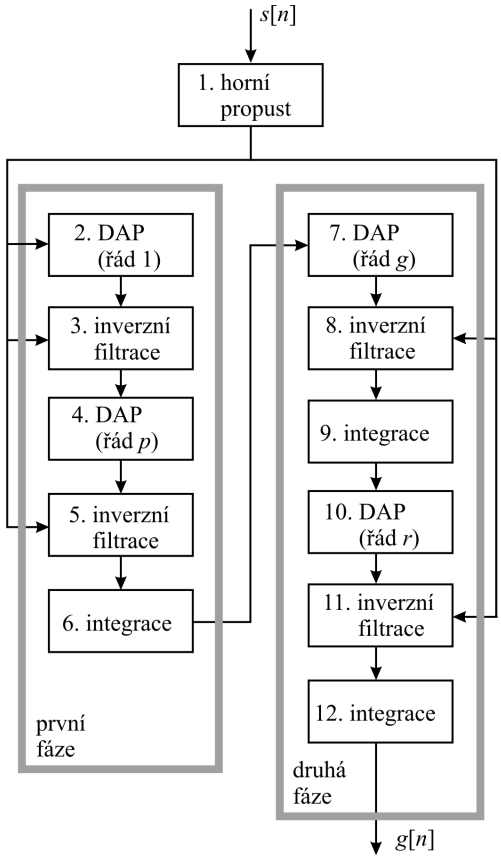
Přesnost inverzní filtrace zhoršuje vysoká základní frekvence řeči, protože rozptýlená struktura spektra buzení zasahuje do formantů, které značí rezonance ve spektru. Nosní samohlásky nejsou také vhodné k inverzní filtraci, protože jejich spektra obsahují antiformanty. I přes tato omezení je inverzní filtrace cenným nástrojem pro základní využití při odhadu hlasivkových pulsů. Inverzní filtrace je neinvazivní technikou a nevyžaduje drahé vybavení a to jsou právě její největší výhody [11].

4.2.1 IAIF

V této kapitole je naznačen postup metody opakované adaptivní inverzní filtrace (Iterative Adaptive Inverse Filtering - IAIF), který je publikován v [11].

IAIF je metoda pro odhad hlasivkových pulsů, kterou vytvořil a publikoval Paavo Alku (1992). IAIF bere jako zdroj segment promluvy $s[n]$ a na výstupu generuje odhad odpovídající proudu hlasivkového signálu $g[n]$. Základní blokové schéma metody je vidět na obrázku 4.1. Metoda pracuje ve dvou fázích. První fáze generuje odhad hlasivkového buzení, které je následně použito jako vstup pro druhou fázi metody, jenž slouží k zpřesnění odhadu.

Pro odhad hlasového traktu je využíváno lineární predikce (LP). Nevýhodou LP je větší zkreslení spektrálních obálek při aplikování na řeč s vyšší frekvencí, proto je zde místo LP použito diskrétní celopólové modelování (DAP). DAP přesněji modeluje odhady spektrálních obálek, které nejsou tolik zkresleny základními harmonickými [15]. Nevýhodou DAP modelování je vysoká výpočetní náročnost.



Obr. 4.1: Základní blokové schéma metody IAIF.

Vstupní signál je nejprve filtrován horní propustí pro odstranění rušivých nízkých frekvencí. Mezní kmitočet filtru by měl být nižší než základní frekvence signálu v segmentu promluvy. Takto filtrovaný signál je teprve použit jako vstup pro následující fáze metody.

V druhém kroku je spočítán počáteční odhad parametrů hlasového traktu. K tomu je použita metoda DAP (Discrete All-Pole Modeling) prvního řádu. Vstupní segment promluvy je filtrován inverzním filtrem, který využívá odhadu parametrů hlasového traktu získaných blokem 2.

Signál vzniklý inverzní filtrací je poté použit pro odhad parametrů hlasového traktu metodou DAP. Řád predikce p je určen jako

$$p = \frac{f_{VZ}}{1000} + 2, \tag{4.1}$$

kde f_{vz} značí vzorkovací kmitočet.

Vstupní segment promluvy je v bloku 5 filtrován inverzním filtrem, který využívá odhadu parametrů hlasového traktu získaných blokem 4.

Vliv rtů lze modelovat jako derivátor, proto je výstup předchozího bloku integrován, čímž se potlačí vliv rtů a na výstupu získáme první hrubý odhad hlasivkových pulsů.

V druhém opakování se nejprve vypočítá DAP analýza řádu g. Tím získáme model spektra popisující vliv hlasivkových pulsů na řečové spektrum. Hodnota řádu g je obvykle mezi 2 až 4. Vstupní řečový signál je inverzní filtrací (blok 8) zbaven vlivu hlasivek a integrací (blok 9) vlivu rtů.

Nový model hlasového traktu získáme pomocí DAP. Řád modelu r volíme stejný, jako řád modelu p v bloku 4 tzn. $r = p$.

Inverzní filtrací segmentu promluvy modelem hlasového traktu z bloku 10 získáme nový odhad hlasivkových pulsů, který integrací zbavíme vlivu rtů.

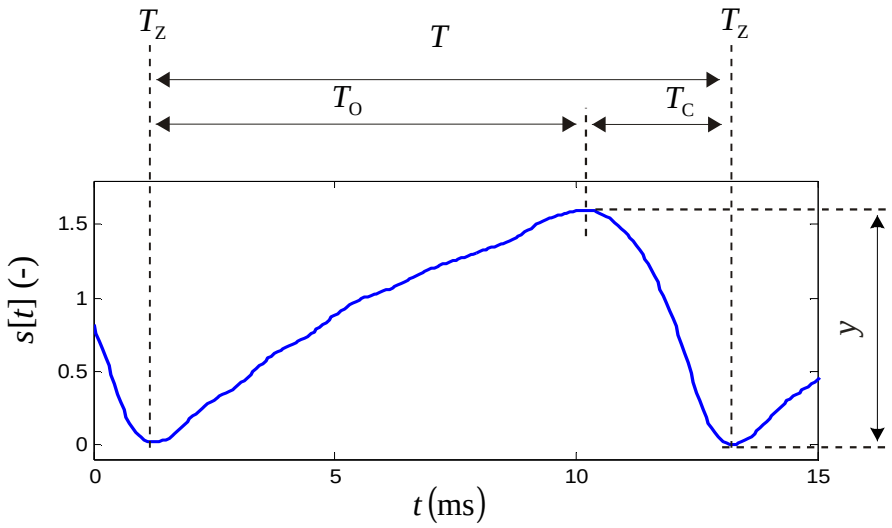
4.3 Parametrizace hlasivkového pulsu

Průběh hlasivkového pulsu je rozdělený do několika fází, viz obr. 4.2. Popis fází hlasivkového pulsu [11]:

- fáze otevírání hlasivek T_O : Hlasivky jsou alespoň minimálně otevřeny, a štěrbina mezi nimi se zvětšuje.
- fáze uzavírání hlasivek T_C : Fáze kdy jsou hlasivky maximálně otevřeny a štěrbina mezi nimi se zmenšuje.
- fáze otevření hlasivek: Hlasivky jsou otevřeny a proud vzduchu prochází skrz. Fáze otevření hlasivek je na obrázku 4.2 označena $T_O + T_C$.
- délka hlasivkového cyklu: Čas mezi dvěma odpovídajícími si okamžiky sousedních cyklů. Na obrázku 4.2 je označena jako T a je převrácenou hodnotou základního kmitočtu F_0 .
- uzavření hlasivek T_Z : Moment kdy jsou hlasivky v kontaktu podél celé jejich délky. Při uzavřené fázi neprochází skrz hlasivky žádný proud vzduchu.
- amplituda hlasivkového pulsu y : Udává míru otevření hlasivek.

Příznaky určené pro popis hlasivkových pulsů jsou uvedeny v tabulce 4.1. Při výběru některých příznaků jsme se inspirovaly prací Petera J. Murphyho [12], který se věnuje analýze hlasivkových pulzů za pomoci elektroglottogramu a následnému výzkumu vlivu emočních stavů na jednotlivé příznaky.

Příznaky sledované v této práci jsou kvocient otevírání, maximální rychlost kontaktu, maximální záporná rychlost kontaktu, rychlostní kvocient, akcelerace kontaktu, chvění, výchylka. Těmito příznaky se snažíme popsat ty nejvýznamnější vlastnosti hlasivkových pulsů. Příznaky je možné použít pro různé účely jako např. výzkum řeči, medicínské analýzy, kódování a syntézu řeči, rozpoznání řeči, identifikace mluvčího atd.



Obr. 4.2: Průběh hlasivkového pulsu.

