Obsah

Seznam obrázků:

Seznam tabulek:

Kapitola 1

1.Úvod

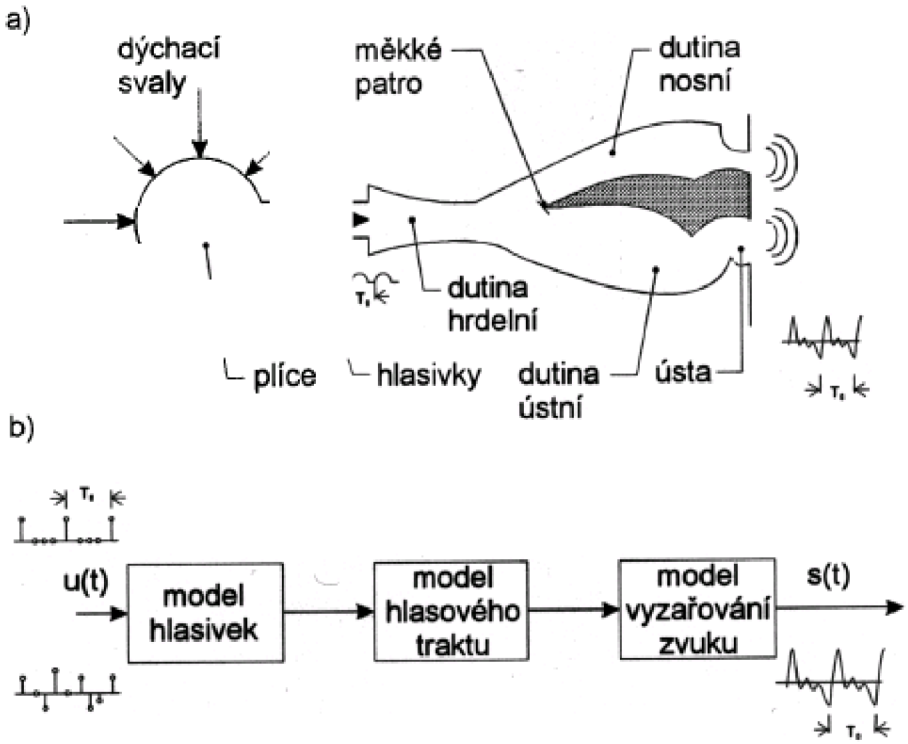
Analýza řečového signálu v dnešní době má veliký význam ve všech možných odvětví. Rozpoznávání a syntéza řeči stále více proniká do zařízení běžných potřeb, a také více nalézají uplatnění poznatky o vlastnostech řeči v nenormálním emočním stavu (např. vlivem stresu, únavy, apod.). Využití nachází například v aplikacích kde je zapotřebí upravit hlas do přirozenější podoby a dále například u hlasového zámku, kde je zapotřebí, aby zařízení ovládal člověk, který je v normálním emočním stavu.

V úvodní části se pokusím přiblížit problematiku analýzy řečového signálu. Dále zde rozvedu metody různých analýz a jejich praktické využití pro zadaný úkol. Dále se zaměřím na emoce a jejich vliv na řeč. Ze zkoumaných parametrů zmíním například vliv na základní tón řeči, na polohu a šířku jednotlivých formantů

V druhé části se seznámím s veřejně dostupnými databázemi řečových signálů pořízených při různých stavech mluvčích. Tuto databázi budu také částečně zkoumat a konečné výsledky použiji na tvorbu diplomové práce.

V poslední části analyzuji některé emoční stavy a pokusím se určit rozhodovací faktor pro tento emoční stav.

1.1. Model hlasového traktu



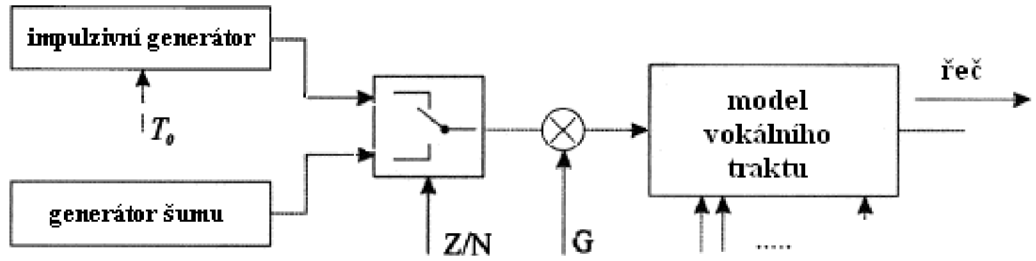
Obr. 1.1.: a) Fyziologický model hlasového ústrojí
b) a jeho blokové znázornění [10]

Člověk používá pro tvorbu řeči svůj hlasový trakt [1][9](dutina hrdelní, ústní a nosní) a hlasivky. Hlasový trakt je dutina, kterou se dostává vzduch ven z plic. Tvoří tak jakýsi dutinový rezonanční filtr, jehož kmitočtová charakteristika odpovídá vyslovenému úseku řeči (doba, po kterou zůstává přenosová charakteristika téměř konstantní, se pohybuje kolem 20 ms). Je tedy zřejmé, že během promluvy se parametry dutiny (její tvar) mění. Na svém vstupu je hlasový trakt buzen buď tlakovými pulsy vzduchu (hlasivkové pulsy), které vytvářejí hlasivky svým kmitáním s periodou řádově jednotky milisekund (v rozmezí od 2,5 ms do 10 ms), nebo je buzen prostým prouděním vzduchu přes otevřené hlasivky, které v tomto případě nekmitají.

Hrtan a hlasivky vytvářejí periodický signál pro znělé a šumový signál pro neznělé hlásky. Budící signál má tedy již při vstupu do hlasového traktu vždy jiné spektrální vlastnosti. Trakt má v každém okamžiku určité impedanční a rezonanční charakteristiky, které dále mění budící signál. Vyzařovací charakteristiky rtů a nosu zesilují v původním budícím signálu vyšší kmitočty (signál je diferencován). Celkový účinek hlasového ústrojí na budící

signál může být zahrnut do jediného filtru s časově proměnnými vlastnostmi, který určité kmitočty tlumí a jiné zesiluje, a to v závislosti na právě vyslovované hlásce. Fyziologický model hlasového ústrojí je nakreslen na obr. 1.1.

Jednotlivé dutiny (dutina hrdelní, ústní a nosní) tvoří akustické dutinové rezonátory, jejichž rezonanční kmitočet závisí na fyzických rozměrech dutiny. Rezonanční kmitočet v tomto smyslu označujeme jako formant. Z velikosti dutin lze pak usuzovat, že první formant (nejnižší rezonanční kmitočet) přísluší dutině hrdelní, druhý a třetí formant dutině ústní a nosní (jednoznačné přiřazení druhého a třetího formantu není možné). Další formanty jsou způsobeny nevýraznými dutinami v hlasovém traktu a propojením jednotlivých dutin.



Obr. 1.2.: Diskrétní model hlasového traktu.

Výše zjištěné poznatky jsou shrnuty v matematickém modelu hlasového ústrojí člověka. Tento model je však jen přibližnou aproximací skutečného fyzikálního modelu, neboť není možné vytvořit takovou soustavu diferenciálních rovnic, která by popisovala vokální trakt v celé jeho rozmanitosti. Dáno je to tím, že vokální trakt člověka obsahuje množství matematicky těžko popsatelných parametrů. Proto se používají jen zjednodušené modely, jako například diskrétní model na obr.1.2. Na základě takovéhoto modelu se dá provést jak syntéza, tak i analýza řeči.

Model je lineární přenosový systém a skládá se z buzení, filtru a bloku realizující regulaci zesílení.

Buzení má pouze dva druhy budících signálů, a to buď šum nebo periodický signál s pilovým či jinak mu podobným průběhem. Šum vzniká v šumovém generátoru s různým typem rozložení šumů, pro jednoduchost se většinou uvažuje gausův šum. Šumový generátor se používá také na tvorbu frikativ. Exploziva vzniknou krátkým připojením impulsního generátoru. Impulsní generátor produkuje periodické buzení, výsledkem jsou znělé signály

(periodické signály). Ve fyziologickém modelu nahrazuje impulsní generátor funkci hlasivek. Hlasivky mají svůj základní hlasivkový tón na kterém kmitají.

Filtr je realizován kaskádou malého počtu dvoupólových rezonátorů. Rezonanční frekvence těchto filtrů odpovídají frekvencím jednotlivých formantů. Pro většinu zvuků v řeči dává dobrý popis hlasového ústrojí pouze pólový model. Přenosová funkce reprezentující hlasový trakt není stacionární, a proto se musí přibližně každých 10 ms obměnit. Pro realizaci takovýchto rychle se měnících systémů se užívá křížových filtrů. Přenosová funkce filtru je určena predikčními koeficienty a má tvar:

$$H(z) = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}} \tag{1.1}$$

Kde M představuje řád prediktoru a a_i jsou predikční koeficienty.

Dalším filtrem z kaskády je filtr realizující přenos odpovídající vyzařování zvuků, tedy útlum zvuku vlivem radiálního šíření zvuku. Tvar přenosové funkce je:

$$L(z) = 1 - \lambda z^2 \tag{1.2}$$

Koeficient λ se volí z intervalu <0.9,1>.

Zesílení před a za filtrem slouží k normování vstupního signálu vstupujícího do filtru a nastavení výstupní úrovně signálu vycházejícího z modelu.

Kapitola 2

2. Zpracování akustického signálu

Základem většiny metod akustické analýzy řeči je předpoklad, že se vlastnosti řečového signálu v průběhu času mění pomalu. To vede k možnosti aplikace tzv. metod krátkodobé analýzy, při kterých se řečový signál zpracovává po jednotlivých krátkých úsecích délky asi 20 ms. Na těchto úsecích je lidská řeč relativně stabilizovaná a_můžeme ji klasifikovat. Výsledkem krátkodobé analýzy je skalár nebo vektor čísel, popisující daný úsek řečového signálu. Protože jednotlivé úseky na sebe navazují, získáme tak časové posloupnosti příznaku či vektorů příznaků, které dostatečně popisují analyzovanou promluvu. Většinou nesou informaci o energetických a kmitočtových změnách akustického signálu v čase a jsou používány k jeho parametrizaci (určení reprezentativních příznaku), která je pro další analýzu přínosná a efektivní. Parametrizací signálu totiž snížíme množství zpracovávaných dat a minimalizujeme jejich redundanci.

2.1 Informační obsah řeči

Řečový signál je nositelem následujících informací [7]:

- Vlastní vyslovené zprávy, která se skládá ze slov.
- Identity mluvčího, který je rozpoznáván podle individuálního stylu promluvy.
- Nálady mluvčího, která se odráží ve stylu promluvy.
- Původu mluvčího (jazyk, dialekt, přízvuk).

Při zpracování řečového signálu musíme mít k dispozici mnoho různých forem znalostí, které jsou specifické pro řeč. V následujícím textu jsou stručně popsány některé zdroje znalosti nutné pro zpracování řeči, bez kterých se ani člověk neobejde.

Fonetika

Fonetika zkoumá zvukový signál lidské řeči v celé jeho složitosti a poskytuje základní znalosti o tvoření mluvené řeči. Definuje možný inventář hlásek a popisuje jejich artikulační a akustické vlastnosti.

Pro věrný záznam skutečných zvuku řeči se užívá *fonetická transkripce*. mezi několika druhy transkripcí je nejrozšířenější transkripce definovaná Mezinárodní fonetickou asociací a označována IPA. Pro transkripci spisovné češtiny však není systém API optimální. Proto byla pro účely automatického rozpoznávání češtiny navržena abeceda označena PAC.

Fonologie

Na rozdíl od fonetiky se zabývá fonologie hláskami z hlediska jejích výskytu a schopnosti vytvářet vzájemné kombinace. Nejmenší fonologickou jednotkou je foném. Nauka o fonémech popisuje možnou variabilitu fonémů ve výslovnosti.

Prozodie

Prozodie je nauka o zvukové stránce jazyka. Určuje melodii ve větě a přízvuk v jednotlivých slovech. Touto oblastí se budeme dále zabývat podrobněji.

Lexikologie

Lexikologie je nauka, která se zabývá slovy a jinými elementy podílejících se na tvorbě slov. Kromě toho se zabývá synonymy (slova s podobným významem), antonymy(slova s opačným významem) a hyponymy (podřazenými pojmy), a dále otázkami ve smyslu etymologie (nauka o původu a vývoji slov).

Syntax

Syntax je nauka o skladbě věty (gramatika), tedy o správném způsobu řazení menších řečových jednotek do vět.

Sémantika

Sémantika určuje význam slov, a tím napomáhá vybrat ze seznamu všech možných slov optimální slovo, které je v dané souvislosti nejvýhodnější.

2.2 Rozpoznávání akustického signálu řeči

Úloha rozpoznávání akustického signálu řeči je již v dnešní době velmi dobře řešena, a to jak pro rozpoznávání izolované tak i spojité promluvy. Metody, algoritmy a postupy pro rozpoznávání akustického signálu řeči jsou intenzivně vyvíjeny již více než třicet let. Tento vývoj umožnila především dostupnost dostatečně spolehlivých osobních počítačů, které se začaly objevovat na trhu v osmdesátých letech minulého století.

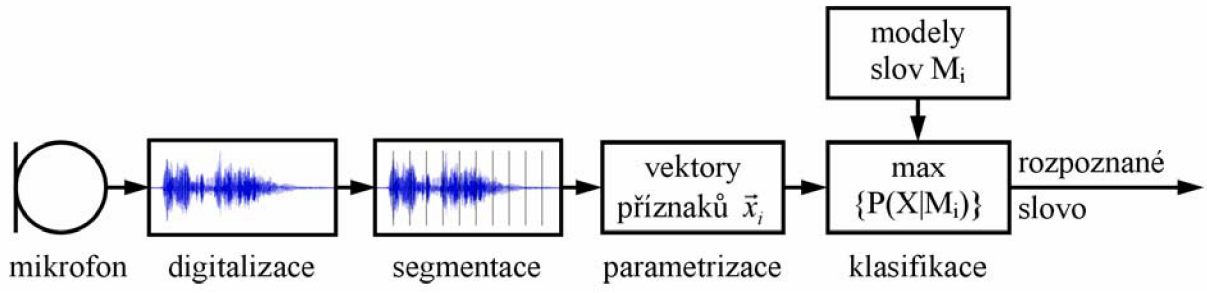
V úloze rozpoznávání izolovaných slov s velkým slovníkem (10 000 až 1 000 000 slov) se dnes již dosahuje rozpoznávacích skóre přes 90% a v úloze rozpoznávání spojité řeči je rozpoznávací skóre větší než 80%[1].

Ve své práci se chci zabývat především audio-vizuálním rozpoznáváním řeči pro český jazyk. Pro audio-vizuální rozpoznávání izolovaných slov a spojité promluvy založené na modelech menších stavebních jednotek řeči (fonémy, vizémy) je však potřeba vytvořit především vhodné modely českých vizémů. Nalezením vhodných českých vizémů se již zabývám a naše audio-vizuální databáze byla pořízená a zpracovávána s tímto záměrem. Přesto největších pokroků jsem zatím dosáhl při audiovizuálním rozpoznávání izolovaných slov založených na celoslovních modelech. Tato úloha vedla k nalezení a extrakci vizuálních příznaků a nalezení metodologie pro vlastní audio-vizuální rozpoznávání řeči.

V této kapitole je proto popsána strategie rozpoznávání řeči především pro akustické rozpoznávání izolovaných slov založená na celoslovních modelech. Zde popsaná digitalizace, segmentace a parametrizace akustického signálu řeči je však obdobná i pro rozpoznávání spojité řeči. O metodách pro rozpoznávání izolovaných slov a spojité řeči založené na modelech menších stavebních jednotek řeči se lze dočíst např. v [HUA01,

PSU95, KOL01].

Na obrázku 2.1 je zobrazen princip rozpoznávání akustického signálu řeči. Nejprve je analogový signál z mikrofonu digitalizován, poté je digitalizovaný signál segmentován na menší segmenty (framy), každý fram je parametrizován a nakonec je provedena klasifikace, kde je sled parametrizovaných framů porovnáván s jednotlivými modely slov a na základě předem daného kritéria je rozhodnuto, který model slova patří ke vstupnímu signálu řeči.



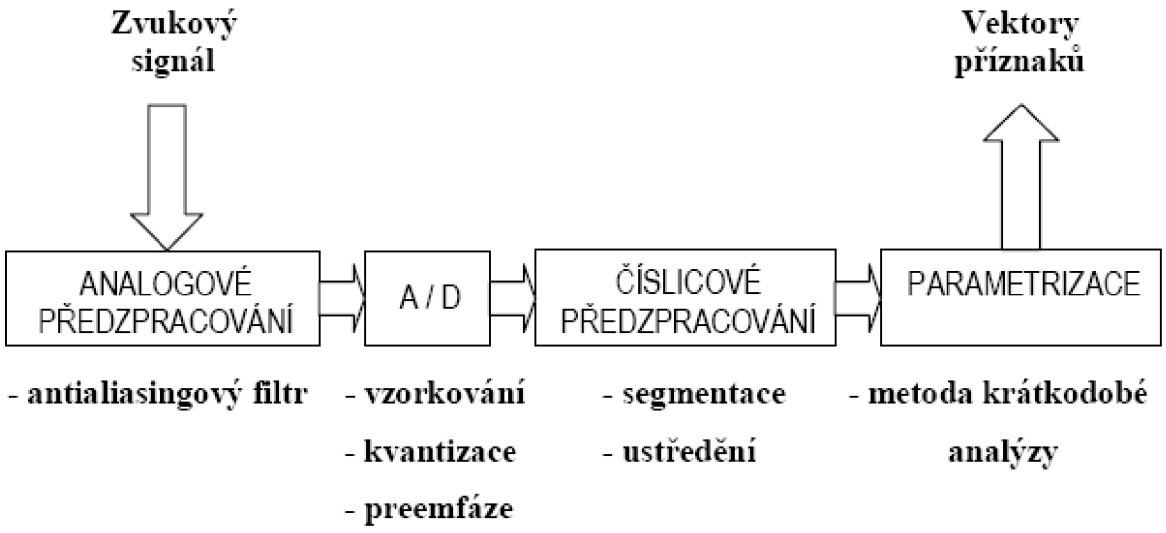
Obr. 2.1**.:** Jednotlivé metody pro rozpoznávání akustického signálu řeči.