Tab. 4.1: Příznaky sledované na hlasivkových pulsech.

symbol	**název**	**vzorec**	**popis měření**
O_Q	kvocient otevírání	$O_Q = T_O / T$	doba otevírání hlasivek dělená délkou trvání hlasivkového cyklu
V_C	maximální rychlost otevírání	$V_C = \max\left(s[t]' \right)$	maximální kladná amplituda v průběhu derivovaného hlasivkového pulsu
V_O	maximální rychlost uzavírání	$V_O = \min\left(s[t]' \right)$	maximální záporná amplituda v průběhu derivovaného hlasivkového pulsu
S_Q	rychlostní kvocient	$S_Q = V_C / V_O$	maximální rychlost otevírání dělená maximální rychlostí uzavírání
D_2	akcelerace kontaktu	$D_2 = \max\left(s[t]'' \right)$	maximální amplituda v průběhu druhé derivace hlasivkového pulsu
J	chvění	$J = \left\| \frac{T_i - T_{i+1}}{(T_i + T_{i+1})/2} \right\|$	rozdíl v délce dvou sousedních hlasivkových pulsů dělená průměrnou délkou těchto pulsů
S	výchylka	$S = \left\| \frac{y_i - y_{i+1}}{(y_i + y_{i+1})/2} \right\|$	rozdíl amplitud dvou sousedních hlasivkových pulsů dělená průměrnou amplitudou těchto pulsů

5 Klasifikátory

Jak bylo uvedeno v kapitole 2, základním principem klasifikátoru je zařazení analyzovaných dat do tříd.

Mezi základními typy klasifikátoru patří:

- Gausovy smíšené modely GMM (Gaussian Mixture Model)
- skryté Markovovy modely HMM (Hidden Markov Model)
- algoritmus podpůrných vektorů SVM (Support Vector Machine)
- algoritmus *k*-nejbližšího souseda KNN (*K*-Nearest Neighbor)
- umělé neuronové sítě ANN (Artificial Neural Network)

5.1 GMM

Gaussovy smíšené modely (GMM) používají Gaussova rozložení pravděpodobnosti, kdy trénovací příznaky modeluje jednou nebo lineární kombinací více Gaussových funkcí rozložení pravděpodobnosti [4]. Vícerozměrná Gaussova funkce rozložení pravděpodobnosti je modelována jako

$$f(\mathrm{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \cdot \exp\left(-\frac{1}{2} (\mathrm{x} - \mu)^{\mathrm{T}} \Sigma^{-1} (\mathrm{x} - \mu) \right), \tag{5.1}$$

kde d značí rozměr Gaussovy funkce, Σ kovarianční matici a μ vektor středních hodnot. Vektor středních hodnot určuje středy jednotlivých shluků dat a kovarianční matice určuje rozptyl Gaussových funkcí v prostoru těchto shluků dat. Rozložení Gausových funkcí s plnou kovarianční maticí znamená, že Gaussovy funkce jsou v prostoru dat libovolně natočeny. Pokud matice obsahuje pouze prvky na hlavní diagonále je rozložení Gaussových funkcí rovnoběžné s osami prostoru. Prvky na hlavní diagonále určují rozptyly ve směru jednotlivých os a prvky mimo hlavní diagonálu určují sklon v ostatních směrech. Matice je vždy diagonální a její velikost záleží na počtu příznaků.

Gaussův smíšený model poté vzniká lineární kombinací více Gaussových funkcí rozložení pravděpodobnosti

$$f(\mathrm{x}) = \sum_{i=1}^{M_\mathrm{g}} \alpha_i N_i(\mathrm{x}), \tag{5.2}$$

kde váhovací parametr α udává důležitost jednotlivých Gaussových funkcí, $N_i(\mathrm{x})$ značí jednotlivé funkce rozložení pravděpodobnosti dle rovnice (5.1) a M_g udává počet Gaussových funkcí rozložení pravděpodobnosti [4].

K definici GMM modelu je tedy třeba určit kovarianční matici Σ, vektor středních hodnot μ a váhovací parametr α. K tomu slouží EM (Expectation-Maximization) algoritmus [4]. EM algoritmus, ale vyžaduje počáteční inicializaci parametrů. K tomu je využíván *k*-means algoritmus, který nepotřebuje žádné počáteční údaje o rozložení hustoty pravděpodobnosti dat. *K*-means algoritmus vyhledá zadaný počet shluků dat (klustrů) a určí jejich středy μ_i dle

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \left\| x_j - \mu_i \right\|^2 . \tag{5.3}$$

Souřadnice středů těchto klustrů slouží jako inicializace pro GMM modely.

Pro samotné trénování GMM modelů je použit EM algoritmus (Expectation-Maximization algorithm). Jedná se o iterační algoritmus opakující dva kroky [4]:

- Expectation (odhad): provede se odhad parametrů GMM modelu a výpočet tzv. „likelihood“ funkce
- Maximization (maximalizace): v tomto kroku jsou aktualizovány parametry GMM modelu za účelem maximalizace pravděpodobnostní likelihood funkce

Cílem EM algoritmu je najít takové parametry Θ^*, pro které je logaritmus likelihood funkce maximální [4]

$$\Theta^* = \arg\max_{\Theta} \left(\log L(\Theta|\mathrm{x}) \right). \tag{5.4}$$

První krok E využívá odhadnutých hodnot *k*-means algoritmu. Další kroky E již využívají hodnot určených podle postupně zpřesňovaného odhadu M kroku.

5.2 HMM

Skrytý Markovovův model (HMM) je matematický aparát podobný konečnému automatu [5]. Jeho využití ve výpočetní technice spadá do oblasti zpracování přirozeného jazyka.

Přístup k modelování lidské řeči pomocí skrytých Markovových modelů je založen na statistických metodách, ve kterých je promluva reprezentována jako pravděpodobnostní funkce a poprvé byl použit začátkem sedmdesátých let (1971) Vintsyukem [5]. Princip této metody vychází z představy o skutečném průběhu generování řeči, to znamená, že hlasové ústrojí se během krátkého časového úseku, tzv. mikrosegmentu, který trvá přibližně 20 ms, dostává do jednoho z konečného počtu stavů artikulačních konfigurací. Můžeme říct, že se generuje například určitý foném. V tomto časovém úseku je produkován krátký zvukový signál, který závisí právě na této konfigurací a je popsán určitými spektrálními charakteristikami. Je nutné, aby počet těchto charakteristik byl konečný. Toho lze dosáhnout procesem vektorové kvantizace, kdy je vytvořena tzv. kódová kniha spektrálních vzorů. Každou analyzovanou spektrální charakteristiku je pak možno nahradit tím vzorem z kódové knihy, ke kterému je nejblíže [5].

Z této představy o vytváření řeči pak vychází její samotné modelování, a to pomocí Markovova procesu, kdy jsou v průběhu generovány dvě navzájem časově závislé konečné posloupnosti náhodných proměnných. Jednou z nich je tzv. podpůrný Markovovův řetězec, který je posloupností výše uvedených stavů a druhou je řetězec spektrálních vzorů [5]. Zde je pak možno využít konečného počtu spektrálních vzorů, což máme zaručeno vytvořením již zmiňované kódové knihy, a ke každému vzoru vytvořit náhodnostní funkci, která pravděpodobnostně ohodnocuje jeho vztah ke všem stavům.

Skrytý Markovovův model je ve své podstatě přímo konečným automatem. Má rovněž pevně zadanou množinu stavů, do kterých se může v průběhu výpočtu dostat, tyto stavy odpovídají stavům hlasového ústrojí během promluvy [5].

5.3 SVM

Algoritmus podpůrných vektrorů (SVM) je metodou strojového učení a hledá nadrovinu, která v prostoru příznaků optimálně rozděluje trénovací data. Optimalita rozdělení je definována jako maximum minima vzdálenosti mezi body z rozdělovaných dat. Tato metoda je ze své přirozenosti binární, tedy rozděluje data do dvou množin a více množin. Rozdělující nadrovina je lineární funkcí v prostoru příznaků [6].

5.4 KNN

Základem pro klasifikátor KNN je výpočet *k*-nejbližšího souseda, kdy neznámý testovaný vzorek se zařadí do té třídy dat, jejíž představitelé jsou nejvíce zastoupeny v prostoru *k*-nejbližších sousedů. Mezi nejčastěji používané metody pro určení vzdálenosti *k*-nejbližších sousedů patří Euklidovská, Hammingova, Kosínová [4].

5.5 ANN

Umělé neuronové sítě (ANN) jsou systémy pro komplexní zpracování dat, jejich vznik byl inspirován přirozenými sítěmi biologických neuronů mozku. Důležitým atributem ANN je schopnost učit se z předloh [7].

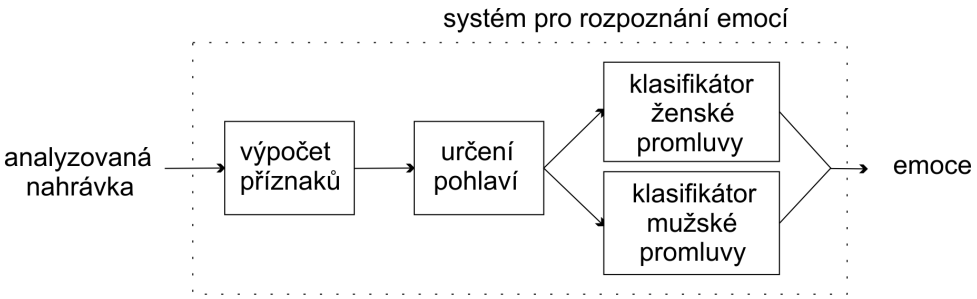
Nevýhodou neuronové sítě je její velký stupeň volnosti, jenž je dán počtem vrstev, počtem neuronů, vstupními vahami a zvolenou topologií. Nevýhodou velkého stupně volnosti je zdlouhavý proces učení [7].