2.3. Předzpracování signálu

Před použitím metod krátkodobé analýzy pro parametrizaci akustického signálu provádíme jeho předzpracování. Na obrázku 2.2. jsou znázorněny jednotlivé etapy předzpracování a pořadí jejich aplikace před vlastní parametrizací.

2.3.1. Analogové předzpracování

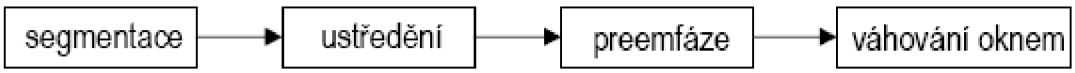
Ve fázi analogového předzpracování používáme antialiasingový filtr. Jedná se o dolní propust pro potlačení vyšších frekvencí signálu. Skutečný zvukový signál totiž není frekvenčně omezený a pro jeho digitalizaci a zabránění vzniku aliasingu (nežádoucích frekvencí vzniklých vzorkováním) jej musíme omezit.



Obr. 2.2.: Předzpracování akustického signálu a jeho parametrizace.

2.3.2. Číslicové předzpracování

Po převedení spojitého signálu do jeho diskrétní podoby můžeme přistoupit k číslicovému zpracování. Číslicové zpracování se obvykle skládá z_následujících bloku:



Obr. 2.3. Číslicové předzpracování.

- **Segmentace**

Při zpracování řečového signálu metodami krátkodobé analýzy požadujeme, aby signál měl, pokud možno, stejné vlastnosti - byl stacionární. Toho dosáhneme pouze na určitých krátkých úsecích - ve skutečnosti je signál nestacionární. Uvažované krátké úseky nazýváme rámce nebo též segmenty a jejich délka musí být dostatečně malá na to, abychom signál mohli považovat za stacionární, ale zároveň i dostatečně

velká pro jeho přesnou analýzu. Velikost rámců volíme v rozmezí 10-30 ms. Rámce se mohou částečně překrývat.

- **Ustředění**

Při snímání signálu může dojít, k jeho stejnosměrnému posunutí v kladném či záporném směru, to může mít nepříznivý dopad na další zpracovaní, především na výpočty funkcí krátkodobé energie, proto je vhodné signál ustředit. Ustředění jednoho rámce probíhá následovně:

$$s'(n) = s(n) - \frac{1}{N}\sum_{i=1}^{N} s(i) \ \ , pro\_1 \le n \le N \tag{2.1}$$

kde $s(n)$ je n-tá hodnota rámce, $s'(n)$ je ustředěná hodnota a N je počet vzorů obsažených v jednom rámci. Eventuelně lze použít číslicového filtru, který zadrží dolní frekvence. Ten je použit hlavně při zpracování signálu neděleného na rámce.

- **Preemfáze**

Lidské hlasové ústrojí produkuje řeč, jejíž amplitudy harmonických složek s frekvencí klesají. Pro zpracování řeči a její následnou klasifikaci je však vhodnější, pokud je spektrum řeči rovnoměrné. K tomu slouží blok preemfáze. Blok preemfáze je číslicový filtr zvýrazňující vyšší harmonické složky. Filtr lze popsat následující rovnicí:

$$y[n] = x[n] - \alpha x[n-1] . \tag{2.2}$$

kde $x(n)$ je vstupní vzorek filtru v čase n, kde $x(n-1)$ je vstupní vzorek zpožděný o vzorkovací periodu, $y(n)$ je výstup filtru a a je parametr. Parametr a bývá volen od 0.9 do 1.

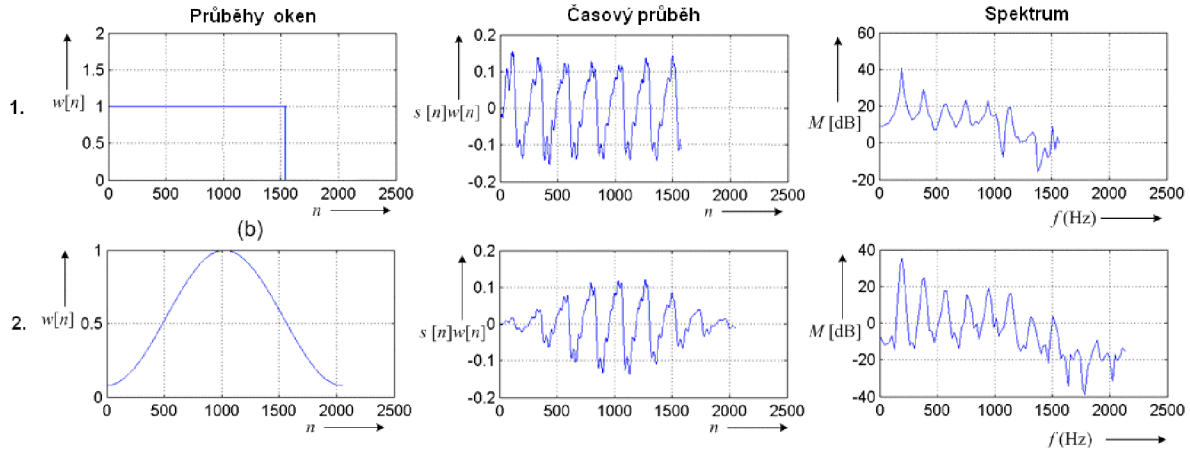
- **Váhování oken**

Při krátkodobé analýze zpracováváme signál po malých intervalech - rámcích. Rozhodneme-li se pro určitou velikost intervalu (většinou 10-30 ms), implicitně předpokládáme, že zvukový signál v okolí je periodický s_periodou uvnitř rámce. Není-li perioda shodná s délkou rámce nebo mají-li rámce nenulové překrytí, dopouštíme se jisté chyby ve zpracování. To ale můžeme částečně opravit použitím tzv. *váhového okénka.* Úkolem okénka je v analyzovaném rámci vybrat příslušné vzorky signálu a při zpracování vyhladit jejich průběh přidělením určité váhy. Aplikace okénka na signál se prakticky provádí vynásobením odpovídající si hodnoty rámce a okna:

$$y(n) = x(n).w(n) \tag{2.3}$$

kde *y(n)* je výstupní posloupnost, *x(n)* je vstupní posloupnost a *w(n)* je aplikované okno. Nejpoužívanějším typem okna je okno Hammingovo, které je definováno následovně:

$$w(n) = 0{,}54 - 0{,}46 * \cos(2\pi \frac{n-1}{N-1}), pro\_1 \le n \le N \tag{2.4}$$



Obr.2.4: Funkce oken: 1.řádek – Obdelníkové okno:(průběhy oken, časový průběh segmentu, spektrum segmentu); 2.řádek – Hammingovo okno:(průběhy oken, časový průběh segmentu, spektrum segmentu)

Existují i jiné typy oken, informace o nich lze najít v publikaci [7].

2.4. Digitalizace akustického signálu řeči

Pro zpracování akustického signálu počítačem je nutné jej převést z analogového tvaru na číslicový. Vzhledem k technickému a vědeckému pokroku posledních desetiletí je digitalizace akustického signálu řeči (vzorkování a kvantování) v současné době již velmi dobře vyřešena a digitalizování signálu řeči dnes zajistí „každá“ zvuková karta v PC. Z různých dřívějších pokusů se signály řeči bylo zjištěno, že k porozumění obsahu digitalizovaného signálu mluvené řeči postačuje frekvenční pásmo 0 až přibližně 3,4 kHz. Z Shannonova vzorkovacího teorému pak plyne, že pro digitalizaci signálu řeči by měla být vzorkovací frekvence Fs větší než 6,8 kHz. Pro účely rozpoznávání řeči se nejčastěji volí vzorkovací frekvence $Fs \geq 8$ kHz a počet úrovní kvantování se obvykle
volí 2_{16} [PSU95, HUA01].

2.5. Metody krátkodobé analýzy

Při analýze akustického signálu pracujeme s velkým objemem značně redundantních dat. Cílem metod krátkodobé analýzy je snížení množství dat tím, že_signál pro další zpracování reprezentujeme ne vlastními nasnímanými vzorky, ale souborem příznaku ze vzorku získaných. Snahou je, aby příznaky co nejlépe popisovaly průběh signálu i jeho vlastnosti, na které se v aplikacích zaměřujeme. Existují různé metody krátkodobé analýzy, které jsou většinou vhodné jen pro určitý okruh úloh. Hlavním kriteriem výběru je především typ řešené úlohy a s tím související požadovaná výpočetní složitost. Pro všechny typy metod je společná jejich aplikace pouze na krátký časový úsek signálu (10-30 ms). Ze stacionarity akustického signálu právě na těchto krátkých úsecích vyplývá nutnost korektní analýzy signálu jen po malých částech. V praxi tedy provedeme segmentaci signálu na jednotlivé rámce a na ně pak aplikujeme některou z metod krátkodobé analýzy. Pro každý rámec získáme jeden příznak - skalár, nebo vektor příznaku – akustický vektor, který popisuje signál v konkrétním rámci. Cílem je, aby získaný popis byl jednoznačný pro každý typ signálu a co nejméně ovlivněn okolními podmínkami. Musí být co nejméně závislý na množství šumu v signálu, na jeho intenzitě a při rozpoznávání řeči i na řečníkovi. Metody krátkodobé analýzy

můžeme rozdělit do dvou kategorií podle toho, zda s_akustickým signálem pracují jako s časovou posloupností navzorkovaných hodnot, nebo zda vnímají spíše jeho spektrální charakter.

2.5.1. Zpracování v časové oblasti

Parametry získané z časového průběhu řeči vyžadují nejmenší výpočetní nároky. Lze z nich však vyčíst pouze základní informace a proto se využívají pro základní rozhodovací mechanizmy jako je klasifikace řeč/šum, znělost/neznělost řeči nebo ve spojení s příznaky vypočtenými z frekvenční oblasti. Mezi příznaky získané z časového průběhu patří například krátkodobá intenzita, krátkodobá funkce středního počtu průchodu signálu nulou nebo krátkodobá autokorelační funkce.

Krátkodobá energie je definována vztahem:

$$E = \frac{1}{N}\sum_{k=1}^{N} x(k)^2 \tag{2.5}$$

Jelikož velká dynamika řečového signálu je při výpočtu krátkodobé energie ještě umocněna na druhou, bývá tento parametr logaritmován. Více informací ke zpracování v časové oblasti se lze dozvědět v literatuře [1].

2.5.2 Zpracování ve frekvenční oblasti

Stejně jako při zpracování v časové oblasti lze v časovém úseku 10-30 ms považovat za konstantní i spektrální charakteristiky řeči. Proto mluvíme o_krátkodobé spektrální analýze. Mluvená řeč je ve frekvenční oblasti reprezentována zastoupením jednotlivých frekvencí, svým spektrem. Základem většiny používaných metod zpracování ve frekvenční oblasti je Fourierova transformace. Jelikož se zde pracuje s diskrétním signálem o *N*_vzorcích, uvedu vztah pro její diskrétní verzi:

$$S(k) = \sum_{n=0}^{N-1} s(n) \times e^{-j\frac{2*\pi*k*n}{N}} \tag{2.6}$$

Spektrum řeči získané diskrétní Fouierovou transformací však stále nese velké množství redundantních informací, je korelované (jednotlivé frekvence jsou na sobě závislé) a_obsahuje v sobě buzení hlasového traktu, proto se spektrum pro klasifikaci rámců nepoužívá. Spektrum je obvykle rozděleno do několika pásem, zpracováno v každém individuálně a vypočítáno tzv. *kepstrum*. Kepstrum můžeme obecně popsat vztahem:

$$C(x) = DTF^{-1} * \left|\ln(\left|DST(s(x))\right|)^2\right| \tag{2.7}$$

kde *DFT* je diskrétní Fourierova transformace, *DFT-1* je inverzní diskrétní Furierova transformace, *s(x)* jsou vstupní vzorky a *C(x)* jsou vypočtené kepstrální koeficienty.

Kepstrální analýza je metoda umožňující ze signálu řeči oddělit parametry buzení a_hlasového ústrojí. Každá harmonická složka spektra je dána součinem buzení a_složky závislé na hlasovém ústrojí. Po zlogaritmování tento součin přejde na součet dvou složek, nebo *ln(ab) = ln(a)+ln(b),* což je umožňuje od sebe jednoduše oddělit. Spektrální obálka je popsána několika koeficienty ze začátku kepstra.

2.6. Dlouhodobé spektrum

Dlouhodobé spektrum patří k nejjednodušším a nejčastějším statistickým parametrům. Vyjadřuje částečně anatomické vlastnosti hlasového traktu. Výpočet dlouhodobého spektra se nejčastěji řeší pomocí LPC spektra nízkého řádu (obvykle řádu 2 nebo 4)[7].

Výpočet spočívá v určení průměrných autokorelačních koeficientů

$$\overline{R}(k) = \frac{1}{J}\sum_{j=1}^{J} R_j(k) \tag{2.8}$$

kde *J* značí celkový počet segmentů v testovaném úseku řeči. Z průměrných autokorelačních koeficientů vypočteme průměrné LPC koeficienty,(pro řád 2) např. metodou Levinson&Durbin [7].

$$\begin{aligned} \overline{a}_1 &= \frac{\overline{R}(1)\cdot\overline{R}(O) - \overline{R}(1)\cdot\overline{R}(2)}{\overline{R}(0)^2 - \overline{R}(1)^2} \\ a2 &= \frac{\overline{R}(2)\cdot\overline{R}(0) - \overline{R}(1)\cdot\overline{R}(1)}{\overline{R}(O)^2 - \overline{R}(1)^2} \end{aligned} \tag{2.9}$$

Dlouhodobé spektrum pak vypočteme ze známých LPC koeficientů vztahem

$$S(f) = \left| \frac{1}{\sum a_m \cdot z^2} \right|^2, z = \exp(\frac{j \cdot 2 \cdot \pi \cdot f}{f_{vz}}) \tag{2.10}$$

Kde je vzorkovací kmitočet signálu, jsou LPC koeficienty a značí kmitočet signálu.

2.7. Prozodie

Prozodie [10] v lingvistice popisuje zvukové vlastnosti jazyka, které se uplatňují na úrovni vyšší než jednotlivý foném. Souhrnně se hovoří o tzv. suprasegmentálních jevech které znamenají veškeré změny základního tónu, intenzity a trvání.

Prozodie je všude přítomná a grafickými prostředky nelze vyjádřit všechny funkce které zajišťuje. V mateřském jazyce dítě napodobuje melodii a rytmus řeči dříve, než začne produkovat první náznaky slov, a také jako první chápe prozodii a informaci o emoci, kterou prozodie nese. Prozodie mateřského jazyka je prelexikální a pregramatická, na rozdíl od jazyka cizího. V něm se o prozodických vlastnostech a funkcích se dozvíme, pokud vůbec až naposled. Následkem tohoto rozporu jsme v cizím jazyce většinou ochuzeni o možnost sdělovat informace pomoci prozodie, obohacovat tak promluvu a jsme omezeni, v produkci i v percepci cizího jazyka, na prvoplánové informace obsažené v lexikální rovině sdělení.

Jak bylo řečeno, celek prozodických vlastnosti promluvy je definován jako souhrn intonace, přízvuku, rytmu a distribuce pauz a může být rozdělen do dvou velkých skupin: rytmus a intonace. Pojem intonace zde odkazuje ke změnám ve frekvenci základního tónu, vnímaným jako změny výšky hlasu. Rytmické změny zahrnují distribuci, silu a pravidelnost přízvuku a také variace intenzity.

Funkce prozodie

Funkce prozodie je možné velmi zhruba rozdělit do dvou skupin na skupinu funkcí jazykových a skupinu funkcí nejazykových. jazykové funkce jsou někdy podrobněji děleny na funkce sémantické a syntetické. Mimojazykové funkce jsou členěny na paralingvistické a extralingvistické. Tyto funkce zajišťují informace o mimojazykových skutečnostech, a to jednak informace poskytované mluvčím vědomě(jako jsou postoje), a dále informace, které jsou v řeči nevyhnutelně přítomné nezávisle na vůli a vědomí mluvčího. Mezi ně spadají informace o věku a pohlaví mluvčího, jeho aktuálním fyzickém a psychickém stavu včetně emocí.

- ***Jazykové funkce prozodie***

Sémantická funkce prozodie zajišťuje rozlišení deklarativní(vyhlašovací nebo vysvětlovací) a interogativní (tázací) věty. Tato funkce se uplatňuje, stojí-li za rozdílem mezi

větou deklarativní a interogativní pouze prozodické změny, nikoliv změny na úrovni gramatické. Je tomu tak prakticky téměř ve všech jazycích.

Čtyřem hlavním modalitám odpovídají čtyři hlavní intonační schémata:

- Deklarativní větu provází použití středního hlasového rejstříku mluvčího a klesavá melodie.
- Přímý pokles finální melodie je charakteristicky pro větu rozkazovací.
- Otázka zjišťovací je zastoupena stoupnutím melodie až do vyšších poloh hlasového rozsahu mluvčího.
- Klesavá melodie spojující vyšší polohu hlasového rozsahu mluvčího s nižší je charakteristická pro otázku doplňovací.

▪ *Fonostylistické funkce prozodie*

Tato skupina funkcí nás bude v této práci zajímat nejvíce. Do této skupiny patří například

- Funkce identifikační: prostřednictvím této funkce se charakterizuje prozodie mluvčího i bez jeho vědomí např. jeho emoční stav, věk a pohlaví nebo jeho sociální a regionální původ.

- Funkce impresivní: Tato funkce odpovídá záměru mluvčího dát promluvě určitý styl jako je emfáze, řečnický styl a podobně.

2.8. Mikroprozodie

Jak již bylo zmíněno, prozodie má suprasegmentální charakter, tzn. že se nevztahuje k jednotlivým segmentům řeči, ale k celým úsekům (úsekem může být slovo nebo dokonce celá věta), přesto je třeba upozornit, že jednotlivé segmenty mohou celkovou prozodii ovlivnit. Akustické parametry, které mají vliv na prozodii, totiž mohou vykazovat určité rychlé změny uvnitř jednotlivých segmentů. Tyto změny se nazývají mikroprozodie, frekvence základního tónu se například vyvíjí v rámci každého fonému a tento lokální vývoj bývá ovlivněn sousedními segmenty. Základní tón pro samohlásky bývá nižší po plozivě a vyšší po frikativě apod. Takové jevy však lze kvůli dominantnímu suprasegmentálnímu charakteru prozodie jen těžko pozorovat. Velmi důležité je s těmito jevy počítat a nepokládat je například za významné nárůsty či poklesy F_0 v rámci celé promluvy apod.