6 Rozpoznání pohlaví

Při úlohách rozpoznání řeči je dokázáno [3], že přístup závislí na mluvčím vykazuje lepší výsledky z důvodu rozdílných rysů mužského a ženského hlasu. Před samotný systém rozpoznání emocí je vložen systém na rozpoznání pohlaví, který určí, zda se jedná o muže nebo ženu. Analyzovaná promluva je poté předána systému pro rozpoznání emocí, natrénovaného pouze na dané pohlaví.

Rozpoznávání pohlaví dosahuje vysoké úspěšnosti a není problém jej aplikovat například do systému pro rozpoznání emočních stavů. Příklad takového systému je zobrazen na obrázku 6.1. Systém pro rozpoznání pohlaví využívá většinou příznaků, které jsou použity k samotnému rozpoznání emočních stavů, tím se ušetří hlavně výpočetní náročnost.

Nejčastěji užívané příznaky pro rozpoznání pohlaví jsou základní tón, frekvence formantů, energie, MFCC koeficienty [3].

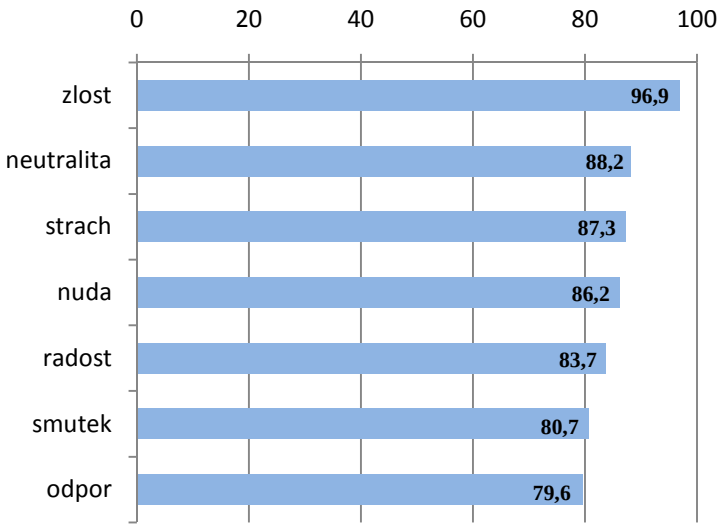


Obr. 6.1: Upravený systém pro rozpoznání.

7 POUŽITÁ DATABÁZE

Jako zdroj audio nahrávek pro semestrální projekt jsme použili Berlínskou databázi nahrávek emočních stavů (Berlin Database of Emotional Speech) [13]. Databáze vznikla jako část výzkumného projektu DFG mezi lety 1997-1999 a byla nahrána v bezodrazové komoře na technické univerzitě v Berlíně. Databáze obsahuje celkem 535 nahrávek a namluvilo ji 10 mluvčí (5 mužů a 5 žen). Databázi tvoří 10 různých vět (5 krátkých a 5 delších) a 7 různých emocí (strach, odpor, radost, nuda, neutralita, smutek, zlost) [13].

Pro zajištění kvality a přirozenosti nahrávek byl uskutečněný subjektivní test databáze. Testu se účastnilo 20 osob, které měli povoleno poslouchat každou nahrávku pouze jednou. Před samotným rozhodnutím o emočním stavu mluvčího museli navíc uvést, jak moc přesvědčivý přednes byl. Na obrázku 7.1 jsou zobrazeny průměrné úspěšnosti rozpoznání pro jednotlivé emoční stavy v procentech [18]. V tabulce 7.1 je uveden počet nahrávek pro jednotlivé emoce.



Obr. 7.1: Úspěšnost subjektivních testů pro jednotlivé emoční stavy

Tab. 7.1: Počet nahrávek pro jednotlivé emoční stavy

emoce	strach	odpor	radost	nuda	neutralita	smutek	zlost
nahrávek	69	46	71	81	79	62	127

8 VÝSLEDKY

8.1 Tempo řeči

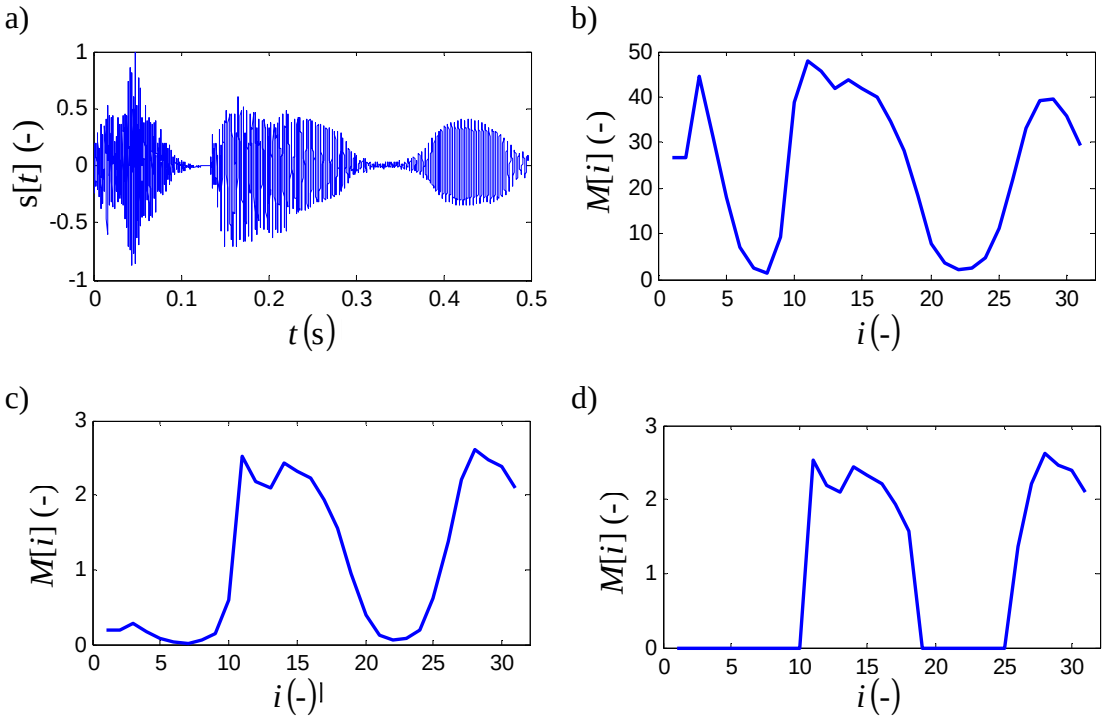
Při zjišťování tempa řeči jsme vycházeli z poznatků v literatuře [2]. Nejvýznamnější informace o rychlosti promluvy nese slabika, která je dále tvořena kombinací samohlásek a souhlásek, kdy na délku slabiky mají největší vliv samohlásky a znělé souhlásky. K oddělení znělých a neznělých segmentů řeči jsme využili rovnice (3.2) pro výpočet krátkodobé intenzity. Dle [2] má signál neznělý daleko více průchodů nulou, než signál znělý, proto pro větší potlačení neznělých segmentů bylo využito počtu průchodu signálu nulou na daném segmentu. Na obrázku 8.1b je znázorněna krátkodobá intenzita pro jednotlivé segmenty *i* z promluvy. Na obrázku 8.1c je zobrazeno potlačení krátkodobé intenzity v závislosti na průchodu signálu nulovou rovinou.

Dále bylo potřeba určit hranici znělosti, která by sloužila pro oddělení intenzity znělých a neznělých částí promluvy. Velikost této hranice jsme určili jako střední hodnotu z hodnot intenzit jednotlivých segmentů, násobenou konstantou k = 0,9. Tuto konstantu jsme určili empiricky a lze dle potřeby snižovat, ale to za cenu možné detekce i neznělé části promluvy. Segmenty *i,* jejichž maximum nedosahovalo na hranicí znělosti, byly určeny jako neznělé a vynulovány, viz obr. 8.1d.

Z takto známého průběhu intenzity a známé délky trvání jednoho segmentu jsme vypočítali celkovou délku trvání znělé promluvy t_Z. Z průběhu intenzity jsme také určili počet vrcholů V, tím tedy počet samohlásek a znělých souhlásek v promluvě. Bohužel z průběhu krátkodobé intenzity nelze rozlišit dvě za sebou jdoucí kombinace znělých souhlásek a samohlásek, jako např. „am“. Dle [2] se maximální délka trvání fonému pohybuje na hranici 130ms pro dvojhlásky. Této hranice jsme využili a nastavili tak maximální délku za sebou jdoucích znělých segmentů pro detekci jednoho znělého fonému. Výsledné průměrná doba trvání jednoho znělého fonému (průměrné tempo řeči) je tedy

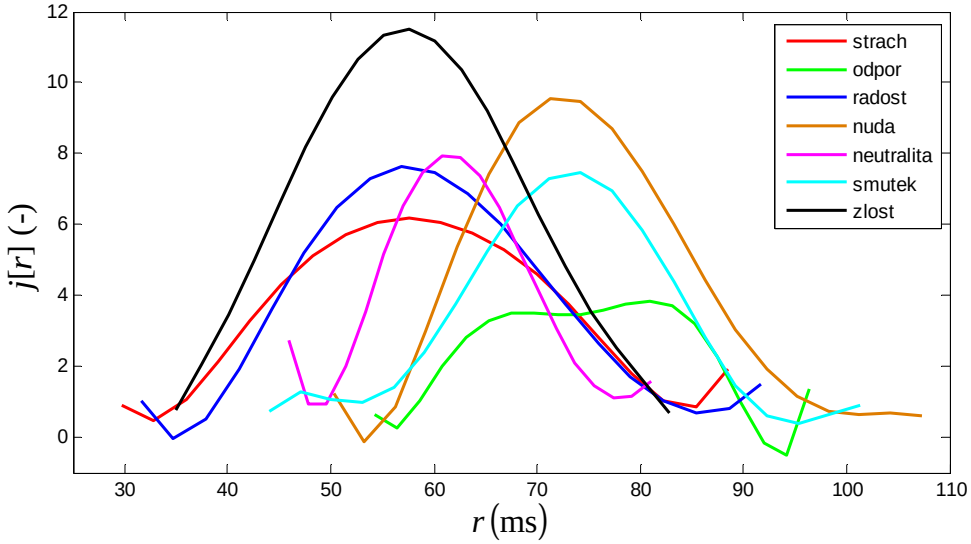
$$r = \frac{t_Z}{V} \quad (\mathrm{ms}). \tag{8.1}$$

Na obrázku 8.2 jsou zobrazeny histogramy průměrného tempa řeči *r* pro jednotlivé emoční stavy. Osa y značí počet nahrávek *j* odpovídajících danému tempu.



Obr. 8.1: Průběh slova „stunden”: a) řečový signál, b) krátkodobá intenzita, c) upravená krátkodobá intenzita, d) omezená krátkodobá intenzita.

Pomocí tempa řeči lze oddělit emoční stavy pasivní a aktivní. Mezi pasivní patří smutek, nuda, odpor a mezi aktivní zlost, radost, strach. Při výpočtu tempa řeči jsme použili délku rámce 30ms a překrytí rámců 50%.



Obr. 8.2: Histogramy průměrného tempa řeči pro jednotlivé emoční stavy.

8.2 MFCC, HFCC

Pro výpočet MFCC koeficientů jsme vytvořili funkci dle kapitoly 3.4. Průběh výpočtu MFCC koeficientů segmentu promluvy je zobrazen na obrázku 3.2. Počet MFCC koeficientů se nejčastěji používá v rozmezí 10-13 [2]. V této práci jsme používali 12 MFCC koeficientů.

Výpočet HFCC koeficientů se od výpočtu MFCC koeficientů liší pouze v použití rozdílné banky filtrů, viz obr. 3.3. Pro výpočet HFCC koeficientů jsme upravili banku filtrů dle kapitoly 3.5. Porovnáním průběhů HFCC a MFCC koeficientů pro segment promluvy je vidět na obrázku 8.3c, kde *i* značí index koeficientu.

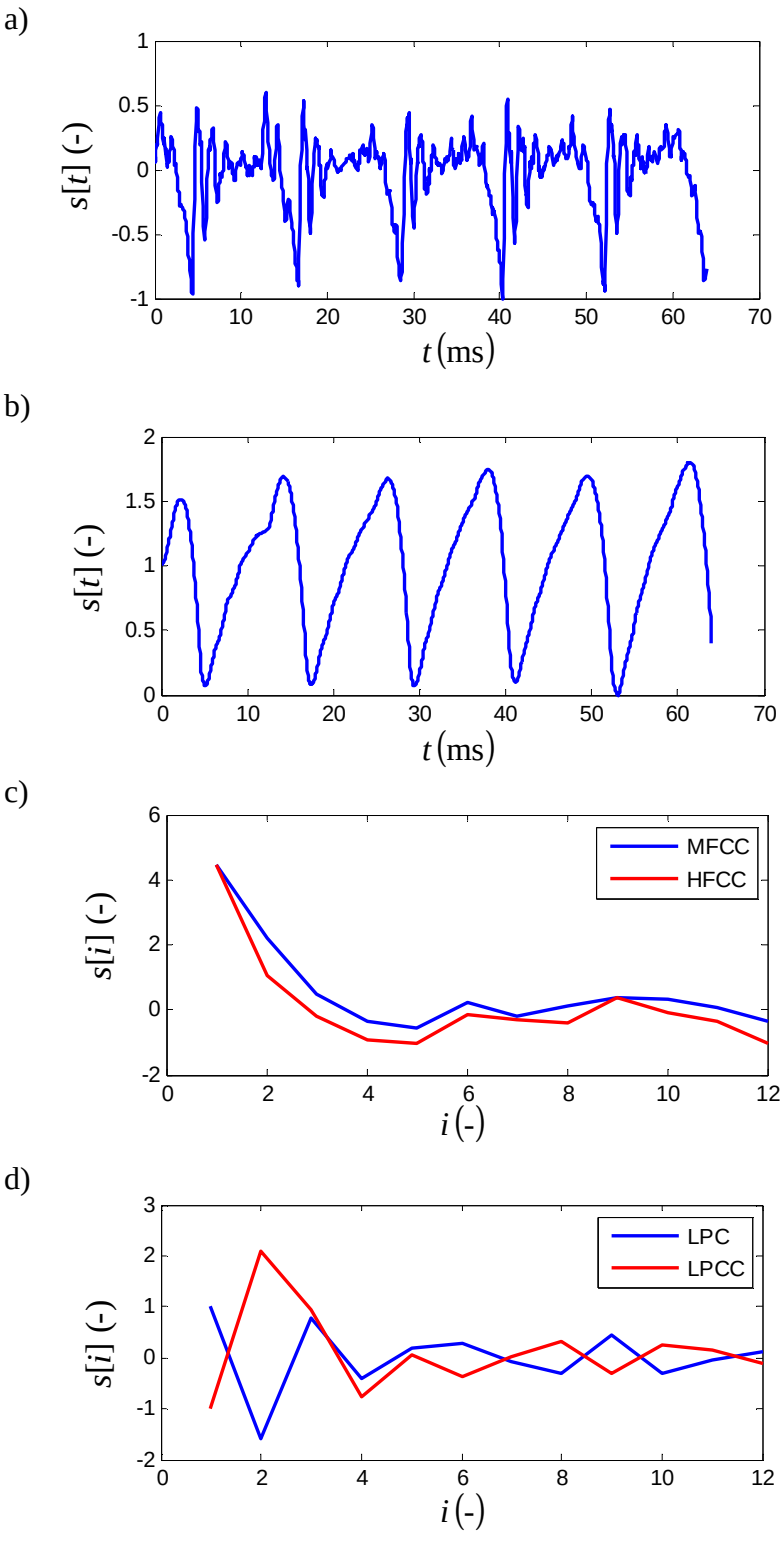
8.3 LPC, LPCC

Pro výpočet LPC koeficientů jsme vytvořili funkci dle kapitoly 3.6 a poté funkci upravili pro výpočet LPCC koeficientů dle rovnice (3.26). Porovnání průběhu LPC a LPCC je vidět na obrázku 8.3d, kde *i* značí index koeficientu.