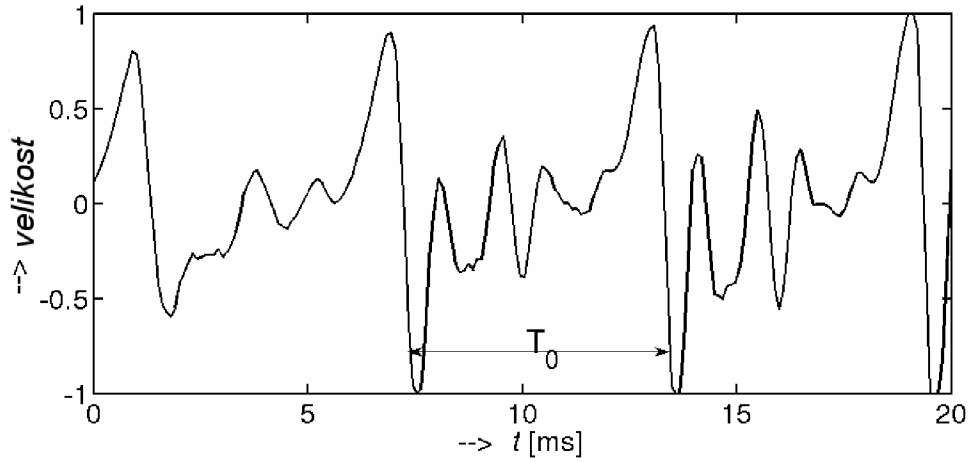
V některých prací zabývající se problematikou rozpoznávání emočních stavů se považuje mikroprozodie jako jedena z příznaků používaných pro rozpoznávání, v této práci tento příznak nebudeme používat.

Kapitola 3

3. Suprasegmentální rysy řečového signálu

3.1 Základní tón řeči

Základní tón řeči představuje hlavní slyšitelný parametr řeči, kterým rozlišujeme (poznáváme) jednotlivé mluvčí. V západních jazycích nemá význam pro rozpoznávání řeči, slouží pouze k vyjádření intonace a tím zvyšuje srozumitelnost. Naproti tomu v některých východních jazycích (např. čínština, vietnamština) má podstatný význam rozpoznávání slov. U většiny populace leží základní kmitočet řeči v intervalu 50-400Hz, přičemž oblast kolem 50Hz odpovídá hlubokým mužským hlasům a vysoké hodnoty 300-400Hz představují dětské hlasy. Průměrný základní kmitočet řeči u mužů je 120Hz, u žen 210Hz. Operní pěvci mohou vyzpívat text se základním tónem do 700-800Hz, ovšem nad 700Hz obvykle přestává být srozumitelný.



Obr. 3.1.: Průběh základního tónu řeči.

Kmitočet základního tónu řeči značíme(fundamental frequency), periodu základního tónu řeči značíme(pitch period). Pro přepočet kmitočtu základního tónu řeči na jeho periodu platí jednoduchý vztah

$$F_0 = \frac{1}{T_0} \quad [\text{Hz}]. \tag{3.1}$$

Při určování základního tónu řeči musíme vzít v úvahu to, že signál generovaný hlasivkami přesně vzato není posloupnost periodických impulsů, ale s časem se mění (s výjimkou zpěvu). Jedná se o kvaziperiodický signál, u kterého se délka periody může měnit i uvnitř analyzovaného segmentu (obvyklá délka segmentu je 20ms). Kdyby byl základní tón řeči konstantní, působila by řeč monotónně a strojově.

Problémy při určovánímohou nastat u signálu s nízkou energií, kde je obtížné oddělit od sebe znělé a neznělé úseky. Je třeba také vzít v úvahu, že příliš vysoký základní tón může být silně ovlivněn nízkou hodnotou formantu. U signálu omezeném na telefonní pásmo (0,3-3,4kHz) se mohou objevovat jen vyšší harmonické základního kmitočtu řeči.

Na určení základního kmitočtu řeči bylo vypracováno několik metod, mezi nejjednodušší patří metoda AMDF[1] (Average Magnitude Difference Function) nebo častěji používaná metoda Center clipping [1].

3.2. Intenzita (Energie)

Intenzita [4] neboli energie signálu je vnímána jako síla hlasu, hlasitost. Její úroveň je spojená s funkcí dýchacího a fonačního systému.

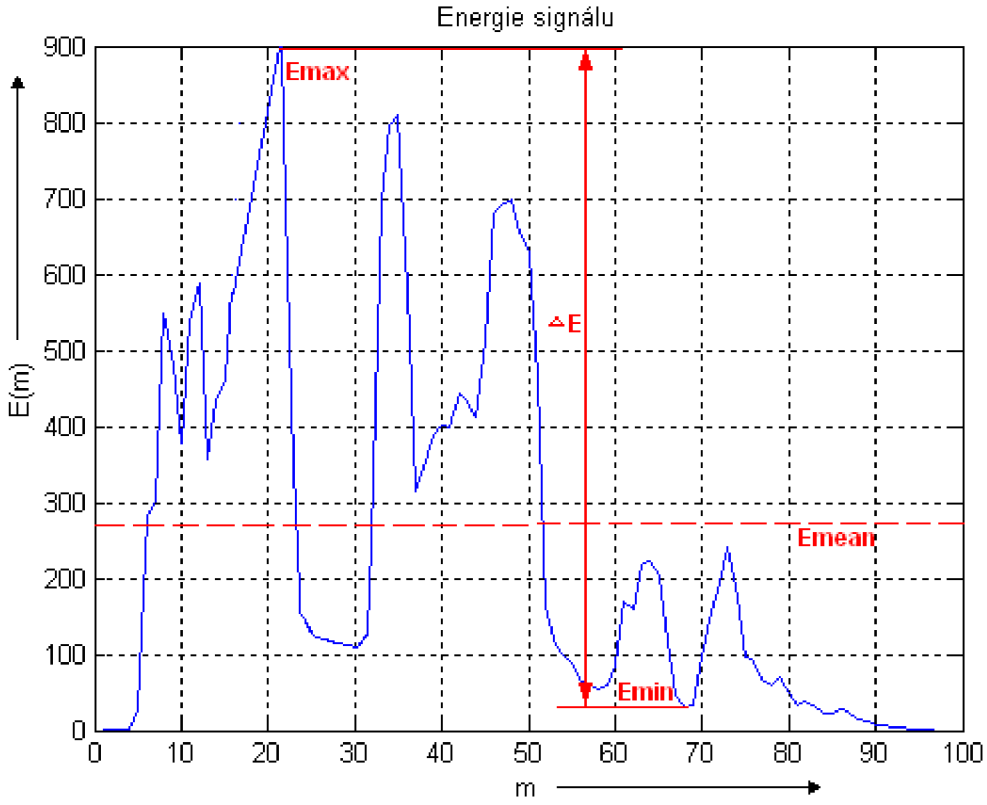
Při měření intenzity nahraného řečového signálu můžou být hodnoty ovlivněny různým nastavením citlivosti nahrávacího zařízení nebo pohybem mluvčího a změnami vzdálenosti od mikrofonu.

Kromě problémů, které můžou vzniknout při nahrávání a jsou spíše rázu technického, je intenzita ve fonetických studiích opomíjena i z dalších důvodů: je často považována za kovariant základní frekvence, protože je stejně jako ona těsně spjata se změnami subglotálního tlaku. Dá se říct , že chování intenzity je v podstatě náhodné.
Technicky budeme popisovat intenzitu pomocí krátkodobé energie. Funkci krátkodobé energie signálu lze definovat vztahem

$$E = \frac{1}{N}\sum_{n=0}^{N-1}(x[n])^2 \qquad [-]. \qquad (3.2)$$

(obr. 3.2) představuje příklad průběhu energie řečového signálu na kterém je znázorněné m.j. střední, maximální a minimální hodnota energie daného signálu, tyto parametry budeme používat později při rozpoznávání emočních stavů. Ze zmíněného obrázku je patrné, že minimální hodnota energie se hledá v oblasti mimo začátku a konci průběhu.

Při měření krátkodobé energie lze doporučit délku rámce 10-20 ms. Hodnoty funkce krátkodobé energie poskytují pro každý rámec informaci o průměrné hodnotě energii v rámci.



Obr. 3.2: Příklad průběhu energie(intenzity) řečového signálu.

3.3. Kepstrum

Cepstrální analýzu můžeme zařadit do tzv. homomorfického, obecně nelineárního zpracování signálů. Použití této techniky pro analýzu a parametrizaci řečového signálu je velice výhodné a efektivní [2]. Tyto metody se hodí pro oddělování signálů, které vznikly

konvolucí nebo násobením více složek. Výsledky kepstrální analýzy slouží nejen pro rozpoznávání řeči, ale jsou cenným materiálem zejména pro rozpoznávání mluvčích a jejich stavů.

Hlasivky vytvářejí kvaziperiodickou nebo šumovou budící funkci a hlasový trakt s impulsní odezvou tuto funkci moduluje. Na výstupu hlasového traktu vzniká řečový signál konvolucí, což v kmitočtové oblasti odpovídá násobení obou funkcí po Fourierové transformaci.

$$s(n) = g(n) * h(n) \tag{3.3}$$

Po přechodu z časové oblasti do kmitočtové se operace konvoluce změní na násobení.

$$\log|S(f)| = \log\left[|G(f)| \cdot |H(f)|\right] = \log|G(f) + \log|H(f)|\| \tag{3.4}$$

Po logaritmování jsou logaritmované součinové složky k dispozici jako součet. Operace logaritmování mění amplitudu spektra, nikoliv však jeho charakteristiku. Inverzní Fourierova transformace transformuje složky zpět do časové oblasti, kde existují dále jako součet, protože součet po Fourierově transformaci zůstává zachován.

$$F^{-i}\left|\log|S(f)|\right\| = F^{-i}\left|\log|G(f)|\right\| + F^{-i}\left|\log|H(f)|\right\| \tag{3.5}$$

Vyjádřeno jako časové funkce, můžeme psát:

$$c_s(\tau) = c_g(\tau) + c_h(\tau) \tag{3.6}$$

Výstupní signál je označován jako cepstrum signálu. Komplexní cepstrum tedy představuje inverzní Fourierovu transformaci logaritmu Fourierova obrazu vstupního signálu.

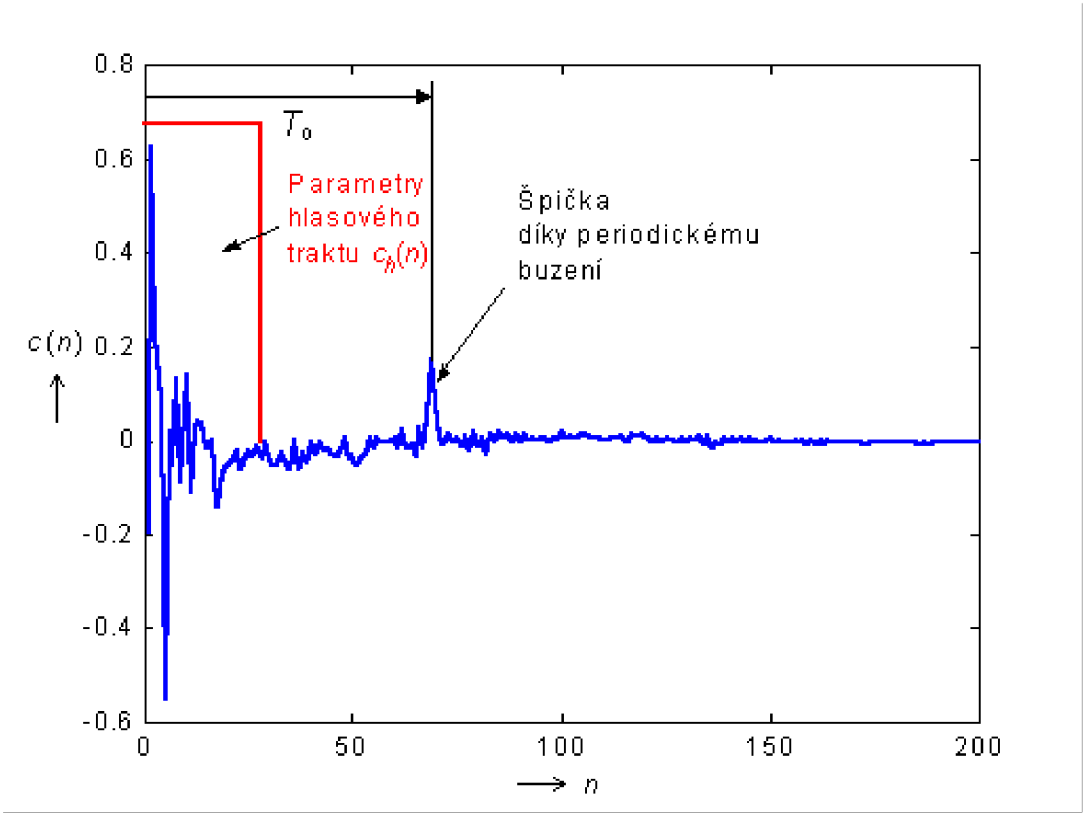
$$c_s(\tau) = F^{-i}\{\log|S(f)|\} \tag{3.7}$$

Použijeme-li při výpočtu pouze reálnou část spektra, hovoříme jednoduše o kepstru. Jelikož nezávislá proměnná má opět rozměr diskrétního času (vzorky), říká se tomuto rozměru slovní hříčkou kvefrence. Cepstrum (kepstrum) vzniklo podobnou hříčkou ze slova spektrum.

Obr. 3.3. Systém k určení cepstra signálu.

Kepstrum je používáno k oddělení budících a přenosových parametrů řečového signálu. Zjednodušené schéma k výpočtu cepstra je na obrázku 3.3.

První koeficient kepstra představuje energii signálu. Koeficienty s nízkým pořadím (dolní kvefrence) popisují pomalé změny ve spektru signálu, tzn. Formantovou strukturu a tím i charakteristiku hlasového traktu. Koeficienty s vyšším pořadím (horní kvefrence) reprezentují rychlé změny ve spektru signálu, čímž specifikují změny buzení hlasového traktu. Rozdělením cepstra na dvě části lze tedy poměrně jednoduše oddělit složky buzení a formování řečového signálu. U znělých úseků řeči se v cepstru vyskytuje výrazná špička (obrázek 3.4.), která svou polohou určuje základní tón řeči.



Obr. 3.4.: Kepstrum znělého signálu [11].

3.4. Mel-frequency cepstrum – MFCC

Mezi hlavní nevýhodu cepstra patří to, že nebere v ohled fyziologické vlastnosti lidského ucha. Lidské ucho má na nízkých kmitočtech větší rozlišení než na kmitočtech vyšších. Lidské ucho spolehlivě rozezná změnu kmitočtu z 50Hz na 60Hz, ale změnu z 10kHz na 10,05kHz vůbec nepostřehne. Proto se chceme při rozpoznávání řeči co nejvíce přiblížit cepstru slyšení.

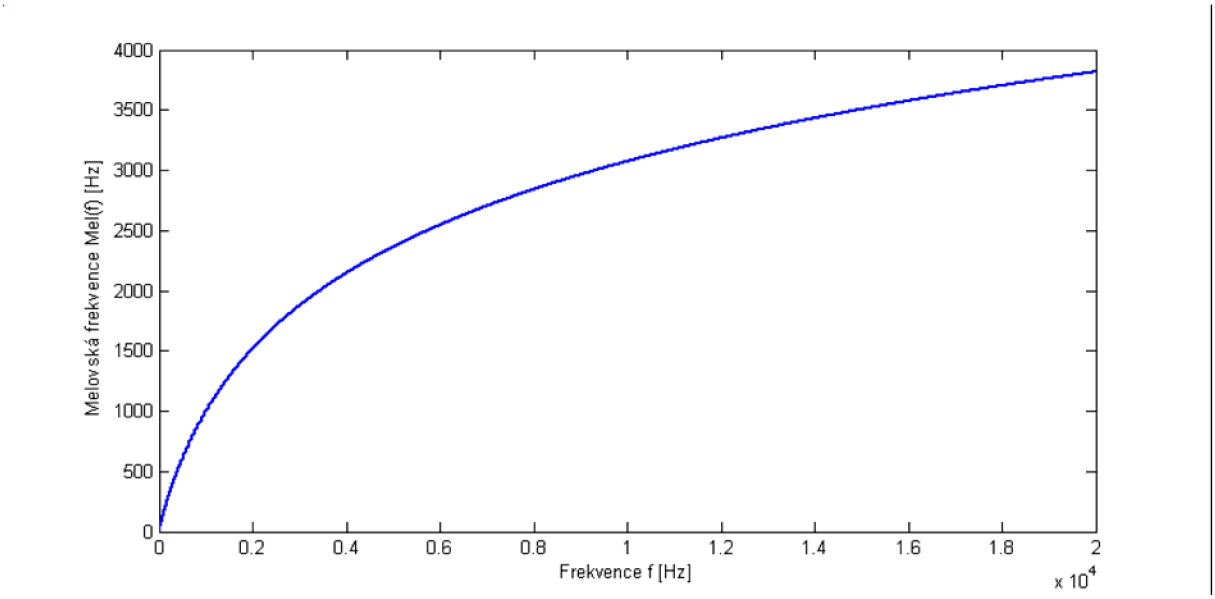
Jedním ze způsobů, jak docílit různého rozlišení na kmitočtové ose, je nelineárně na ni umístit filtry, změřit energii na jejich výstupu a použijeme je místo DFT při výpočtu cepstra.

Jiným způsobem je frekvenní osu nelineárně upravit a na upravené ose pak umístit filtry lineárně.