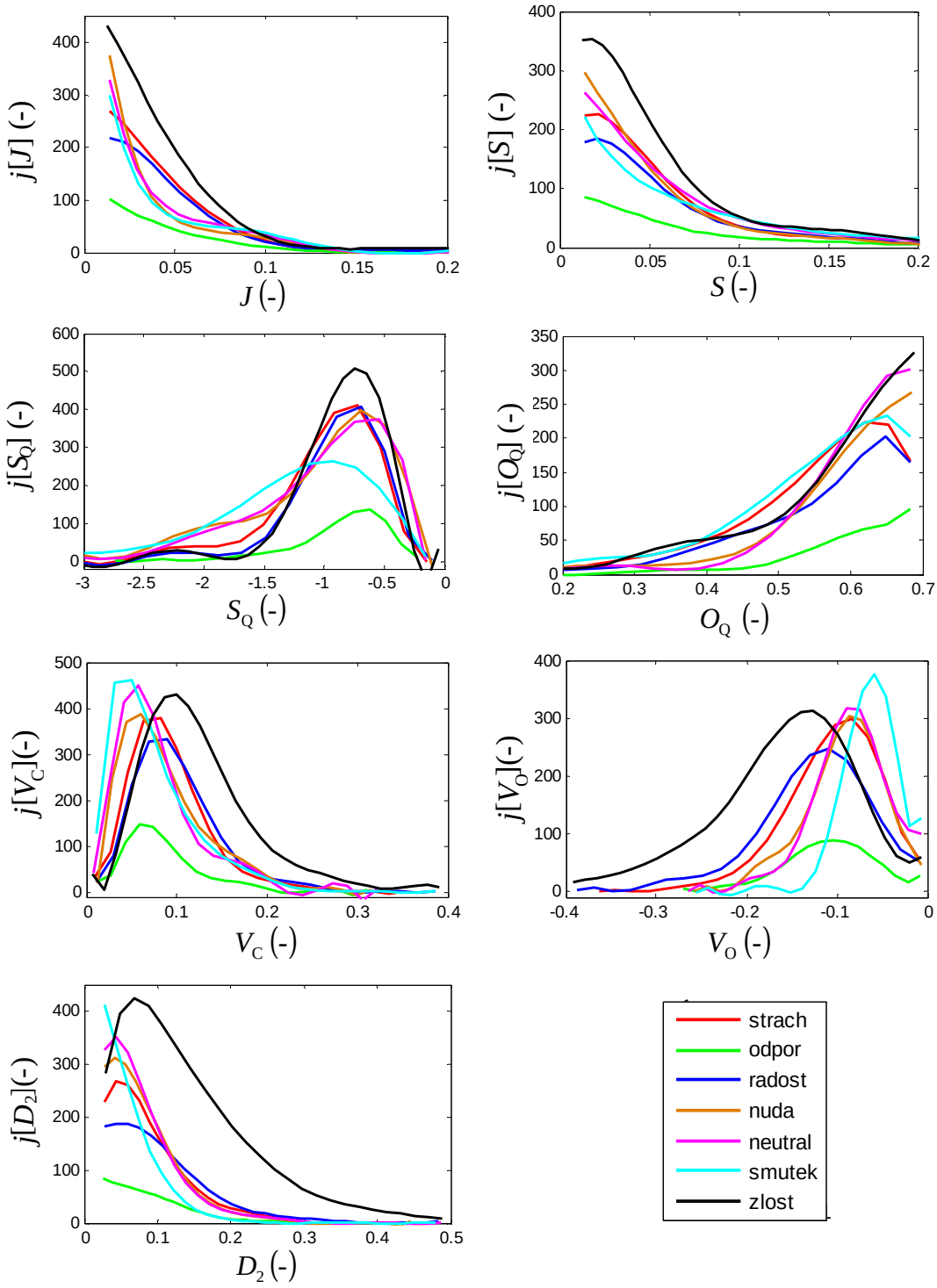
8.4 Hlasivkové pulsy

Pro odhad hlasivkových pulsů jsme využili metody IAIF. Dle znalostí z kapitoly 4.2.1, jsme navrhli funkci pro odhad hlasivkových pulsů. V rámci této metody je potřeba zjistit základní tón segmentu promluvy, k tomu bylo použito metody centrálního klipování, viz kap. 3.3.2. Pro výpočet parametr; hlasového traktu pomocí DAP metody jsme využily volně dostupnou sadu funkcí MATSIG [16]. Na obrázku 8.3b je vidět odhadnutý průběh hlasivkových pulsů ze segmentu promluvy.

Pro výpočet příznaků uvedených v tabulce 4.1 jsme vytvořili skript a aplikovali ho na použitou databázi nahrávek z důvodu zjištění vlivu emočních stavů na jednotlivé příznaky. Na obrázku 8.4 jsou zobrazeny histogramy příznaků pro jednotlivé emoční stavy, osa y zde značí počet segmentů *j*.



Obr. 8.3: a) Segment řeči, b) odhad hlasivkových pulsů, c) průběh MFCC a HFCC koeficientů, d) průběh LPC a LPCC koeficientů.



Obr. 8.4: Histogramy sledovaných příznaků na hlasivkových pulsech pro jednotlivé emoční stavy.

8.5 Klasifikace

Za účelem klasifikace byla použita volně dostupná sada funkcí NETLAB pro GMM klasifikátor [14].

Možné způsoby klasifikace:

- nezávislá na mluvčím (speaker independent): Je zapotřebí použít jiných mluvčí pro testování než pro trénování.
- závislá na mluvčím (speaker dependent): Není zapotřebí trénovat a testovat rozdílnými mluvčími.
- nezávislá na pohlaví (gender independent): Klasifikace probíhá pro muže a ženy společně.
- závislá na pohlaví (gender dependent): Klasifikace probíhá odděleně pro muže a ženy

Aby bylo zabráněno vlivu mluvčího na výsledky, použili jsme při klasifikacích metodu výměny mluvčích (leave 2-speaker out), kdy testovací skupina obsahuje vždy dva mluvčí a při další klasifikaci jednotlivé mluvčí v testovacích a trénovacích skupinách střídáme. Výsledná úspěšnost je dána průměrem z jednotlivých klasifikací. Při klasifikaci na celé databázi nahrávek jsme nechali pro testování vždy dva mluvčí (muže a ženu), pro oddělené databáze mužů a žen testovací skupina obsahovala vždy jednoho mluvčího. Výsledky jsou zobrazeny pomocí matice záměn (confusion matrix). Zde řádky značí zkoumaný emoční stav a sloupce rozpoznaný emoční stav systémem. Hodnoty v těchto tabulkách jsou uváděny v procentech.

8.5.1 GMM klasifikátor

Jak je uvedeno v kapitole 5.1, klasifikátor GMM je tvořen lineární kombinací Gaussových funkcí. V několika testech jsme zkoušeli, jak moc se počet těchto funkcí projeví na úspěšnosti klasifikace. Test jsme provedli pro různé délky segmentů promluvy a pro MFCC, HFCC, LPC a LPCC koeficienty. Výsledek je uveden v tabulce 8.1. Na obrázku 8.5 a 8.6 jsou výsledky pro přehlednost vyneseny do grafu.

Tab. 8.1: Vliv počtu Gaussových funkcí na úspěšnost klasifikace pro různé délky segmentů promluvy: a) MFCC koeficienty, b) HFCC koeficienty, c) LPC koeficienty, d) LPCC koeficienty

a)

počet funkcí	délka segmentu [ms]		
	32	64	128
3	12	20	38
5	20	24	40
10	21	30	43
15	28	31	41
20	26	37	43
25	30	35	45
30	32	33	41
35	31	35	42
45	32	37	37
55	30	35	39
65	33	37	39
75	30	34	38
85	33	30	37
95	31	34	38

b)

počet funkcí	délka segmentu [ms]		
	32	64	128
3	17	28	37
5	23	29	43
10	24	33	40
15	28	31	44
20	30	34	41
25	29	33	43
30	29	35	45
35	30	33	39
45	33	34	41
55	28	41	45
65	32	34	41
75	29	36	46
85	27	38	39
95	31	37	39

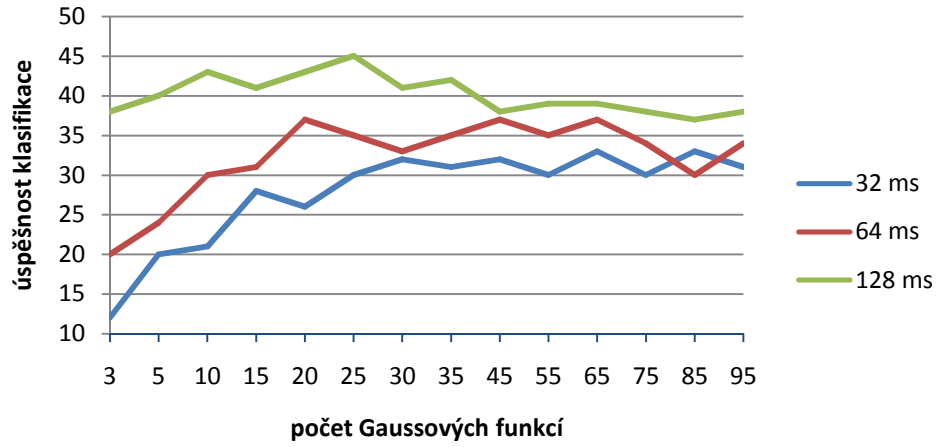
c)

počet funkcí	délka segmentu [ms]		
	32	64	128
3	11	12	19
5	12	12	22
10	11	15	22
15	13	14	21
20	11	19	22
25	11	15	23
30	12	16	24
35	12	19	26
45	11	18	23
55	11	17	23
65	13	18	26
75	16	18	26
85	14	17	24
95	14	19	30

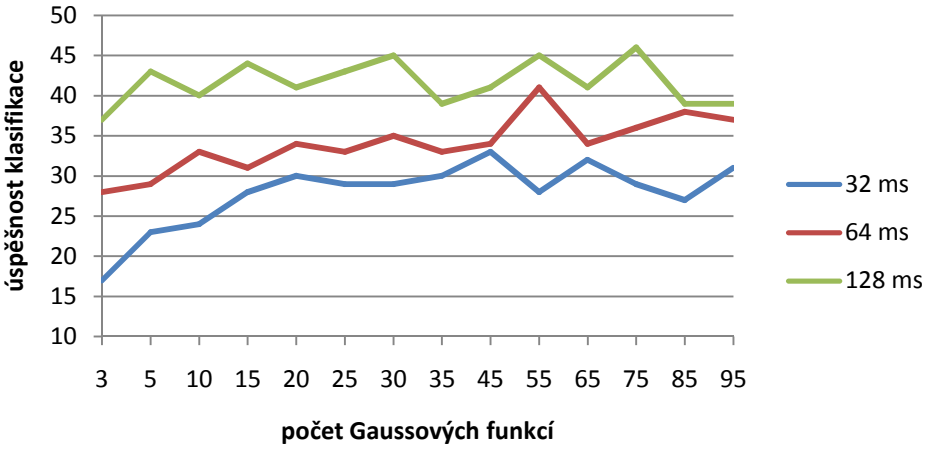
d)

počet funkcí	délka segmentu [ms]		
	32	64	128
3	6	15	14
5	10	13	13
10	11	11	15
15	12	12	20
20	11	12	20
25	11	14	20
30	11	13	21
35	11	13	24
45	12	13	26
55	13	17	27
65	11	15	23
75	12	17	27
85	12	15	28
95	12	17	22