Používaná nelineární úprava při převodu Hertzů na Mely [8]:

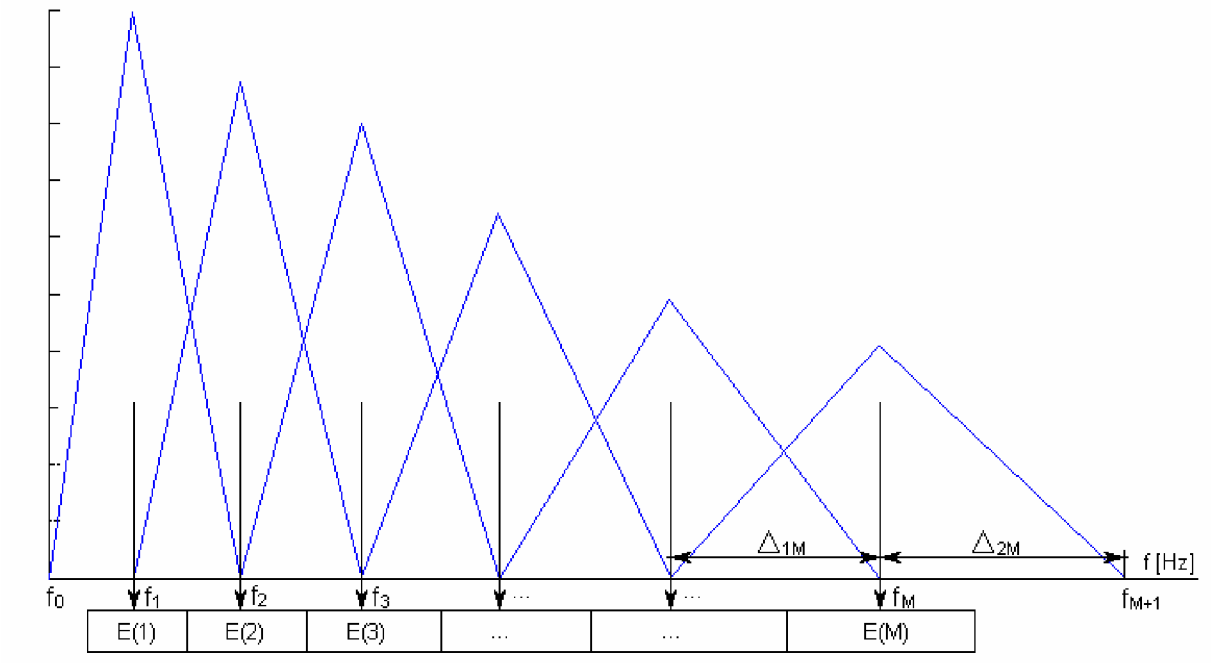
$$f_m = 2595 \log_{10}(1 + \frac{f}{700}) \ . \quad [\text{mel}] \qquad (3.8)$$

Grafická závislost potom vypadá:



Obr. 3.5.: Převodní charakteristika hertzů na mely.

Lineární rozmístění filtrů na Mel-ové ose má za následek nelineární rozmístění na standardní kmitočtové ose (obr. 3.6).



Obr. 3.6.: Pásma banky filtrů pro aplikaci melovské stupnice frekvencí.

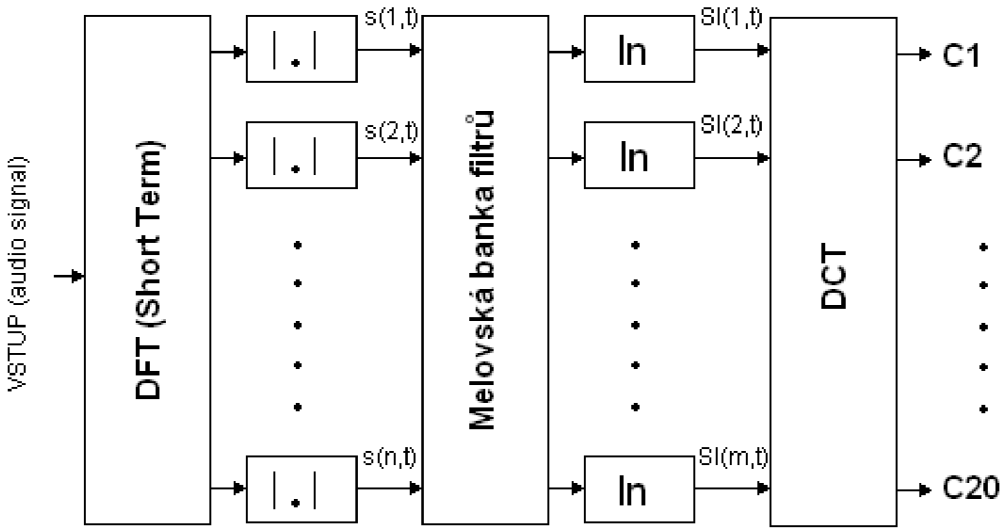
Výpočet energie lze provést dvěma způsoby:

1. Zkonstruujeme banku filtrů, vstupní signál filtrujeme v časové oblasti a počítáme energie. Tento způsob je složitý a nepoužívá se.

2. Provedeme DFT, umocníme, vynásobíme trojůhelníkovým oknem a sečteme. Zpětnou FT můžeme realizovat pomocí disktrétní cosinové transformace [8] (DCT).

$$c_{mf}(n) = \sum_{k=1}^{K} \log m_k \cdot \cos\left[n(k-0.5)\frac{\pi}{L} \right] \tag{3.9}$$

kde jsou Mel-frekvenční cepstrální koeficienty (MFCC), udává pořadí segmentu řeči, je celkový počet MFCC.



Obr.3.7.: Schéma výpočtu MFCC [8].

Postup určení melovských kepstrálních koeficientů je následující[4]:

1. Na vstup systému jsou přiváděny vzorky řečového signálu $s[n]$. Je provedena preemfáze a segmentace signálu na rámce délky 10-30 ms.

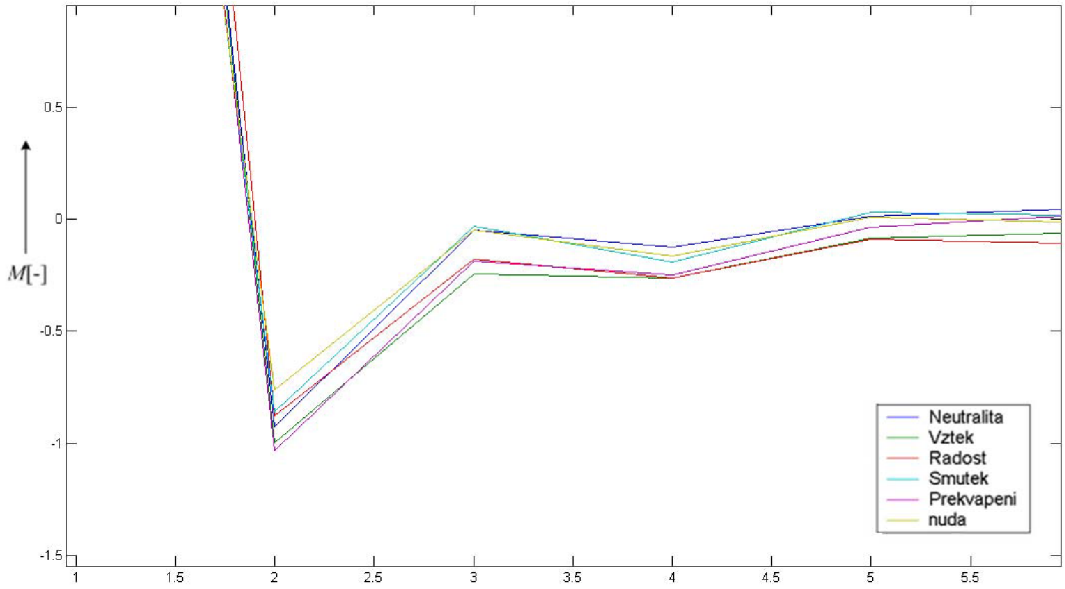
2. Pomocí FFT se vypočítá amplitudové spektrum $|S[k]|$.

3. Klíčová část celého procesu je melovská filtrace. Tato filtrace je realizována bankou trojúhelníkových filtrů s rovnoměrným rozložením středních frekvencí podél frekvenční osy s měřítkem v melovské škále. Počet pásem banky filtrů se volí v závislosti na počtu a umístění kritických pásem a to při respektování vzorkovacího kmitočtu f_{vz} a celkové šířky přenášeného pásma B_w. **V tabulce 4.2** jsou pro často užívané vzorkovací kmitočty a odpovídající frekvenční pásma uvedeny typické hodnoty počtu pásem neboli počet filtrů v melovské bance filtrů.

Trojúhelníkové filtry lze popsat následujícím způsobem

$$\begin{aligned} u(f,i) &= \frac{1}{b_i - b_{i-1}}(f - b_{i-1}) \quad \text{pro} \quad b_{i-1} \le f < b_i, \\ u(f,i) &= \frac{1}{b_i - b_{i-1}}(f - b_{i+1}) \quad \text{pro} \quad b_i \le f < b_{i+1}, \\ u(f,i) &= 0 \quad \text{pro} \quad \text{ostatní}. \end{aligned} \tag{3.10}$$

Průchod koeficientu FFT vstupního signálu trojúhelníkovými filtry znamená, že každý koeficient je násoben odpovídajícím ziskem filtru a výsledky pro příslušné filtry jsou sečteny.



Obr. 3.8.: Průběhy melovských koeficientů pro různé emoční stavy.

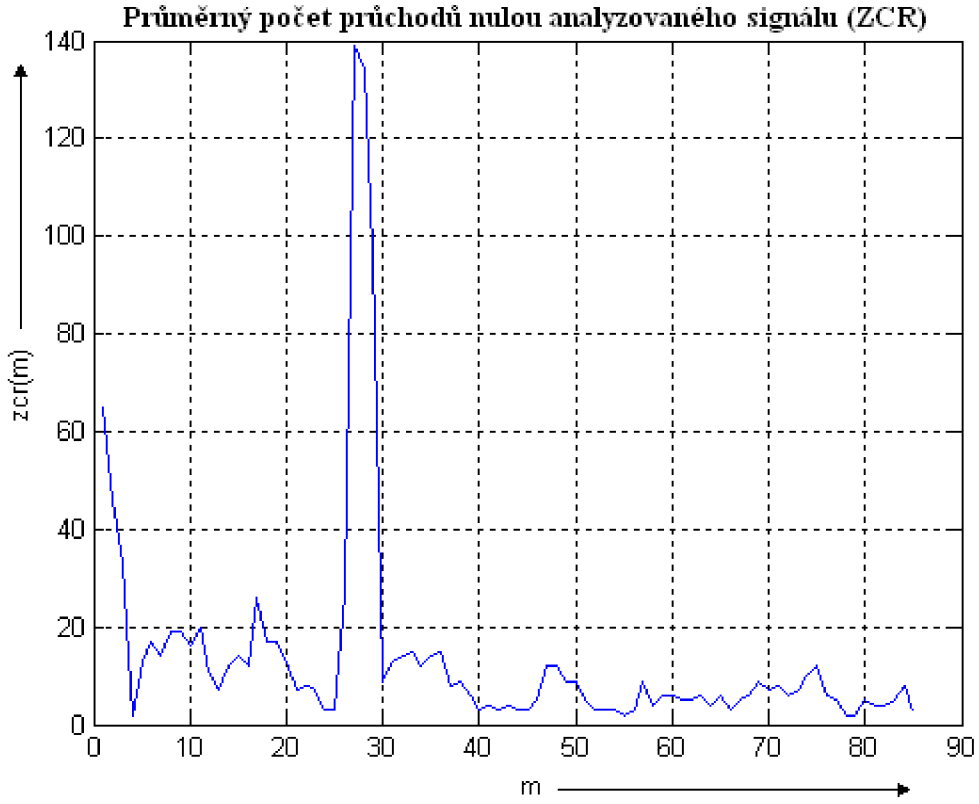
3.5. Střední počet průchodů signálu nulovou rovinou (ZCR)

Tento parametr je brán jako počet průchodů signálu nulovou rovinou, kterou můžeme s rezervou chápat jako jednoduchou charakteristiku popisující frekvenční vlastnosti řečového signálu, tato charakteristika se často označuje *zcr* (vycházející z anglického názvu), Její užití je motivováno tím, že v případě sinusového průběhu o frekvenci *f* je průměrný *zcr* 2*f*. Tento fakt musíme brát při práci s řečovým signálem jen s rezervou protože tento signál můžeme chápat spíše jako širokopásmový, funkci *zcr* lze definovat jako

$$zcr(m) = \sum_{n=0}^{N-1} \left|\mathrm{sgn}(s(n)) - \mathrm{sgn}(n-1)\right| w(m-n) \qquad [-] \qquad (3.11)$$

Kde N je délka jednotlivých rámců.

Z praktických testů bylo zjištěno, že při řešení problematiky rozpoznávání emočních stavu nebo obecně při zpracování řečového signálu může být hodnota *zcr* jakýmsi jednoduchým odhadem frekvence základního tónu. Na tuto hodnotu lze spolehnout pro hrubý odhad změny F_0. Pomocí referenční hodnoty lze zjistit zda došlo k nárůstu či poklesu střední hodnoty základního tónu řeči.



Obr. 3.9.: Střední počet průchodů signálu nulovou rovinou.

Kapitola 4

4. Emoční stavy mluvčího

Přibližně 10% informace přenášené řečovým signálem nese informaci o stavu mluvčího. Tato část je nejvíce ovlivňována psychickým stavem mluvčího.

Na téma emoce v mluvené řeči již bylo uděláno mnoho výzkumů, první z nich se datují až na počátek 20. století, kdy byly k určitým projevům v řeči přiřazovány jednotlivé emoce. První výzkumy byly spíše na poli psychologie, ale s počátkem syntézy řeči a automatického rozpoznávání mluvčích (ASR), bylo toto odvětví doprovázeno již více technickými aplikacemi.

Mnoho výzkumů na poli rozpoznávání řeči se v minulém desetiletí soustředilo na různé aspekty zpracování řečových signálů. Jedním z cílů bylo vytvořit syntetizovanou řeč více přirozenou, dalším z cílů bylo rozpoznat citový obsah z krátkého úseku řeči, i když byl význam z této věty odstraněn, tj. jedinou informací zde byla intonace, nezávisle na obsahu tohoto projevu.

Emoce dávají řeči barvu a dělají význam více komplexní. Jako posluchači reagují na emotivní stav mluvčích a přizpůsobují své chování závisející na druhu emocí, které mluvčí přednáší, tj. např. můžeme ukázat empatie smutným lidem, nebo jestli někdo váhá, pokoušíme se objasnit, co s kterými prostředky myslí nebo chce. Pro klasifikaci citového stavu mluvčího na základě prozodie a kvality hlasu musíme roztřídit zvukové rysy v řeči a přiřadit je k náležejícím emocím. Toto také předpokládá, že hlas opravdu nese plnou informaci o citovém stavu mluvčího. Tento předpoklad je často považován za samozřejmý.

Nicméně, nalezení akustických korelátů k emocím není zrovna jednoduché a také ne zrovna homogenní. Výsledky si někdy odporují a je těžké definovat, které údaje se vztahují k emotivní řeči.

V následujícím textu udělám krátký přehled nad výzkumem v oboru rozpoznávání emocí v řeči. Zaměřím se na tyto aspekty: řečová data, kategorie emocí a příznaky pro rozpoznávání emocí. [5].

4.1. Volba vhodných dat

V této části zmíním, jaké druhy dat lze použít při analýze emocí v řeči.

Prvním problémem stojícím před začátkem výzkumu emocí v řeči je problém výběru vhodných dat. Zpravidla materiály pro výzkum emocí v řeči dělíme do tří skupin:

- spontánní řeč
- předstíraná (hraná) řeč
- přivozená řeč

všechny tři skupiny mají svá pro i proti, a žádná ze skupin nemůže být považována za nejvýhodnější. V následující části provedu srovnání těchto skupin a uvedu příklady, ve kterých studiích byla která data použita.

4.1.1. Spontánní řeč

Pro použití spontánní řeči je často argumentováno tím, že obsahuje nejvíce přímé a autentické emoce. Sběr takovéhoto druhu řeči je ale velmi obtížný. V ideálních podmínkách by se mluvčí vůbec neměl dozvědět o tom, že je nahráván, aby se choval úplně přirozeně. Při tomto druhu sběru informací vyvstává otázka, zda je takovýto postup etický. Při jiném druhu sběru informací, např. záznam v televizi či rádiu, mohou nastat potíže s porušením autorského práva.

I když je obtížné shromažďovat tento druh řečových záznamů spontánní řeči, existují databáze s tímto druhem projevu, ale nejsou veřejně dostupné. Např. The Belfast database, The Leeds-Reading Emotion in Speech Corpus. Obě zde jmenované se skládají z klipů z televizních pořadů. Jiným příkladem je třeba JST databáze, v které jsou nahrávky přirozené řeči v přirozených situacích. Nebo databáze SUSAS, která obsahuje záznamy konverzace

pilotů v kokpitu, kteří jsou také v „přirozené situaci“, ale stále ne tak běžné, jako mnoho každodenních situací.

Letecká data, např. konverzace posádky v případech, kdy letadlo havaruje, je také užívána právě tak dobře, jako nahrávky radiového zpravodajství při pádu vzducholodi Hindenburg.

Dalšími výzkumy využívajícími spontánní řeč jsou např. Scherer a Ceshi (1997,2000), kteří používají nahrávky pasažerů čekajících na letišti na svá zavazadla.

4.1.2. Hraná řeč

Hraná řeč nemá stejné problémy při sběru dat jako řeč spontánní, ale je zde sporný stupeň přirozenosti. V úvahu se také musí vzít rozdíly v „hraní“. Některé sbírky hrané řeči jsou složeny z nahrávek pořízených profesionálními herci (např. Banse & Scherer, 1996, Mozziconacci, 1995, Scherer, 1996). V ostatních případech jsou použiti amatérští herci (např. Fairbanks & Hoaglin, 1941) nebo studenti divadelních akademií (např. Green & Cliff, 1975), nebo jen studenti (např. Levy, 1964). Samozřejmě kvalita záznamu se mezi takto pořízenými nahrávkami výrazně liší.

V první řadě věrohodnost hraného emočního projevu je závislá na kvalitě řečníka, který určuje projev jednotlivých emocí. Nicméně jsou zde ještě nezodpovězené otázky ohledně využití hrané řeči. Jedna z nich je, zda hraná řeč je skutečně schopna odrážet autentické emoce. Je jasné,že hraná řeč obsahuje emoce, ale jak významné jsou? Hranou řeč je jednodušší ovládat, ale na úkor přirozenosti.

Stibbard (2001) uvedl, že hraná řeč je pouze představa o tom, jak lidé věří, že by se emoce měli v řeči projevovat, ne jak se skutečně v řeči projevují. Také ukazuje, že řeč je více stereotypní a že výraz emocí je více extrémní než ve spontánní řeči.

Pro syntézu řeči by tyto nevýhody nemuseli znamenat problém. Je vhodné, když syntetizovaná řeč dostává více originální a lépe vyložitelný emotivní nádech, namísto skutečných, ale komplexních a nesnadno interpretovatelných emocí. Stupeň stereotypicity nemůže být příliš vysoký, pak by její znění připadalo komické. Nicméně při rozpoznávání řeči

a emocí mluvčího způsobuje tato záměna mezi ideálem a realitou značné problémy. Při rozpoznávání řeči musíme zvládat složitost reálné řeči.

4.1.3. Přivozená řeč

Při přivozené řeči se využívá nápadu, že jisté emoce mluvčímu přivodíme. Postup může být takový, že subjekty sledují film, který by měl vyvolávat specifické emoce. Poté musí znovu převyprávět obsah celého filmu výzkumnému pracovníkovi. Představa je ta, že řeč bude zabarvena přivozenými emocemi. Je také možné umístit subjekt do situace vyvolávající dané emoce a vytvořit záznam jeho/její řeči. Nicméně tato metoda trpí obdobnými etickými problémy, jako spontánní řeč, tj. jestli je etické někoho vyděsit a nahrávat jeho řeč. Možná je to ještě více neetické, než nahrávat někoho, kdo je skutečně vyděšen. Výsledkem této úlohy jsou vyvolané nebo přivozené emoce příliš mírné, a když nejsou mírné, jsou často neetické.

Tato metoda má také své pozitivní rysy, které dávají kontrolu nad podnětem, na druhé straně různé osoby mohou reagovat odlišně na stejné podněty. Přesnost, s jakou byla vyvolána nebo přivozena řeč, závisí z velké části na tom, jak úspěšný „vyvolávací“ proces byl.

Studie, které využívají vynucenou emoční řeč jsou např. Skinner (1935), Friedhoff(1962), Hecker (1968), Iida (1998).

4.1.4. Shrnutí

Jaký druh dat je nejlepší použít? Odpověď na otázku není jednoduchá. Na jedné straně je spontánní řeč nejvíce věrohodný zdroj, ale jak ukázal Stibbard (2001), také spontánní řeč je svým způsobem omezena např. společenskými omezeními. Čisté „výbuchy“ emocí, které např. můžeme vidět u malých dětí, nejsou společensky akceptovány jako výraz hněvu, smutku, štěstí, atd. Jak dospíváme, očekává se, že se naučíme ovládat po jazykové stránce své emoce. Nicméně stále se zajímáme o jemné stopy, které po těchto emocích mohly zůstat v hlase.

Hraná řeč dává rozsáhlé možnosti, jak ovládat řeč, ale není skutečně přirozená. Ale zdá se, že výzkumníci nejsou plně přesvědčeni, zda ji soudit jako špatnou nebo dobrou.

4.2. Emoce a jejich dělení

Asi nejjednodušší přístup k popisu emocí je použít kategorie používané v běžném hovorovém jazyce – označení jako strach, hněv, radost atp. Toto dělení dává na první pohled různé způsoby, jaké kategorie mohou být použity pro popis emocí a emočních stavů.

Hlavní téma této části je spojení mezi běžným dělením emocí a problémy s jejich ohraničením. Běžné třídění (jako používáme v jazyce) je ošidný složitý systém, který byl vyvinut k popisu mimořádně složitého souboru projevů

4.2.1. Základní emoce

Pravděpodobně nejznámější představa u emočního výzkumu je, že jistá kategorie emocí je primární a ostatní jsou sekundární.

Představa primárních emocí měla velký vliv na popis emocí. Představa je, že přirozené východisko pro výzkum je získat seznam primárních emocí a pak zkoumat, jak se odráží v řeči. Často se předpokládá, že jednou budou sekundární emoce spadat také na toto místo.

Představou je, že emoce se dělí na dvě části, primární a sekundární. První skupina obsahuje takové emoce, které jsou „čisté“ a „jednoduché“. Druhá skupina emocí je odvozena od skupiny první. Obsahuje emoční stavy, které jsou odvozeny z primárních emocí jejich smícháním (tak jako se míchají základní barvy).

Široký okruh teorií se shoduje, že „čistokrevné“ emoce mají jen několik forem, které jsou kvalitativně od sebe odlišné. Každá forma má příznaky, kterými se od ostatních odlišuje.

Tato situace je důsledkem pro sestavení základního seznamu emocí. Na jedné straně, když základní emoce mají kvalitativně odlišné příznaky, tak vhodná cesta, jak je popsat, je seznam těchto kategorií. Na druhé straně pohled na základní emoce nenabídne podporu pro představu, že znalosti o položkách na seznamu budou přeneseny v přímém směru do jiného emočního stavu.

Zkrátka seznam základních emočních kategorií je jen startovací bod pro výzkum, který má za cíl zkoumat řečové vzory přidružené k základním emocím.

Stojí za zmínku, že neexistuje žádný konečný seznam základních emocí. Je to jenom dohoda „velké pětky“ - strach, vztek, štěstí/radost, smutek, nuda. Běžně se stává, že základní skupina obsahuje též překvapení, teplý a studený hněv, pohrdání a lásku, kterou můžeme dělit na sexuální a další typy.

Pro tuto práci jsem si vybral analýzu 6 emočních stavů včetně neutrální promluvy. A to jelikož jsem pracoval již z hotovými databázemi nahrávek českých emočních stavů.

A to tyhle emoční stavy:

- neutrální projev
- vztek
- smutek
- radost
- překvapení
- nuda

4.2.2. Sekundární emoce

Běžný jazyk obsahuje velké množství názvů pro různé druhy emocí. Např. sbírka, kterou spravoval Whissell [5] obsahuje 107 slov popisujících emoční stavy. Nebo Plutchik [5] vedl seznam obsahující 142 různých slov. Tyto slova pokrývají velmi velký rozsah emočních stavů, velmi málo z nich by však mohlo být považováno za základní emoce.

Užívaný termín pro emoční stavy, které nejsou základní, je „sekundární emoce". Tento výraz značí ne příliš šťastný význam, že tyto slova jsou méně důležitá. Emoce druhé v pořadí zní přirozeněji, odráží se v nich rozumný předpoklad, že jsou tyto emoce v některém smyslu více komplexní než emoce základní.

Žal

Žal se projevuje nižším průměrným základním tónem řeči a rozsah hondot zkladního tónu řeči je zde nejmenší. Projevuje se pomalým tempem řeči s dlouhými přestávkami a prodlužováním samohlásek.

Zalíbení / něžnost

Tyto emoce jsou charakterizovány nižším základním tónem řeči F_0, jemným průběhem hlasitosti, zvučným témbrem, pomalým tempem řeči a konstantní nebo mírně rostoucí intonací řeči. Je zde také pravidelná a rytmická výslovnost.

Sarkasmus / ironie

Sarkasmus je charakterizován rozporem mezi slovní a mimoslovní úrovní. V řeči se projevuje pomalejším tempem.

Překvapení / údiv

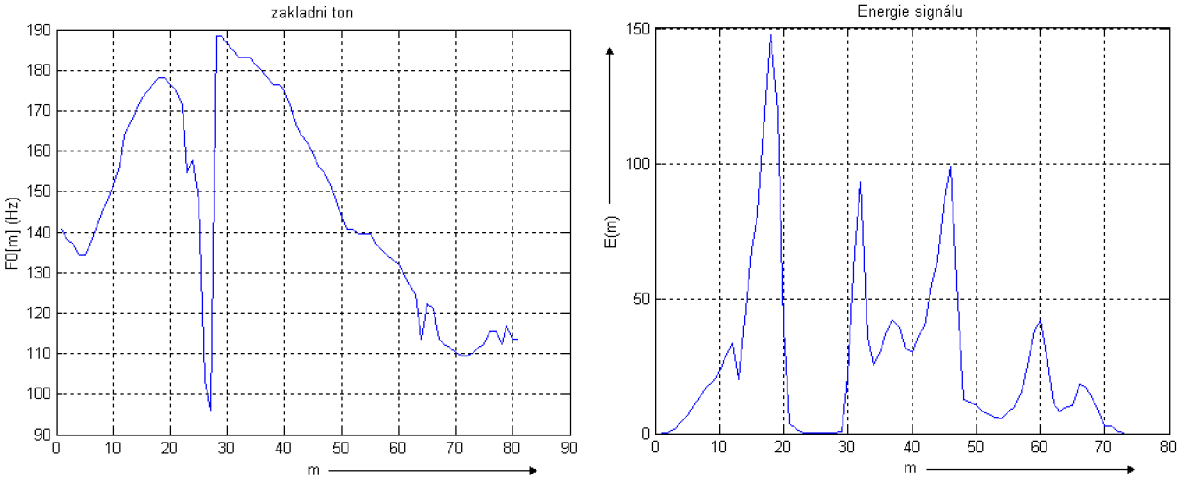
Projevuje se rychlejším tempem. Intonace charakteristická pro překvapení obsahuje jeden nebo dva vrcholy, kde F_0 dosahuje několikanásobných hodnot oproti F_0 v neutrálních větách.

Nenávist / odpor / pohrdání

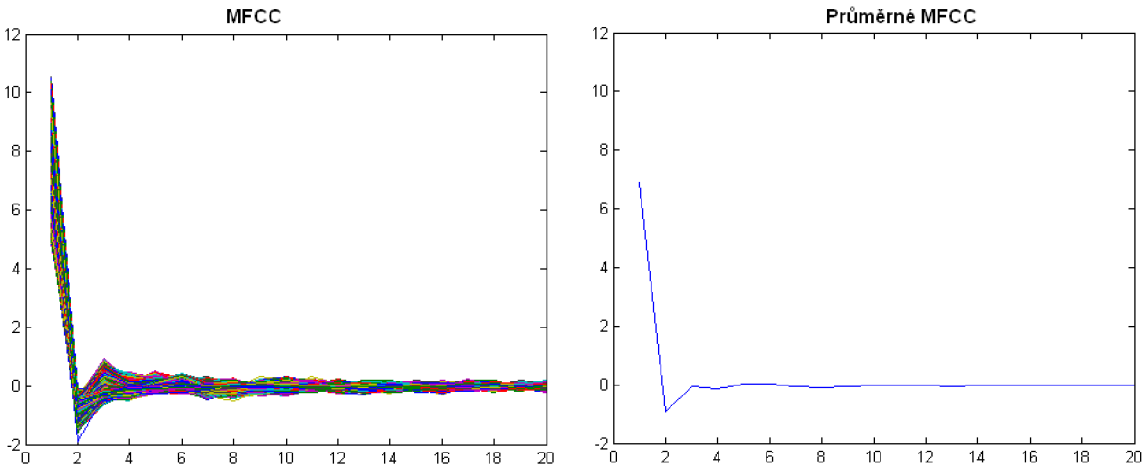
Tyto emoce jsou charakterizovány větší šířkou pásma formantů než u neutrální řeči. Hovorová rychlost je nižší, je to způsobeno prodloužením řeči a většími přestávkami mezi slovy. Pohrdání je charakterizováno sestupnou intonací.

4.3. Neutralita

Neutrální projev je brán jako referenční pro výzkum emocí. K jednotlivým emocím v této práci hledáme odchylky od neutrálního emočního stavu. Maxima frekvence se obvykle nacházejí uprostřed analyzované nahrávky a také nýbrž minima se často pohybují v okolí středu. Neutrální projev má spíše klesající průběh F_0 a to zejména po dosažení svého maxima.



Obr. 4.1.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu - Neutralita.



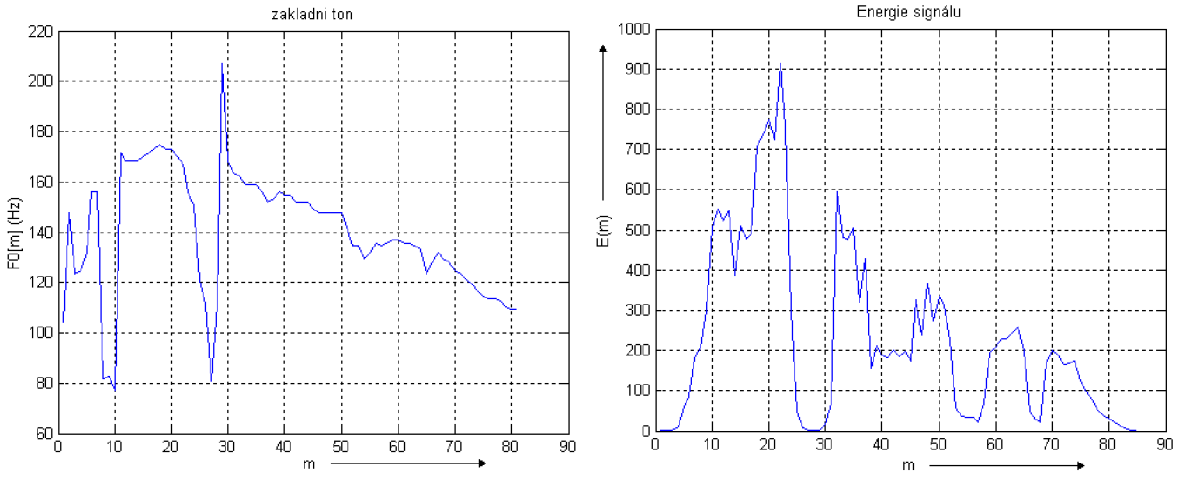
Obr. 4.2.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu - Neutralita.

4.4. Vztek

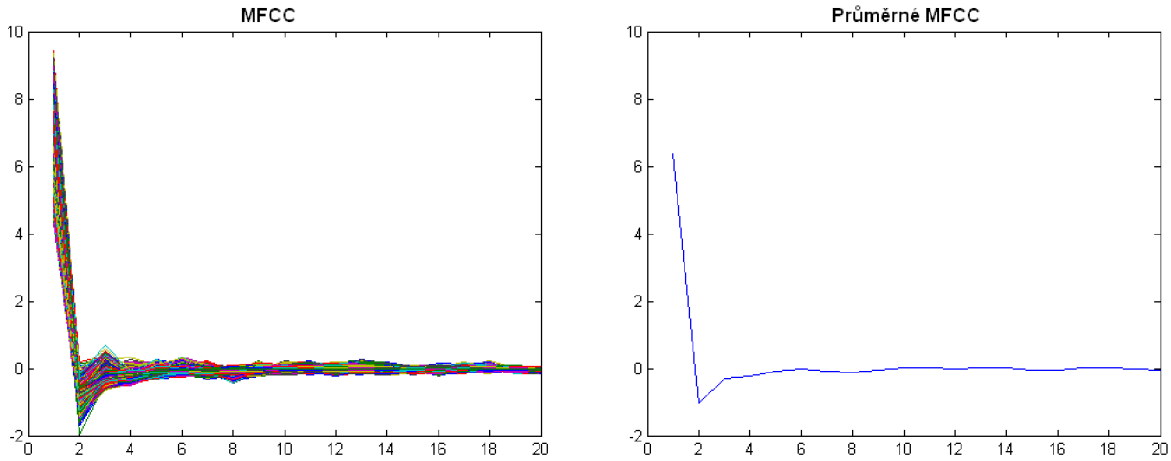
Vztek je emoční kategorie, kde výsledky výzkumu obou typů řeči (jak spontánní, tak i vyvolané), dávají stejné výsledky.

Při vzteku je průměrný základní tón vyšší než v neutrální promluvě. Také jeho rozsah je větší nežli u normální řeči a zpravidla o přibližně 15%.

Průběh intenzity je spíše klesavý, kdy mluvčí dává důraz na začátek promluvy a tím mluvčí dává hned ze začátku věty důraz na jeho nespokojenost či jiný aspekt.



Obr. 4.3.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu – Vztek



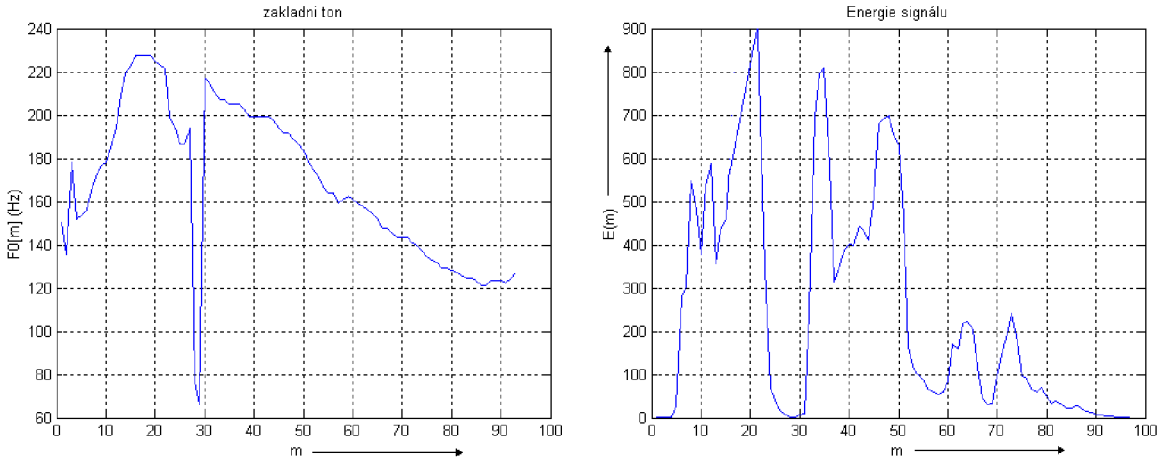
Obr. 4.4.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu - Vztek.