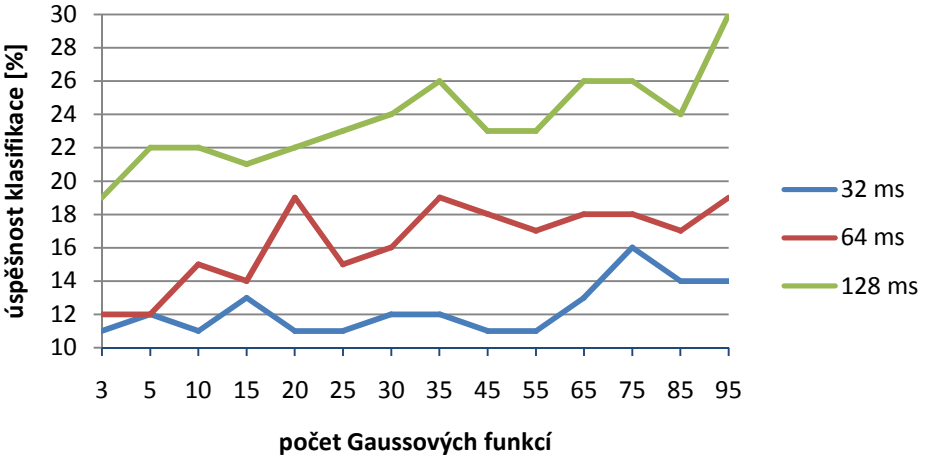
a)



b)

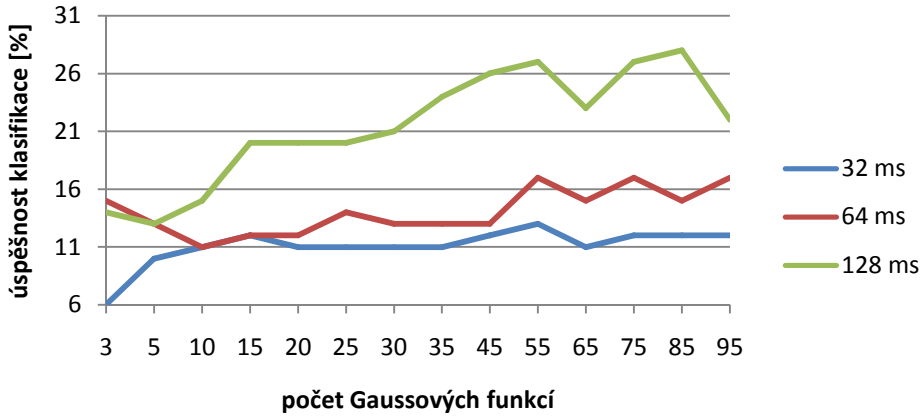


c)

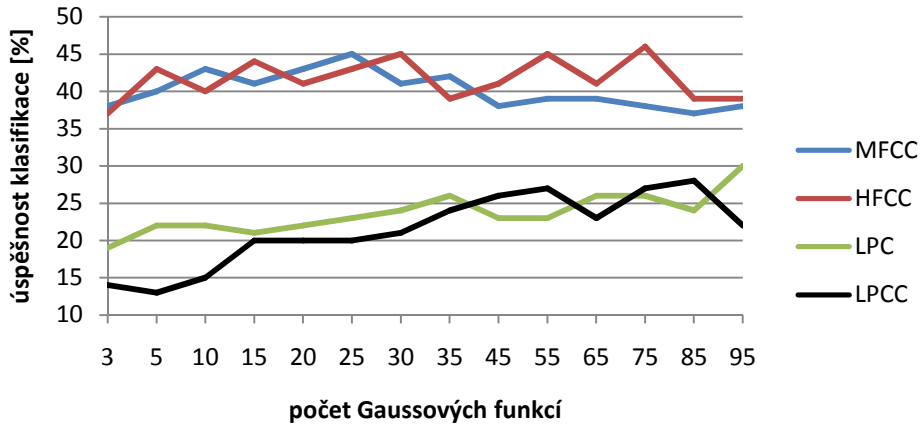


Obr. 8.5: Závislost úspěšnosti klasifikace na počtu Gaussových funkcí: a) MFCC koeficienty, b) HFCC koeficienty, c) LPC koeficienty.

d)



d)



Obr. 8.6: Závislost úspěšnosti klasifikace na počtu Gaussových funkcí: a) LPCC koeficienty, b) porovnání.

8.5.2 Nezávislá na mluvčím

Trénovácí skupina obsahovala vždy čtyři muže a ženy, testovací vždy jednoho muže a ženu. Úspěšnost klasifikace nezávislé na pohlaví pro 7 emocí byla 38%, viz tab. 8.2. Při klasifikaci v závislosti na pohlaví obsahovala trénovací skupina vždy čtyři muže (ženy) a testovací skupina vždy jednoho muže (ženu). Průměrná úspěšnost klasifikace na databázi mužů byla 47%, viz tab. 8.3. Průměrná úspěšnost klasifikace na databázi žen byla 44%, viz tab. 8.4.

Tab. 8.2: Úspěšnost klasifikace nezávisle na pohlaví.

	zlo	nud	odp	str	rad	smu	neu
zlo	**52**	4	9	10	21	1	3
nud	1	**21**	7	6	2	38	25
odp	3	12	**24**	16	8	28	9
str	6	10	10	**37**	5	15	17
rad	20	6	16	18	**31**	3	6
smu	0	6	2	0	0	**89**	3
neu	0	24	7	8	3	30	**28**

Tab. 8.3: Úspěšnost klasifikace na databázi mužů.

	zlo	nud	odp	str	rad	smu	neu
zlo	**68**	0	9	5	15	2	2
nud	0	**34**	6	3	6	26	26
odp	18	9	**8**	18	27	19	0
str	3	17	6	**39**	8	17	11
rad	26	4	11	15	**26**	4	15
smu	0	0	4	0	0	**84**	12
neu	3	31	3	13	3	18	**31**

Tab. 8.4: Úspěšnost klasifikace na databázi žen.

	zlo	nud	odp	str	rad	smu	neu
zlo	**54**	1	12	4	25	0	3
nud	2	**33**	0	13	4	20	28
odp	3	9	**31**	14	14	23	6
str	12	12	15	**33**	12	3	12
rad	32	5	16	14	**32**	2	0
smu	0	8	8	0	0	**81**	3
neu	0	33	10	13	0	5	**40**

8.5.3 Závislá na mluvčím

Zde trénujeme všemi mluvčími a testujeme vždy jen částí z nich. Průměrná úspěšnost klasifikace nezávisle na pohlaví byla 78%, viz tab. 8.5. Průměrná úspěšnost klasifikace na databázi mužů byla 94%, viz tab. 8.6. Průměrná úspěšnost klasifikace na databázi žen 89%, viz tab. 8.7.

Tab. 8.5: Úspěšnost klasifikace nezávisle na pohlaví.

	zlo	nud	odp	str	rad	smu	neu
zlo	**78**	0	5	6	8	1	2
nud	0	**75**	4	2	0	9	9
odp	0	3	**70**	2	0	22	3
str	0	3	2	**81**	1	9	4
rad	4	4	6	6	**74**	4	2
smu	0	2	0	0	0	**98**	1
neu	0	13	1	1	0	10	**76**

Tab. 8.6: Úspěšnost klasifikace na databázi mužů.

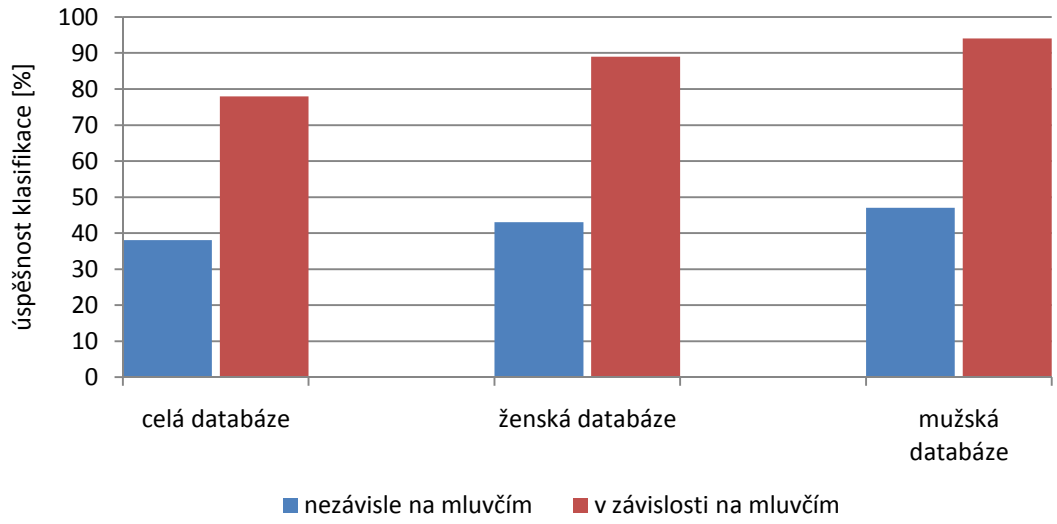
	zlo	nud	odp	str	rad	smu	neu
zlo	**88**	0	3	2	3	2	2
nud	0	**94**	0	0	0	3	3
odp	0	0	**100**	0	0	0	0
str	0	0	0	**97**	4	3	0
rad	0	0	0	4	**93**	4	0
smu	0	0	0	0	0	**100**	0
neu	0	0	0	0	0	3	**97**