4.5. Radost

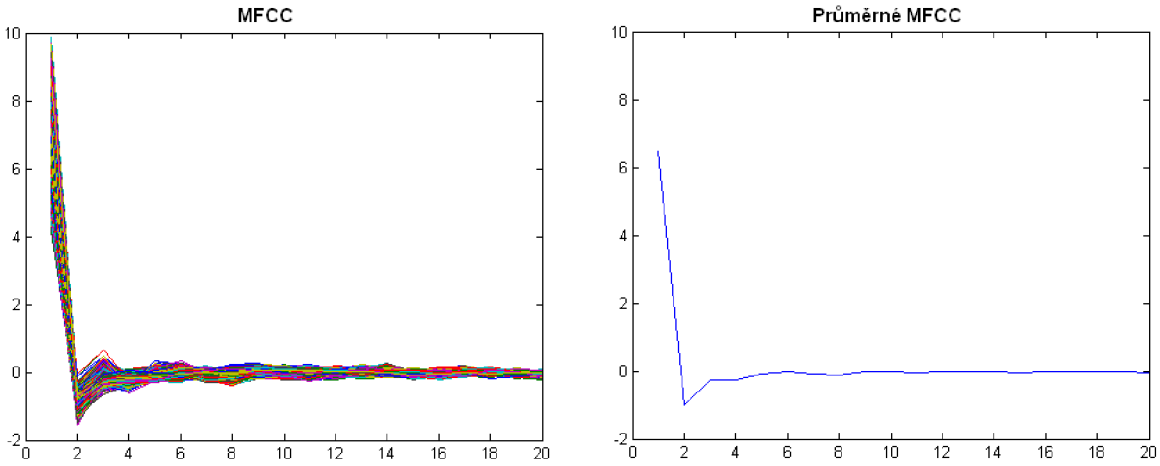
Obecně se výzkum soustředí na více forem radosti či štěstí, ale i tak se oznámené nálezy shodují u většiny, kdo výzkum prováděli.

Radost je v češtině realizována výraznou melodičností a variacemi intonace a intenzity. Průměrná hodnota základního tónu F0 je vždy vyšší než průměrná hodnota F0 při neutrální promluvě. Intenzita je první části rostoucí a dále pak již klesá. Rozsah hodnot základního tónu řeči je větší a průměrná hodnota intenzity je od neutrálního stavu větší přibližně o 3-5dB.

Se vzrůstající hodnotou rychlosti se zvyšuje i intenzita. Tempo bývá obvykle rychlejší, ale není to podmínkou a záleží na již zmiňované formě radosti či štěstí. Intonace není tak výrazně vyšší nežli u vzteku.



. Obr. 4.5.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu – Radost.



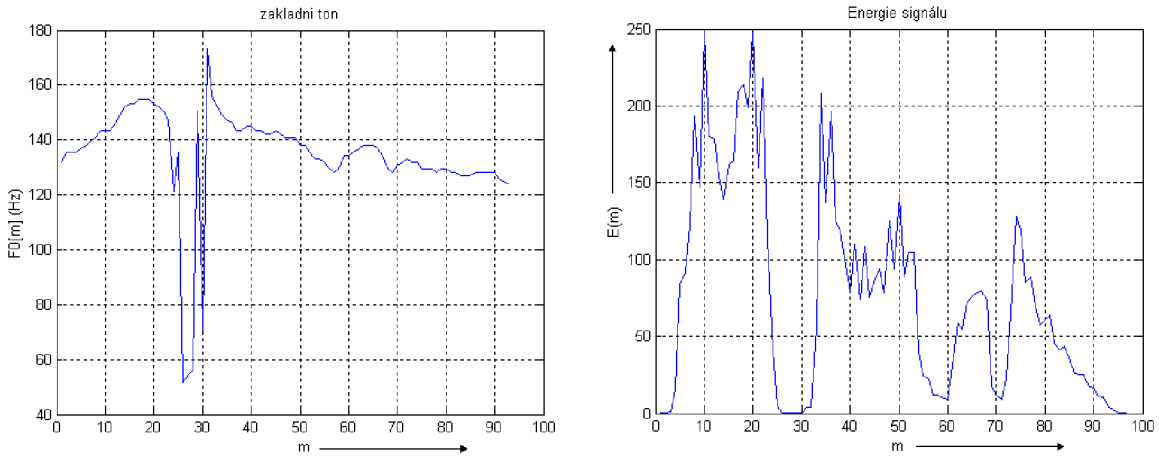
Obr. 4.6.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu – Radost.

4.6. Smutek

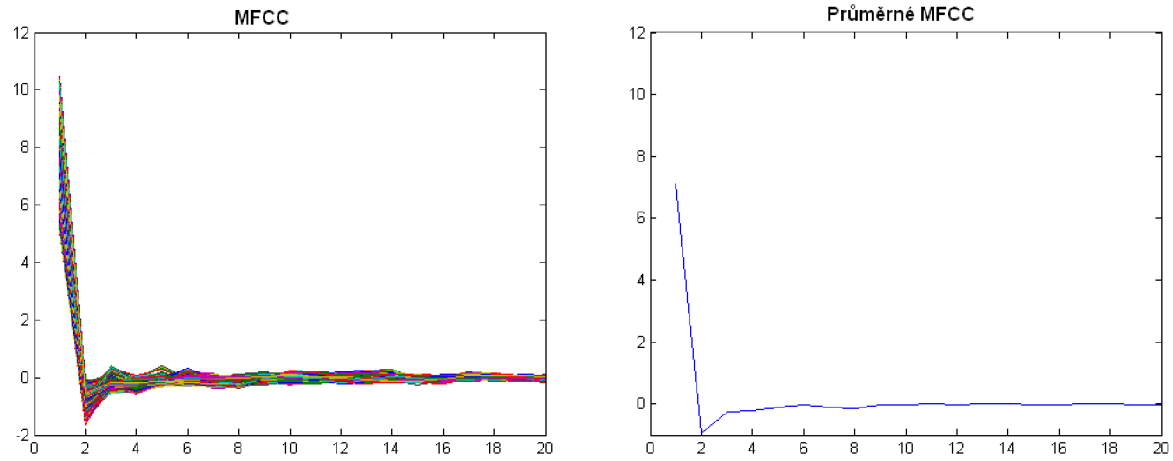
Pocity smutku se v řeči projevují stejným nebo nižším základním tónem řeči F_0 jako u neutrální řeči jen přibližně v rozsahu –3% . Rozsah hodnot základního tónu je také menší a tempo řeči je pozvolnější.

Po akustické stránce je velice podobný poslednímu zkoumanému emočnímu stavu a to nudě, akorát průměrné F_0 je vyšší než v neutrálním vyjádření. Tento rozdíl spolu s labializovanou artikulací v češtině, umožňuje posluchačům rozdíl mezi smutkem a nudou. Nicméně je velice těžké tohoto také docílit programově. Často se zamění výsledek emočního stavu smutek za nudu.

Tempo je blízké normálnímu projevu a také intenzita je zde slabší a monotónní. Pozice maxim intenzity i základného tónu se ve většině případů vyskytuje na začátku promluvy. U tohoto emočního stavu je nejlepší volbou příznakový vektorů volit intenzitu signálu, která je pro tento emoční stav monotónní, základní tón řeči (pokles) a také rozsah F_0 (menší). Kepstrální koeficienty se od neutrálního projevu téměř nelišší, proto je zbytečné je u tohoto emočního stavu brát jako směrodatné při rozpoznávání, ale spíše jen jako potvrzení správnosti rozpoznaného emočního stavu.



Obr.4.7.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu – Smutek.

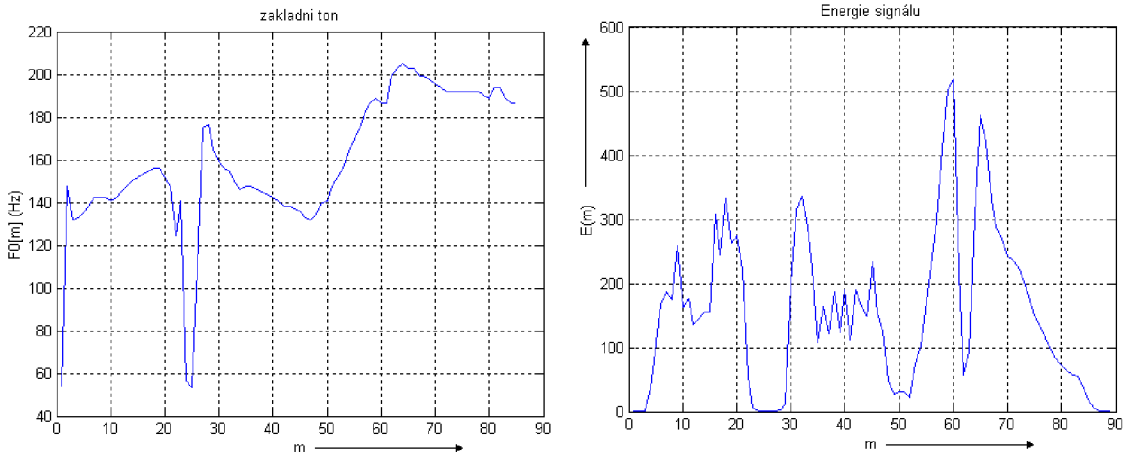


Obr. 4.8.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu – Smutek.

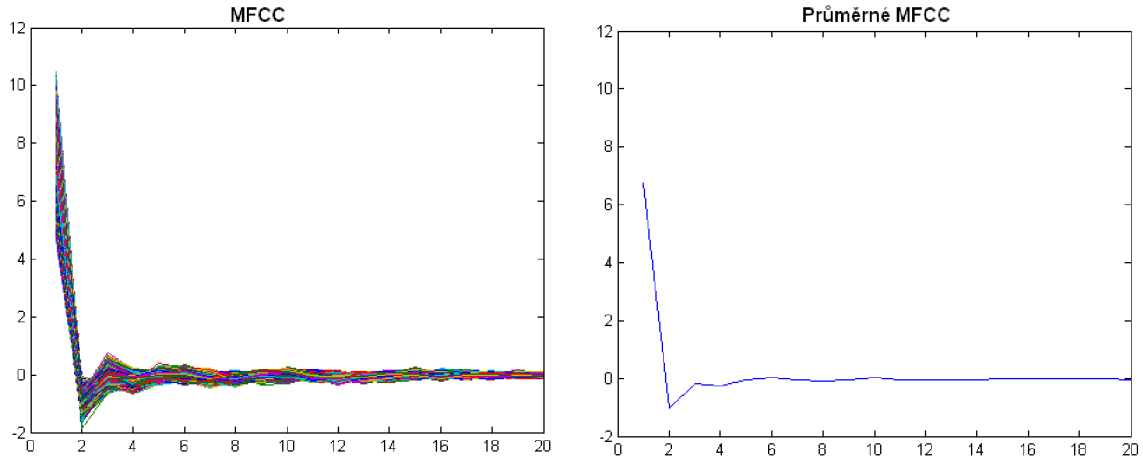
4.7. Překvapení

Tento emoční stav se liší od ostatních v mnoha faktorech, tím pádem byl pro rozpoznáván asi nejméně náročný. Prozodické vlastnosti jsou zde výrazné a to například maximum a minimum frekvence základního tónu řeči, kde se navíc maxima vyskytovala zejména na konci řeči jak je tomu patrné pro tento emoční stav podobný v tomto ohledu na řečnickou otázku, která je zakončena na konci také vyšší intenzitou a frekvencí základního tónu. Také minima se často vyskytovali na začátku promluvy.

Dalším dobře odlišným prvkem při zkoumání emočního stavu překvapení bylo nalezeno ve variabilitě jak energie tak i frekvence základního tónu. Kepstrální koeficienty zde byli celkem podobné normální promluvě a tedy nejsou nijak důležité pro rozpoznání tohoto emočního stavu.



Obr.4.9.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu – Překvapení.



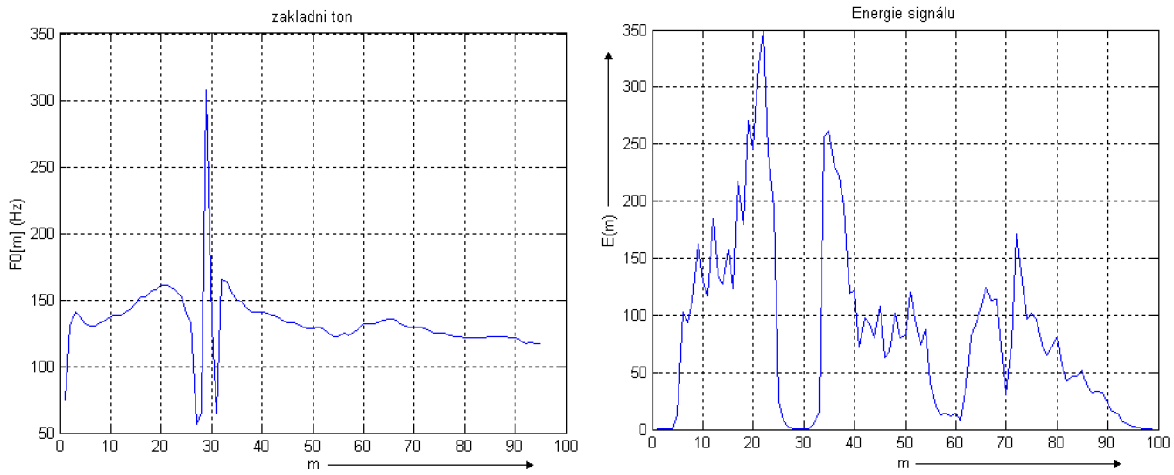
Obr. 4.10.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu – Překvapení.

4.8. Nuda

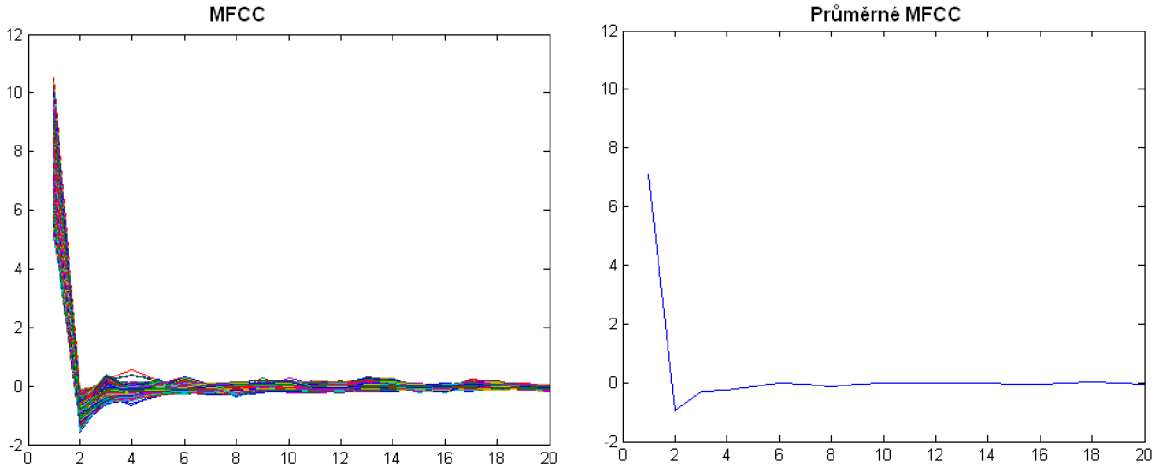
Tento emoční stav se projevuje velice podobnými parametry jako emoční stav smutek, jak jsem již psal výše.

Emoční stav nuda se v řeči projevují stejným nebo nižším základním tónem řeči F_0 jako u neutrální řeči a to jen přibližně v rozsahu –2% . Rozsah hodnot základního tónu je také menší a tempo řeči je pozvolnější. Po akustické stránce je velice podobný poslednímu zkoumanému emočnímu stavu a to nudě, akorát průměrné F_0 je vyšší než v neutrálním vyjádření. Tento rozdíl spolu s labializovanou artikulací v češtině, umožňuje posluchačům rozdíl mezi smutkem a nudou. Nicméně je velice těžké tohoto také docílit programově. Často se zamění výsledek emočního stavu smutek za nudu.

Tempo je blízké normálnímu projevu a také intenzita je zde slabší a monotónní. Pozice maxim intenzity i základného tónu se ve většině případů vyskytuje na začátku promluvy. U tohoto emočního stavu je nejlepší volbou příznakových vektorů volit intenzitu signálu, která je pro tento emoční stav monotónní, základní tón řeči (pokles) a také rozsah (menší). Kepstrální koeficienty se od neutrálního projevu téměř nelišší, proto je zbytečné je u tohoto emočního stavu brát jako směrodatné při rozpoznávání, ale spíše jen jako potvrzení správnosti.



Obr. 4.11.: Příklad průběhů intonačních křivek (vlevo) a energií (vpravo) Em.stavu – Nuda.



Obr. 4.12.: Příklad průběhů kepstrálních koeficientů (vlevo) a jejich středních hodnot (vpravo) Em.stavu – Nuda.

4.9. Shrnutí

Jako vhodné příznaky pro rozpoznávání emočního stavu mluvčího se jeví základní tón řeči a rozsah jeho hodnot, průběh intenzity signálu na jednotlivých fonémech, energie ve spektru a mel-kepstrální koeficienty.

Nalezení jednoho příznaku pro spolehlivé rozpoznání emočního stavu mluvčího je takřka nemožné, vždy je třeba použít příznakový vektor složený z více příznaků. A podle jejich kombinace a jejich vzdáleností pak určit nejpravděpodobnější stav mluvčího.

Tab. 1.: Vyhodnocení průměrů prozodických rysů zkoumaných signálů

Emoční stav		Neutral	Vztek	Radost	Smutek	Překvapení	Nuda
$F0_{mean}$	Hz	121,03	129,91	150,66	120,43	138,49	123,87
$F0_{max}$	Hz	336,8	336,12	368,91	390,57	324,56	342,56
$F0_{min}$	Hz	55	56,55	55,21	50,5	59,12	54,18
$F0_{max}$-$F0_{min}$	Hz	281	279,54	313,7	340	256,43	288,38
V_{F0}	Hz	13	12	17,2	14,3	12,2	11,1
σ_{F0}	-	41,3	39,9	48,7	45,2	42,7	36,3
$F0_{maxpos}$	%	44	45	37	44	61	31
$F0_{minpos}$	%	50	41	39	48	44	39
E_{mean}	-	127,76	231,9	131,55	55,02	89,15	113,45
E_{max}	-	505,91	1275,3	580,6	319,5	363,5	450,3
E_{min}	-	0,001	0,0117	0,002	0,0043	0,0024	0,0077
E_{max}-E_{min}	-	112,13	119,1	126,33	112,21	119	116,65
V_E	dB	340	821,26	374,34	221,15	167,9	230,37
σ_E	-	0+15375i	0+23297i	0+14877i	0+5855i	0+9618i	0+12569i
E_{maxpos}	%	20	29	39,5	30	35	25
E_{minpos}	%	44	45	50	50	60	36
ZCR_{mean}	-	26,4	31,3	32	28,7	29,8	26,2
ZCR_{max}	-	146	161,5	171,3	152,4	155,5	156,4
ZCR_{min}	-	1,3	1,53	1,98	1,44	2,05	1,66

Tab. 2.: Srovnání jednotlivých emočních stavů vůči neutrální promluvě

Emoční stav	Vztek	Radost	Smutek	Překvapení	Nuda
Střední hodnota F0	↑↑	↑↑	↓	↑↑	↑
Variabilita F0	↑	↑↑	↑	↑	=
Střední hodnota Intenzity	↑↑	↑	↓	↓↓	↓
Variabilita Intenzity	↑↑	=	↓↓	↓	↓↓
Počet průchodů nulou	↓	↓	↓↓	↓↓	↓↓

↑↑ Větší nárůst
↑ Menší nárůst
= velice podobné
↓ Menší pokles
↓↓ Větší pokles

Kapitola 5

5. Databáze

V této kapitole se budeme zabývat databázemi, které jsou pro řešení tohoto problému velice důležité. Pro určení správných rozhodovacích členů je největší požadavek kladen právě na správný výběr a správné rozčlenění emoční řečové databáze. Zde je uveden přehled databází emoční řeči, pokud je databáze volně dostupná, je u ní uveden odkaz ke stažení. V popisu jednotlivých databází je uveden jazyk, počet mluvčích, velikost, typ záznamu a popis tvorby záznamů. Popis jednotlivých databází je k nalezení zde [6].

5.1.Seznam existujících databází emoční řeči

Geneva Airport Lost Luggage Study (Scherer & Ceschi, 1997, 2000)

Audiovizuální databáze obsahující nahrávky pasažérů na letišti v Ženevě u výdeje zapomenutých zavazadel. Zde byly pořizovány nahrávky rozhovorů s pasažéry.

URL: http://www.unige.ch/fapse/emotion
Počet mluvčích: 109
Emoce: hněv, lhostejnost, stres, smutek, dobrá nálada
Druh záznamu: přirozený

SALAS database

Databáze obsahující audiovizuální záznamy. Různých emočních stavů u mluvčího bylo dosaženo interakcí s různými osobnostmi posluchačů. Tato databáze obsahuje velký počet různých emocí, ale jejich projevy nejsou příliš intenzivní.

URL: http://www.image.ntua.gr/ermis
Jazyk: angličtina, řečtina
Počet mluvčích: 20
Emoce: velký rozsah
Druh záznamu: přivozený

Capital Bank Service and Stock Exchange Customer Service

Ve svých výzkumech ji využili Devillers & Vasilescu (2004), obsahuje hlavně negativní emoce (strach, hněv, stres). Záznamy jsou pořizovány na call centrech.

Jazyk: angličtina
Emoce: strach, hněv, stres
Druh záznamu: přirozený

5.2. Vlastní databáze českých mluvčích

Pro výzkum emočních stavů je zapotřebí zkoumat emoční stavy na vzorcích, které je možno jednoznačně přiřadit k jednotlivým emočním stavům. Jako nejlepší volba se jeví vlastní databáze česky mluvených vzorků. Bohužel není jednoduché takovou databázi vytvořit. A jelikož ještě neexistuje žádná pořádná databáze českých emočních stavů použil jsem databázi vytvořenou studentem školy VUT v roce 2007 Ing.Hichamem Atassim, kterou jsem doplnil o některé vlastní nahrávky, které jsou ale lehce s příměsí šumu. Tento šum, který při nahrávání těchto emočních stavů ovlivnili okolní aspekty a fakt, že nebyli nahrány v odhlučněné komoře nýbrž jen v uzavřeném pokoji profesionální technikou zapůjčené od muzikální kapely, které patří můj dík.

Jediným problémem při použité již této existující databáze byl fakt, že nelze analyzovat více emočních stavů jelikož sehnat tyto ostatní emoční stavy od již nahraných mluvčích je takřka nemožné. Z tohoto důvodu jsem se rozhodl rozdělit analýzu na tyto již nahrané emoční stavy, které jsem jen doplnil o pár vlastních nahrávek pro testování nahrávek s lehce větší hladinou šumu.

Jazyk: čeština
Počet mluvčích: 12
Emoce: Neutrální, vztek, smutek, radost, překvapení, nuda
Druh záznamu: přirozený

Problémy při tvorbě emoční řečové databáze:

- Řečový signál má horší kvalitu, neboť je na pozadí podbarven rušivými elementy (hudba, publikum...). Sehnat „čistý“ řečový záznam, ve kterém jsou ještě obsaženy emoce, je velmi časově náročné.
- Pro různé emoce se text řečený mluvčím liší. Délka záznamu je ale příliš krátká pro statistické příznaky.
- Možnost sehnat pro jednotlivé mluvčí jen některé druhy emocí.
- Pro výzkum řečových signálů je vhodné, když je záznam pořizován stejnou aparaturou (stejný mikrofon, zvuková karta, prostředí...), což při sběru těchto dat není možno docílit.
- autentičnost nahrávek je třeba ověřit na nezávislých posluchačích, kteří formou testu přiřadí k nahrávkám jednotlivé emoce. Pokud dojde k neshodě mezi posluchači, musí být nahrávka vyřazena.

Kapitola 6

6. Rozpoznávání emočních stavů

Již od dávných dob se plno odborníků zabývalo touto problematikou k docílení nejlepších výsledků v oblasti analýzy řeči. Řeč je základním parametrem k přenosu informací mezi lidmi. Existuje již od pradávna a touto formou sdělíme jednak obsah informace,ale také pro nás již samozřejmé také emoční vyjádření této informace. Často se zapomíná na to,že stejná věta vyslovená v jiném emočním stavu může ve výsledku mít úplně odlišný význam. Je více důvodů proč lidé chtějí tyto emoční stavy zkoumat a snažit se jej co nejdokonaleji analyzovat jak už pro vědecké či komerční účely.

Je plno systémů,které již dokáží analyzovat z řečového signálu např. pohlaví, věk či systémy TTS a STT apod., většina těchto systémů byla navrhnuta zejména ke komunikaci člověka s počítačem a naopak. Tyto systémy se dají využít zejména také v mobilní komunikaci, kdy můžete například ovládat svůj mobilní telefon svým hlasem nebo různé zařízení v domácnosti. Systémy TTS a STT umožňují počítači přečíst celou knihu před audio výstup či můžete číst knihu do mikrofonu a ten potom zpracuje řečový signál a převede vám jej do textového formátu a uloží jej v počítači (systém STT – Speach to Text – v překladu z Angličtiny: Řeč do textu)

V dnešní době patří zpracování řeči a její následná analýza k velmi rozebíranému tématu a většina technických univerzit již v této oblasti pokročila natolik,že se již začínají objevovat první implementace systémů do reálného světa. Například v Americe již systémy využívají na zákaznických linkách k zjištění a statistice spokojenosti svých zákazníků. Velice lehce zde jde potom analyzovat zdali volal zákazník spokojený či nespokojený a zdali na konci rozhovoru se jeho emoční stav změnil či ne. Také firma SONY, APPLE a plno dalších již využívají tento druh rozpoznávačů k různým účelům.

Navržený automatický rozpoznávač emočních stavů lze využít pro řadu aplikací jako například:

- Zábava: Na trh se čím dál víc dostávají osobní roboti, kteří jsou určeni pro domácí práce, zábavu a další možné účely. Nejvíce potřebné je zde kladen důraz na komunikaci mezi robotem a člověkem. Pro lepší přirozenost robota v domácnosti je také jeho vystupování a to hlavně v lidskosti jeho řečového projevu. A proto kladou důraz na analýzu a jeho zpětný reprodukovaný hlas.
- Bezpečnost: Na základě nahrávek telefonního rozhovorů lze dodatečně zjistit jakým způsobem byl hovor nebezpečný či jinak podstatný pro daný účel zjišťování. Ať už se jedná o policejní či komerční účely. Také lze tento systém využít u tzv. virtuálního psychiatra, který v podobě počítače může analyzovat váš aktuální emoční stav a také lze potom vyhodnotit zdali vám pomohl či ne.

6.1. Matematické vyjádření použitých rysů

Zde ukázka výpočtů jednotlivých prozodických rysů. Zaměříme se speciálně na tři základní a to: Frekvence základního tónu řeči, Intenzita(Energie) a střední hodnota počtu průchodů nulou.

6.1.1. Frekvence základního tónu řeči

Střední hodnota frekvence základního tónu

$$F_{0\,mean} = \frac{1}{N}\sum_{i=0}^{M-1} F_0[i] \quad \text{[Hz]}, \qquad (6.1)$$

kde N je počet rámců,
$F_0[i]$ je frekvence základního tónu $i$tého rámce.

Maximální hodnota frekvence základního tónu

$$F_{0\max} = \max(\overline{F_0}) \quad \text{[Hz]}. \tag{6.2}$$

kde $\overline{F_0}$ je vektor frekvencí základního tónu vypočítané pro jednotlivé rámce.

Minimální hodnota frekvence základního tónu

$$F_{0\min} = \min(\overline{F_0}) \quad \text{[Hz]}. \tag{6.3}$$

Rozdíl maximální a minimální hodnoty frekvence základního tónu

$$F_{0\max-\min} = F_{0\max} - F_{0\min} \quad \text{[Hz]}. \tag{6.4}$$

Koeficient variability frekvence základního tónu

Tento parametr je podobný rozptylu, avšak počítá i s maximální a minimální hodnotou.

$$V_{F0} = (F_{0\max} - F_{0\min}) \frac{|F_{0mean} - F_{0median}|}{F_{0mean}} \quad \text{[Hz]}. \tag{6.5}$$

Rozptyl frekvence základního tónu

$$D_{F0} = \frac{\sum_{i=0}^{N-1} (F_0[i])^2}{N} - (F_{0mean})^2 \quad \text{[Hz]}. \tag{6.6}$$

Směrodatná odchylka frekvence základního tónu

$$\sigma_{F0} = \sqrt{D_{F0}} \quad \text{[-]}. \tag{6.7}$$

Pozice maxima frekvence základního tónu

$$F_{0\max\ pos} = 100\ \frac{find\ (F_{0\max})}{N} \quad [\%], \qquad (6.8)$$

kde *find* je funkce určující index vstupní hodnoty.

Pozice minima frekvence základního tónu

$$F_{0\min\ pos} = 100\ \frac{find\ (F_{0\min})}{N} \quad [\%]. \qquad (6.9)$$

6.1.2. Intenzita(Energie) řeči

Průměrná hodnota Energie

$$E_{mean} = \frac{1}{N}\sum_{i=0}^{N-1} E_0[i] \quad [-], \qquad (6.10)$$

kde $E_0[i]$ je energie vypočtená pro *i*tý rámec.

Maximální hodnota energie

$$E_{\max} = \max(\overline{E}) \quad [-], \qquad (6.11)$$

kde $\overline{E}$ je energie vypočítaná pro jednotlivé rámce.

Minimální hodnota energie

$$E_{\max} = \min(\overline{E}) \quad [-]. \qquad (6.12)$$

Rozdíl maximální a minimální hodnoty energie v dB

$$E_{dB} = 10\log(\frac{E_{\max}}{E_{\min}}) \quad [\mathrm{dB}]. \qquad (6.13)$$

Koeficient variability energie

Význam této veličiny je stejný jako význam stejné veličiny pro základní tón řeči

$$V_E = (E_{\max} - E_{\min}) \frac{| E_{mean} - E_{median} |}{E_{mean}} \quad [-]. \qquad (6.14)$$

Rozptyl energie

$$D_E = \frac{\sum_{i=0}^{N-1} (E[i])^2}{N} - (E_{mean})^2 \quad [-]. \qquad (6.15)$$

Směrodatná odchylka energie

$$\sigma_E = \sqrt{D_E} \quad [-]. \qquad (6.16)$$

Pozice maxima energie

$$E_{\max\ pos} = 100 \frac{find(E_{\max})}{N} \quad [\%]. \qquad (6.17)$$

Pozice minima energie

$$E_{\min\ pos} = 100 \frac{find(E_{\min})}{N} \quad [\%]. \qquad (6.18)$$

6.1.3. Střední hodnota počtu průchodů nulou

Střední hodnota počtu průchodů nulou

$$zcr_{mean} = \frac{1}{N}\sum_{i=0}^{N-1} zcr[i] \quad [-]. \qquad (6.19)$$

$zcr[i]$ je počtu průchodů nulou i-tého rámce.

Maximální hodnota počtu průchodů nulou

$$zcr_{\max} = \max(\overline{zcr}) \quad [-], \qquad (6.20)$$

kde $\overline{zcr}$ je vektor počtu průchodů nulou pro jednotlivé rámce.

Minimální hodnota počtu průchodů nulou

$$zcr_{\min} = \min(\overline{zcr}) \quad [-], \qquad (6.21)$$

kde $\overline{zcr}$ je vektor počtu průchodů nulou pro jednotlivé rámce.

Pozice maxima počtu průchodů nulou

$$zcr_{\max\ pos} = 100\ \frac{find\ (zcr_{\max})}{N} \quad [\%]. \qquad (6.22)$$

6.2. Navržený systém analýzy emočních stavů z řečového signálu

Celý systém je založen na analýze řeči při použití dostupných vzorců, které jsou zmíněny v předchozí kapitole. Pomocí matematického programu MATLAB (6.1.0.450) je vytvořen program, který se skládá ze dvou částí. A to samotná analýza řečového signálu a druhá část, která shromažduje trénovaní vzorky a po jejich načtení je zapíše do trénovaní matice k pozdějšímu určení klasifikátorů. K tomuto se vrátíme později.

Jak jsem již zmínil dříve k docílení co nejpřesnější analýzy řeč signálu je zapotřebí mýt co nejvíce obsáhlou databázi jednotlivých emočních stavů. Také je třeba při shromaždování těchto nahrávek dbát na to, že je zapotřebí mít od každého mluvčího nahrávky stejných emočních stavů. V mém případě jsem použil již dříve zkoumanou databázi vytvořenou studentem školy VUT v roce 2007 Ing.Hichamem Atassim, kterou jsem doplnil o některé vlastní nahrávky, které jsou ale lehce s příměsí šumu a tudíž některé jejich výsledky se odlišují od ostatních analyzovaných v jednotlivých emočních stavech.

Samotný systém tvoří tyto bloky:

1. Načtení vzorku řeči: Tento blok načítá zkoumaný vzorek řečového signálu. Při ohledu na funkci rozpoznávání v reálném čase by tento blok byl ošetřen trochu jiným způsobem, který by čekal na hodnoty z vnějšího zdroje a samotné načítání vzorku by zde nebylo nutné.

2. Nastavení (INI): Blok načte veškeré potřebné hodnoty, které budou dále v systému zapotřebí. Veškeré tyto hodnoty jsou uloženy v souboru INI.mat , který je pro celý systém nepostradatelný. Tento blok se mění v průběhu učení jednotlivých vah příznaků pro správnou analýzu řečových signálů.

3. Výpočet příznaků F_O: Tento blok analyzuje veškeré hodnoty základního tónu řeči a uloží hodnoty do rozpoznávací matice. Blok analyzuje tyto příznaky: Střední hodnota F_O, maximální hodnota F_O, minimální hodnota F_O, rozdíl Max.a Min. hodnoty F_O, variabilitu F_O , směrovou odchylku F_O, pozice maxima a minima F_O udávanou v procentech.

4. Výpočet příznaků Energie: Tento blok analyzuje veškeré hodnoty intenzity řečového signálu a taktéž je uloží do rozpoznávací matice. Blok analyzuje tyto příznaky: Střední hodnota energie, maximální a minimální hodnota energie, jejich rozdíl Max. a Min. hodnot, variabilitu energie, směrovou odchylku energie a v poslední řadě také pozice maxima a minima energie udanou v procentech vzhledem k celkovému času signálu.

5. Výpočet příznaků ZCR: Tento blok složí k výpočtu parametrů funkce počtu průchodů nulou. A to následující: průměrný počet průchodu nulou, maximální a minimální počet průchodu nulou

6. Výpočet příznaků MFCC: Tento blok slouží k analýze milovských kepstrálních koeficientů analyzovaného vstupního signálu. V našem případě analyzujeme prvních 2Okeptrálních koeficientů, které jsou zapsány do rozpoznávací matice. Tento blok také spočítá průměrné kepstrální koeficienty celého signálu.

7. Medián: Tento blok je realizován funkcí median.m , která vypočítává mediány jednotlivých příznaků F_O a energie,které jsou zapotřebí k výpočtu některých veličin.

8. Průměr vzorků: Tento blok je realizován funkcí prumervzorku.m , která u některých dříve analyzovaných příznaků spočítá jejich průměr a zpětně je zapíše do rozpoznávací matice.