Tab. 8.7: Úspěšnost klasifikace na databázi žen.

	zlo	nud	odp	str	rad	smu	neu
zlo	**93**	1	3	1	1	0	0
nud	0	**87**	2	2	0	7	2
odp	0	0	**83**	3	0	11	3
str	0	0	0	**91**	0	6	3
rad	2	2	2	7	**80**	0	7
smu	0	0	3	0	0	**97**	0
neu	0	5	0	0	0	3	**93**

8.5.4 Porovnání

Na obrázku 8.7 jsou pro přehlednost shrnuty jednotlivé úspěšnosti klasifikace. Je zde vidět, že klasifikace v závislosti na mluvčím má daleko vyšší úspěšnost při rozhodování o emočním stavu mluvčího, ale to pouze za cenu toho, že příznaky od mluvčího jsou obsaženy již v trénovací datech. Je zde také vidět, že pokud se klasifikace rozdělí pro muže a ženy odděleně je dosaženo vyšší úspěšnosti klasifikace.



Obr. 8.7: Úspěšnost klasifikace dle závislosti na mluvčím.

8.5.5 Rozpoznání pohlaví

Pro systém rozpoznání pohlaví jsme využili příznaků: MFCC koeficienty, základní tón F_0 a frekvence tří prvních formantů F_1, F_2, F_3. Celkem jsme sestrojili tři systémy rozpoznání pohlaví. První systém využíval MFCC koeficienty, druhý systém F_0, F_1, F_2, F_3, třetí systém MFCC společně s F_0, F_1, F_2, F_3. Úspěšnost rozpoznání pohlaví v závislosti na emočním stavu je uveden pro jednotlivé systémy v tabulkách 8.8, 8.9, 8.10.

Při rozpoznání pohlaví nezávisle na emočním stavu byla průměrná úspěšnost 80%.

Tab. 8.8: Systém 1, úspěšnost rozpoznání pohlaví.

	zlo	nud	odp	str	rad	smu	neu
muži	93	97	63	94	88	88	81
ženy	92	76	65	75	93	78	87
celkem	**92,5**	**86,5**	**64**	**84,5**	**90,5**	**83**	**84**

Tab. 8.9: Systém 2, úspěšnost rozpoznání pohlaví.

	zlo	nud	odp	str	rad	smu	neu
muži	68	72	72	77	78	70	89
ženy	96	95	88	87	95	86	97
celkem	**81,5**	**83,5**	**80**	**82**	**86,5**	**78**	**93**

Tab. 8.10: Systém 3, úspěšnost rozpoznání pohlaví.

	zlo	nud	odp	str	rad	smu	neu
muži	91	82	40	82	84	68	84
ženy	80	65	72	78	81	75	64
celkem	**65,5**	**73,5**	**56**	**69**	**82,5**	**71,5**	**74**

8.5.6 Testy na hlasivkových pulsech

Pro zjištění nejlepších příznaků jsme použily funkci „*rankfeatures*“. Tato funkce seřadí příznaky dle jejich oddělitelnosti mezi třídami od nejlepšího po nejhorší. Funkce je již součástí programu MATLAB.

Příznaky s nejlepší oddělitelností mezi jednotlivými emočními stavy patří V_O, V_C, a D_2. Tyto tři příznaky jsme použili pro následnou klasifikaci. Na obrázku Obr. 8.4 je vidět, že u ostatních příznaků se histogramy překrývají a pro klasifikaci nám dané příznaky nevyhovují. Dle histogramů příznaků V_O, V_C a D_2 je možné z části oddělit emoční stavy pasivní a aktivní. V rámci těchto skupin jsou emoční stavy těmito příznaky neoddělitelné, viz tab. 8.11. V tabulce 8.12 je uvedena úspěšnost klasifikace pro dva emoční stavy.

Tab. 8.11: Úspěšnost klasifikace na databázi mužů pro 7 emočních stavů.

	zlo	nud	odp	str	rad	smu	neu
zlo	**18**	28	18	9	9	18	0
nud	20	**40**	10	10	10	0	10
odp	0	17	**13**	30	10	13	17
str	50	33	0	**0**	17	0	0
rad	37	27	0	10	**4**	4	18
smu	0	5	0	33	0	**0**	72
neu	21	18	5	12	9	5	**30**

Tab. 8.12: Úspěšnost klasifikace na databázi mužů pro 2 emoční stavy.

	zlo	smu
zlo	**71**	29
smu	20	**80**

9 ZÁVĚR

V diplomové práci jsme se seznámili se základními vlastnostmi a různými metodami analýzy řečového signálu. V dnešní výpočetní době se stále více rozmáhá odvětví počítačové analýzy řečového signálu, protože pomocí počítačové analýzy je možné sledovat vlastnosti a změny hlasového traktu, které jsou pouhým uchem nepostřehnutelné. Počítačem zpracovávaný řečový signál je možno využít v mnoha odvětví např. v komunikacích pro kódování řeči, zdravotnictví pro rozpoznání logopedických vad malých dětí a také v zabezpečovací technice při identifikaci mluvčího.

Za účelem klasifikace byla použita volně dostupná sada funkcí NETLAB [14] pro GMM klasifikátor využívající Gaussova rozložení pravděpodobnosti a testy zde probíhané byly na audio nahrávkách z Berlínské databáze emocí [13].

Při práci jsme se zaměřili na analýzu několika příznaků. Tempo řeči nám popisuje, jak člověk mluví rychle. Pro tempo řeči jsme vytvořili algoritmus, jenž spočítá průměrnou dobu trvání znělých souhlásek a samohlásek v promluvě. Dle histogramů průměrného tempa řeči, pro jednotlivé emoční stavy, je možné oddělit emoční stavy pasivní a aktivní, viz obr.Obr. 8.2.

Dalším analyzovaným příznakem jsou v této práci MFCC koeficienty. Zpracování pomocí MFCC je navrženo tak, aby do jisté míry respektovalo nelineární vnímání frekvencí lidským uchem a to využitím banky trojúhelníkových pásmových filtrů. Nejnovější úpravou MFCC koeficientů jsou HFCC koeficienty, které mají jiné rozmístění banky pásmových filtrů, viz obr. 3.3. Pro výpočet MFCC a HFCC koeficientů jsme vytvořili vlastní funkce. Dále jsme vytvořili funkce pro výpočet LPC a LPCC koeficientů, které jsme otestovali na emoční databázi nahrávek.

Na obrázku Obr. 8.6b. je zobrazeno porovnání MFCC, HFCC, LPC a LPCC koeficientů. Je zde vidět, že MFCC a HFCC koeficienty jsou vhodnějším příznakem pro rozpoznání emočního stavu. Při porovnání MFCC koeficientů a novějších HFCC koeficientů se hodnoty úspěšnosti klasifikací nijak výrazně nelišily. Ani mezi LPC a LPCC koeficienty není patrný rozdíl úspěšnosti klasifikace. Při porovnání úspěšnosti klasifikace zvlášť pro jednotlivé koeficienty je vidět, že při použití delšího segmentu promluvy je úspěšnost klasifikace vyšší, což vypovídá o tom, že emoční stav jako celek je příznakem spíše suprasegmentálním, viz obr. Obr. 8.5, Obr. 8.6a.

Úspěšnost klasifikace je do jisté míry dána i věrohodností počátečních odhadů klustrů *k*-means algoritmem. Pokud odhadne nepřesné hodnoty je poté i odhad za pomoci EM algoritmu méně přesný. Jedna z věcí, která má při klasifikaci s GMM modely velký podíl na úspěšnosti výsledku je počet použitých Gaussových funkcí pro modelování příznakového prostoru. Na obrázku Obr. 8.5 je vidět, že za pomoci malého počtu Gaussových funkcí není možné příznakový prostor dostatečně modelovat a úspěšnost klasifikace je tak malá, při použití vyššího počtu Gaussových funkcí se úspěšnost klasifikace pohybuje na stejné úrovni, a dalším přidáváním funkcí jenom zvyšujeme výpočetní náročnost, proto je nejvhodnější volit počet Gaussových funkcí maximálně 2 krát vyšší než je velikost příznakového prostoru.

Pro další testování jsme si vybrali pouze MFCC koeficienty. První testy probíhaly nezávisle na mluvčím. Základem je použití trénovací databáze, která nebude obsahovat testovací mluvčí. Pro zabránění vlivu mluvčí na výsledky bylo využito testování metodou „leave 2-speaker out“. Průměrná úspěšnost klasifikace nezávislé na pohlaví byla 38%. Při klasifikaci v závislosti na pohlaví bylo dosaženo úspěšnosti 47% pro databázi mužů a 44% pro databázi žen.