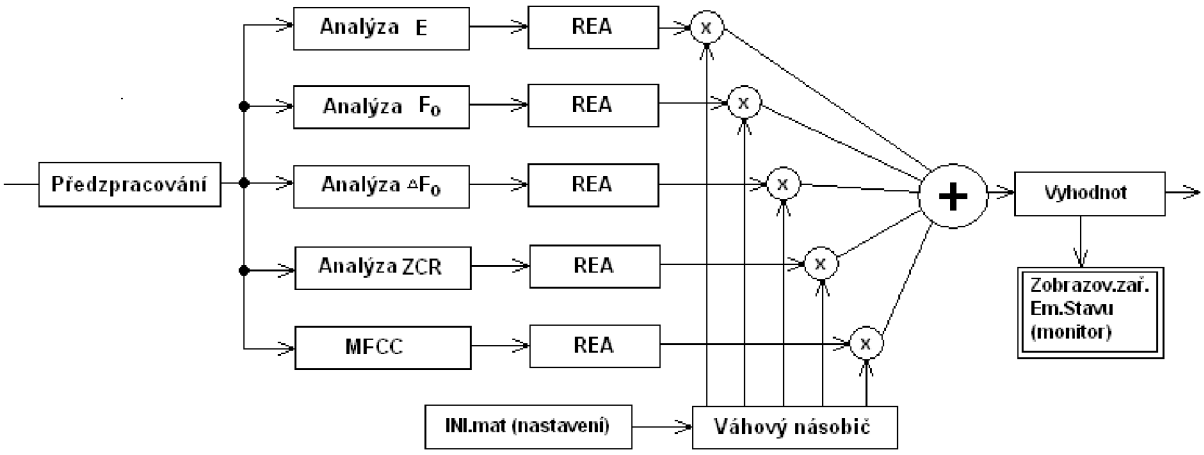
9. Porovnání: Tento blok je realizován funkcí porovnej.m , která slouží k samotnému porovnání analyzovaného vzorku řeči. V prvním kroku tato funkce načte databázi všech již vypočtené prozodické rysy jednotlivých emočních stavů,které jsou uloženy v databázi matic:
    a. MATICE_Neutral.mat - matice pro emoční stav: Normální
    b. MATICE_Vztek.mat - matice pro emoční stav: Vztek
    c. MATICE_Smutek.mat - matice pro emoční stav: Smutek
    d. MATICE_Radost.mat - matice pro emoční stav: Radost
    e. MATICE_Prekvapení.mat - matice pro emoční stav: Překvapení
    f. MATICE_Nuda.mat - matice pro emoční stav: Nuda

Jednotlivé příznaky z těchto matic jsou uloženy do jedné rozpoznávací matice společně s příznaky získaných analýzou daného vzorku, který chceme vyhodnotit. Funkce taktéž načte z INI.mat hodnoty vah jednotlivých příznaků. Tyto příznaky jsou samozřejmě každý jinak důležitý pro samotnou analýzu a o tyto váhy se stará blok k tomuto určený.

Pro vyhodnocování jednotlivých databázových matic nebyli použity veškeré vzorky, nýbrž jen ty nejlepší a také ty, které nezávislé osoby označily za správný emoční stav mluvčího. Ostatní vzorky byli jen zkušebně testovány jelikož by mohli zkreslovat konečný výsledek analýzy. Navíc díky absenci ženských nahrávek je systém konstruován pro analýzu zejména mužských emočních nahrávek. Toto se dá lehce eliminovat a to tím,že do sytému přidáme nahrávky ženských emočních stavů.

Počet sloupců jednotlivých matic záleží na počtu analyzovaných příznaků a to v našem případě 8 hodnot F_0, 8 hodnot Energie, 3 hodnoty počtu průchodů nulou a v poslední řadě mel-kepstrální koeficienty (20 hodnot)

Schéma vyhodnocovací funkce:



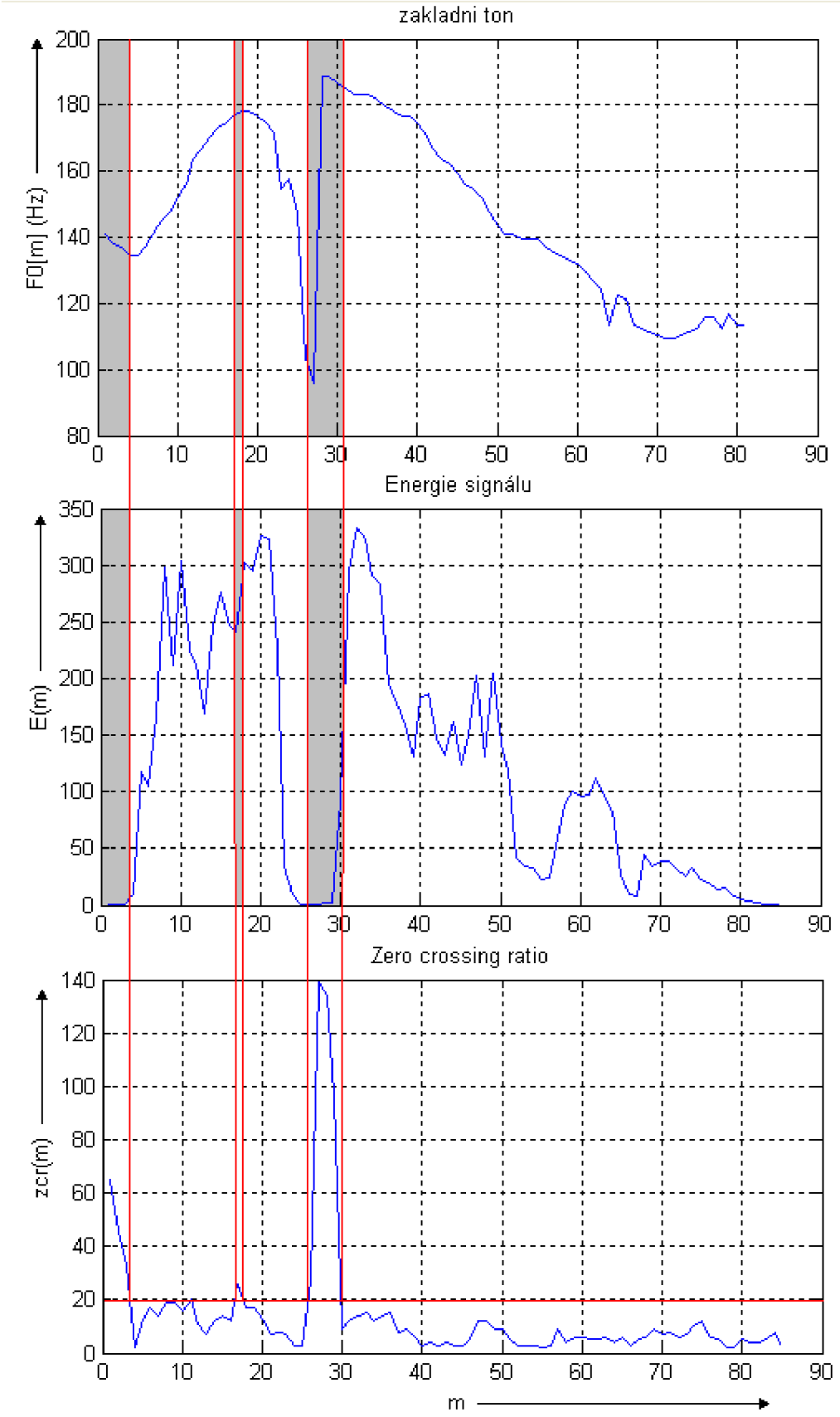
Obr.6.1.: Schéma vyhodnocovací funkce systému pro analýzu emočních stavů.

V závěru funkce je každému emočnímu stavu přidán dle rozdílů od různých emočních stavů a vypočtených rozhodovacích vah (systémem REA) přiřazeny hodnoty násobku těchto jednotlivých vah příznaků a jejich rozdílu od daného příznaku. Důvod těchto vah je prostý a to, že né všechny vypočtené příznaky mají stejnou váhu v rozhodování o emočním stavu. Funkce vrací 6 hodnot, které jsou vyjádřením subjektivního rozdílu od daných emočních stavů.

10. Vyhodnocení: Tento blok se zabývá a přímo navazuje na předchozí blok. Pomocí těchto 6-ti hodnot vypočtených v předchozím bloku přiřadí jednotlivému emočnímu stavu hodnotu shody s jednotlivými emočními stavy a to hodnotu v procentech.

11. Výstup na monitor: Tento blok je určen k výpisu vypočtených analýz. V prvním kroku vypíše jednotlivé shody s danými emočními stavy a posléze vyhodnocený emoční stav a to ten,který se svoji shodou nejvíce odpovídá jednotlivým emočním stavům.

Celý systém také odfiltrovává nepotřebné analyzované segmenty řeči,které by jen zkreslovali výsledné rozpoznávání. Je zbytečné analyzovat začátky a konce nahrávek a také jejich pauzy. Z tohoto důvodu byl do systému implementován systém pro rozpoznávání tichých mezer(pauz,apod.). Dle obrázku 6.2. si lze všimnout, že při rozhodovací hranici určitých parametrů lehce zjistit jaké segmenty nebudou pro analýzu nijak potřebné a kvůli lepším výsledkům jsem je z vyhodnocování vyřadil. Jednotlivé prahy těchto parametrů jsou uloženy v inicializačním souboru (INI.mat).



Obr. 6.2.: Systém pro filtraci nepotřebných segmentů analýzy.

6.3. Vyhodnocení

K samotnému testování mého systému pro analýzu emočních stavů byli použito 12 různých řečníků pro každý emoční stav. Dále byli testovány nějaké pokusné vzorky, které byly nahrány jen v některých emočních stavech, aby bylo možné ověřit jakou přesnost má systém pokud nemáme od mluvčího předem před-nahrané veškeré emoční stavy.

Od toho se odvíjely výsledky pro různé situace analýzy:

1. Vzorek od mluvčího, který je součástí rozpoznávací databáze
2. Vzorek od mluvčího, který není součástí databáze

Tab.3: Výsledky testů systému kdy mluvčí je v databázi rozpoznávacích matic.

	Neutralita	Vztek	Radost	Smutek	Překvapení	Nuda
Neutralita	**68**	3	6	4	4	4
Vztek	1	**74**	3	0	1	2
Radost	3	7	**71**	1	8	3
Smutek	9	5	2	**76**	2	23
Překvapení	4	6	13	3	**83**	1
Nuda	16	4	3	16	1	**67**
Průměr: 73 %						

Tab.4: Výsledky testů systému kdy mluvčí není v databázi rozpoznávacích matic.

	Neutralita	Vztek	Radost	Smutek	Překvapení	Nuda
Neutralita	**54**	5	7	4	121	13
Vztek	4	**71**	17	6	3	4
Radost	6	14	**61**	7	11	4
Smutek	16	2	7	**54**	4	21
Překvapení	6	5	3	3	**66**	7
Nuda	14	3	5	26	4	**51**
Průměr: 59 %						

Kapitola 7

7. Závěr

Veškerá diplomová práce se zaměřuje na analýzu a rozpoznávání emočních stavů českých nahrávek emočních stavů. Podařilo se mi také vyzkoušet systém na již profesionálně zpracované databázi emočních stavů (viz. Kapitola 5), ale tyto databáze byli v cizích jazycích a také s trochu odlišnými kategoriemi jednotlivých emočních stavů nežli jsem testoval u českého jazyka.

Při testování i vyhodnocování byli brány v potaz vypočtené suprasegmentální rysy daných emočních nahrávek neboli prozodické příznaky. V mém případě se jedná o frekvenci základního tónu řeči, intenzitě (energii), počty průchodu nulou a spektrálními vlastnostmi zde zkoumané a to mel-kepstrálními koeficienty. Výpočtu MFCC bylo zapotřebí využít milovských bank filtrů (viz. Kapitola 3). Pro naši analýzu jsem vypočítávali pouze prvních 2O milovských koeficientů, jelikož další již takřka neovlivní výsledky analýzy.

Systém jsem vybavil filtračním zařízením, které odfiltruje veškeré nepotřebné vzorky v signálu, které by zbytečně ovlivňovali výsledek analýzy. Tento systém je naznačen na obrázku 6.2. kde lze vidět červeně označené místa, které jsou označeny jako nevýrazné, či jako pauzy řečového signálu. Systém tímto zvýší o nepatrnou část účinnost analýzy, nicméně nemá až tak zásadní vliv na konečný výsledek.

Menší problém se vyskytuje ve výpočtu základního tónu řeči, kde u některých segmentů nelze přesně určit frekvenci základního tónu řeči a to vlivem mikroprozodie. Lze ji trochu eliminovat výběrem vhodné metody pro výpočet této frekvence základního tónu, nýbrž nelze úplně eliminovat. Ostatní prozodické rysy byly takřka bez jakýchkoliv problémů analyzovány.

Celý systém je postaven na použitém prostředí MATLAB (verze. 6.1.0.450) a díky tomuto vyspělému prostředí lze některé výpočty vyhodnotit ve velice krátkém čase. V mé práci se zabývám implementací tohoto systému do reálného času, nicméně i přes výkonnost

tohoto programu dle mého názoru nestačí rychlost procesoru, která je v dnešní době nabízena na našem trhu. V budoucnu avšak věřím, že bude počítačová technika natolik vyspělá, aby tyhle výpočty zvládla ve velice krátkém čase.

S ohledem na implementaci do reálného času jsem práci nijak zbytečně nezatěžoval různými grafickými prostředími či grafických výstupů. Grafické výstupy si ovšem můžete zapnout v nastavovacím souboru nastavení.m , kde při přiřazení veličině zobraz hodnotu 1 vám bude systém zobrazovat některé ze základních prozodických veličin. Některé grafické výstupy jsou znázorněny v této diplomové práci.

Celý systém byl navržen pro krátké segmenty řečového signálu. Řádově okolo několika sekund. Pokud by jsme systém chtěli implementovat do reálného času, bylo by zapotřebí k němu vyvinout další funkci, která by z plynulé nahrávky vysekávala jednotlivé věty či jednotlivé slova, které by systém pro analýzu emočních stavů zpracovával zvlášť. Při samotném testování systému jsem brali v potaz dvě základní situace a to když analyzovaný řečník je již součástí porovnávacích databází emočních stavů či v druhém případě se jedná o cizího řečníka, který v rozpoznávací databázi ještě není. V prvním případě, kdy je řečník v databázi se průměrná pravděpodobnost správného určení emočního stavu pohybuje okolo 73%. Tento výsledek je ovšem zkreslen malým počtem nahrávek a také u některých vzorků příměsí šumu a tedy nemůže být brán jako směrodatný pro přesné určení úspěšnosti. Při druhé situaci, kdy řečník není v databázi byla průměrná úspěšnost okolo 59% u vzorků s menší mírou šumu. U vzorků, které obsahovali větší hladinu šumu a příměsí okolních zvuků byla úspěšnost značně snížena a to pod hranici 50%.

Systém dosáhl lepších výsledků za použití některých se zmíněných databází v kapitole 5, ale tyto existující databáze jsou v cizích jazycích a tudíž jsem je zde neuváděl.

Na závěr pro další vývoj tohoto systému by bylo vhodné jej rozšířit o:

- Rozsáhlejší databázi českých emočních stavů (ve více emočních stavech)
- Zoptimalizovat příznaky pro rozpoznávání jednotlivých emočních stavů
- Použít některé z dalších příznaků: např. tempo či mikroprozodii apod.
- Pro zasazení do reálného času – vytvořit funkci na vyseknutí vět či slov ze signálu
- Použít výkonou neuronovou síť pro klasifikátory jednotlivých emočních stavů

Seznam použité literatury

[1] **PSUTKA, J.** *Komunikace s počítačem mluvenou řeči*. Academia, Praha 1995.

[2] **SMÉKAL, Z.** *Číslicové zpracování signálů.* Skripta, Brno VUT 2003.

[3] **PRCHAL, J., ŠIMÁK, B**. *Digitální zpracování signálů v telekomunikacích.* ČVUT 2001, IBSN 80-01-02149-1

[4] **ATASSI, H**. Porovnání analýzy emočních stavů v závislosti typu jazyka.DP,VUT Brno 2007

[5] **COWIE, R.** *Describing the Emotional States expressed in Speech.* ISCA Workshop, 2000

[6] **Internetové stránky,** emotional-research.net: *http://emotion-research.net/wiki/Databases*

[7] **SIGMUND, M.** *Analýza řečových signálů.* VUTIUM, Brno 2000.

[8] **ČERNOCKÝ, J.** *Předzpracování řeči, tvorba řeči, cepstrum.* ÚPGM FIT VUT Brno

[9] **BOŠTÍK, M.** *Analýza hlasu pro rozpoznávání stresu: kandidátská disertace.* VUT Brno, 2005

[10] **VLČKOVÁ-MEJVALDOVÁ, J.** *Prozodie, cesta i mříž porozmění.* Karolinum, Praha 2006.

[11] **VONDRA, M.** *Prozodie.* Text do předmětu MZPR, Brno VUT 2007.

Přílohy

1.CD ROM disk, který obsahuje:

- Diplomovou práci ve formátu DOC a PDF
- Vytvořené programy v prostředí MATLAB:
  - Hlavní program lze spustit souborem ¨analýza.m¨
- Databázi řečových signálů

2. Seznam souborů systému pro analýzu emočních stavů.

Příloha 2: Seznam souborů systému pro analýzu emočních stavů

soubor	popis
energie.m	Funkce pro výpočet průběhu energie a jeho parametry (střední hodnota, medián, atd.)
rea.m	Funkce pro klasifikaci vah koeficientů jednotlivých prozodických rysů
melcepst.m	Funkce pro výpočet melovských koeficientů
median.m	Funkce pro výpočet mediánu posloupnosti
nastaveni.n	Vytvoří základní parametry INI souboru (default)
shrp.m	Funkce pro výpočet frekvence základního tónu
porovnej.m	Funkce pro porovnání shody s jednotlivými emočními stavy
prumervzorku.m	Funkce pro výpočet průměru některých prozodických stavů
spektrum.m	Funkce pro výpočet spektra signálu
vyhodnot.n	Funkce vyhodnotí shodu s jednotlivými rysy procentuálně
print.m	Funkce pro výstup vypočtených shod emočních stavů na monitor
zcr.m	Funkce vypočtu průběhu počtu průchodů nulou signálu
Analyza.m	Hlavní program – Analýza Em.stavů
Databázové soubory	
INI.mat	Databáze všech potřebných parametrů systému
MATICE_Vztek.mat	Databáze prozodických příznaků vzteku pro český jazyk
MATICE_Nuda.mat	Databáze prozodických příznaků nudy pro český jazyk
MATICE_Radost.mat	Databáze prozodických příznaků radosti pro český jazyk
MATICE_Neutral.mat	Databáze prozodických příznaků neutrality pro český jazyk
MATICE_Smutek.mat	Databáze prozodických příznaků smutku pro český jazyk
MATICE_Prekvapeni.mat	Databáze prozodických příznaků překvapení pro český jazyk

Poznámka: Systém byl navrhnut a vyvíjen v programu MATLAB verze.6.1.0.450.