Další testy probíhaly v závislosti mluvčím. Zde využíváme pro trénování všech mluvčí a pro testování vždy jen část z nich. Úspěšnost klasifikace na celé databázi byla 78%. Při oddělené klasifikaci na mužské databázi bylo dosaženo úspěšnosti 94% a 89% na databázi žen. Zde vidět, že testy v závislosti na mluvčím mají vyšší úspěšnost při rozhodování o emočním stavu mluvčího, ale to za cenu toho, že příznaky od testovaného mluvčího jsou obsaženy v trénovacích datech. Této skutečnosti se využívá například v zabezpečovacích systémech při identifikaci a verifikaci mluvčího. Při oddělené klasifikaci pro muže a ženy je dosaženo vyšší úspěšnosti klasifikace.

V rámci diplomové práce jsme také sestrojili tři systémy pro rozpoznání pohlaví pracující s kombinací příznaků: MFCC koeficienty, základní tón a frekvence tří prvních formantů. Úspěšnost sestrojených systému jsme otestovali na databázi nahrávek metodou „leave 1-speaker out“, viz tab. 8.8, 8.9, 8.10.

Dle kapitoly 4.2.1 jsme sestrojily funkci pro výpočet hlasivkových pulsů a zkoumali vliv emočních stavů na příznaky z tabulky 4.1. Dle výsledků na obrázku 8.4 příznaky neoddělují dostatečně jednotlivé emoční stavy a pro jejich rozpoznání nám nevyhovují, viz tab. 8.11.

LITERATURA

[1] KUKAČKA, M. *Modely emocí pro autonomní agenty a umělé bytosti.* Dostupné z*:* <http://ksvi.mff.cuni.cz/~brom/eseje/Emoce2005.pdf>

[2] PSUTKA, J., RAIDA, L., MATOUŠEK, J., RADOVÁ, V. *Mluvíme s počítačem česky.* 1. vydání. Praha: Academia 2006.

[3] VOGT, T., ÁNDÉ, E. *Improving Automatic Emotion Recognition from Speech via Gender Differentiation.* Dostupné z: < http://mm-werkstatt.informatik.uni-augsburg.de/files/publications/124/lrec06.pdf >

[4] REYNOLDS, A., THOMAS, F., DUNN, R. *Speaker verification using Adapted Gaussian mixture models*. Digital Signal Processing 2000.

[5] Laboratoř zpracování přirozeného jazyka. *Skryté Markovovy modely.* Dostupné z: <http://nlp.fi.muni.cz/nlp/nlp-prace/referaty/xkrivan/HMM.html>

[6] WIKIPEDIA. *Support vector machines.* Dostupné z: <http://cs.wikipedia.org/wiki/Support_vector_machines>

[7] SCIENCEWORLD. *Co jsou to neuronové sítě.* Dostupné z: <http://scienceworld.cz/technologie/co-jsou-to-umele-neuronove-site-4077>

[8] KRČMOVÁ, M. *Fonetika.* Dostupné z: <http://is.muni.cz/elportal/estud/ff/js07/fonetika/materialy/index.html>

[9] ATASSI, H. *Metody detekce základního tónu řeči.* Elektrorevue 2008. Dostupné z: <http://www.elektrorevue.cz/cz/clanky/zpracovani-signalu/0/metody-detekce-zakladniho-tonu-reci/>

[10] GANCHEV, T., FAAKOTAKIS, N., KOKKINAKIS, G. *Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task.* Dostupné z: < http://www.wcl.ee.upatras.gr/ai/papers/ganchev17.pdf >

[11] PULAKKA, H. *Analysis of Human Voice Production Using Inverse Filtering, High-Speed Imaging, and Electroglottography.* Helsinki: University of Technology, 2005. Dostupné z: < http://recherche.ircam.fr/equipes/analyse-synthese/beller/doc/degottex/pulakka_mst_2003_Analysis_Voice_production_inverse_filter_egg.pdf >

[12] MURPHY, P., LAUKKANEN, AM. Electroglottogram Analysis of Emotionally Styled Phonation.*Multimodal Signals: Cognitive and Algorithmic Issues.* Springer, Berlin. pp. 264-270, 2008

[13] Berlin Database of Emotional Speech.
Dostupné z: <http://pascal.kgw.tu-berlin.de/emodb/>

[14] Netlab neural network software.
Dostupné z: <http://www.ncrg.aston.ac.uk/netlab/index.php>

[15] VILLAVICENCIO. F., RÖBEL, A., RODET, X. *All-Pole Spectral Envelope Modelling with Order Selection for Harmonic Signals* Dostupné z: <http://mediatheque.ircam.fr/articles/textes/Villavicencio07a/>

[16] Matsig - An object-oriented signal processing class library for MATLAB.
Dostupné z <http://matsig.sourceforge.net/>

[17] HOLUBEC, B. *Použití PSO metody pro selekci příznaků*. FEL ČVUT v Praze, 2008. 65s. Vedoucí diplomové práce: Lhotská Lenka Doc.Ing., CSc.
Dostupné z <https://dip.felk.cvut.cz/browse/pdfcache/holubb1_2008dipl.pdf>

[18] BURKHARDT, F., PAESCHKE, A., ROLFES, M., SENDLMEIER, W., WEISS, B. *A Database of German Emotional* Proceedings Interspeech 2005, Lissabon, Portugal.
Dostupné z < http://pascal.kgw.tu-berlin.de/home/publications/p1078.pdf>

SEZNAM ZKRATEK, VELIČIN, A SYMBOLŮ

ANN	Artificial Neural Network (umělé neuronové sítě)
ACF	Autocorrelation Function (autokorelační funkce)
DCT	Discrete Cosine Transform (diskrétní kosinová transformace)
FFT	Fast Fourier Transform (rychlá Fourierova transformace)
GMM	Gaussian Mixture model (Gaussovy smíšené modely)
HFCC	Human Factor Cepstral Coefficients (kepstrální koeficienty typu HF)
HMM	Hidden Markov Model (skryté Markovovy modely)
IFFT	Inverse Fast Fourier Transform (inverzní rychlá Fourierova transformace)
KNN	*K*-Nearest Neighbor (algoritmus *k*-nejbližšího souseda)
LPC	Linear Predictive Coding (lineární predikční kódování)
PLP	Perceptual Linear Predictive (perceptivní lineární prediktivní analýza)
MFCC	Mel-Frequency Cepstral Coefficients (mel-frekvenční kepstrální koeficienty)
SVM	Support Vector Machine (algoritmus podpůrných vektorů)
α	váhovací paramentr
Σ	kovarianční matice
μ	vector středních hodnot
a_i	*i*-tý koeficient lineární predikce
A_n	*n*-tý antiformant
$c[m]$	*m*-tý mel-frekvenční kepstrální koeficient
$c_{\mathrm{H}}[m]$	*m*-tý kepstrální koeficient typu HF
$c_{\mathrm{L}}[m]$	*m*-tý kepstrální koeficient lineární predikce
D_2	akcelerace kontaktu
E	krátkodobá energie
ERB_i	šířka *i*-tého filtru v bance typu HF
$f(\mathrm{x})$	vícerozměrná Gaussova funkce rozložení pravděpodobnosti
f_{lin}	frekvence v lineární škále
$\hat{f}_{\mathrm{mel}}$	frekvence v melovské škále
f_{L}	nejnižší hranice banky filtrů

f_{H}	nejvyšší hranice banky filtrů
f_{b_i}	střední frekvence *i*-tého melovského filtru v lineární skále
f_{c_i}	střední frekvence *i*-tého filtru v bance typu HF v lineární skále
F_n	*n*-tý formant
F_0	frekvence základního hlasivkového tónu
G	koeficient zesílení
J	Jitter (chvění)
k	redukční faktor
L	délka periody ve vzorcích (lag)
M	krátkodobá intenzita
M_{b}	počet bank filtrů
M_{g}	počet Gaussových funkcí rozložení pravděpodobnosti
O_{Q}	Opening Quotient (kvocient otevírání)
r	tempo řeči
$R[m]$	*m*-tý koeficient autokorelací funkce
S	Schimmer (výchylka)
S_{Q}	Speed Quotient (rychlostní kvocient)
$s[n]$	*n*-tý vzorek signálu
T_0	perioda základního hlasivkového tónu
t_{z}	délka trvání znělé promluvy
V	počet vrcholů
V_{C}	maximum Velocity of Contact (maximální rychlost otevírání)
V_{O}	maximum negative Velocity at opening (maximální rychlost uzavírání)
neu	neutralita
nud	nuda
odp	odpor
rad	radost
str	strach
smu	smutek
zlo	zlost

PŘÍLOHY: Obsah přiloženého CD

Přiložené CD obsahuje tři adresáře:

/nahravky

obsahuje roztříděnou databázi nahrávek

/text

obsahuje text diplomové práce ve formátu *.pdf

/funkce

obsahuje čtyři podadresáře

1. podadresář: **/hlasivkove_pulsy**

obsahuje m-funkce MATLABu potřebné pro testování na hlasivkových pulsech

2. podadresář: **/pomoc_fun**

obsahuje m-funkce MATLABu, které využívají vytvořené systémy rozpoznání

3. podadresář: **/systemy**

obsahuje m-funkce MATLABu vytvořených systémů rozpoznávání

4. podadresář: **/tempo_reci**

obsahuje m-funkce MATLABu potřebné pro výpočet tempa řeči