Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

DIPLOMOVÁ PRÁCE

PLZEŇ, 2012 **JAKUB KOPŘIVA**

P R O H L Á Š E N Í

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

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

V Plzni dne *18. 5. 2012* .....

vlastnoruční podpis

Poděkování

Děkuji vedoucímu diplomové práce Doc. Ing. Jindřichu Matouškovi, Ph.D. za hodnotné rady, připomínky, trpělivost a odborné vedení práce. Dále děkuji své rodině za celkovou podporu.

Anotace

Hlavním tématem je problém automatické detekce momentů, ve kterých došlo k uzavření hlasivek, v digitalizovaném řečovém signálu. V začátku práce je vyložen vědomostní základ pojmů a principů, na který je dle potřeby odkazováno. Následuje popis a porovnání vybraných algoritmů, které řeší úlohu nalezení základního hlasivkového tónu v řečovém signálu. V hlavní části této práce jsou popsány principy různých algoritmů určených pro detekci pitch marků v řečovém signálu a nechybí srovnání jejich vlastností a dosažených výsledků. Některé z těchto algoritmů byly implementovány.

Klíčová slova

Pitch mark, automatická detekce pitch marků, okamžik uzavření hlasivek, řeč, řečový signál, hlasivky, glotální signál, neuronová síť, backpropagation, základní hlasivkový tón, kontura F_0, autokorelace, klasifikace znělosti, spektrum, Fourierova transformace, syntéza řeči, PSOLA.

Abstract

Automatic glottal closure instant detection problem is the goal of the thesis. At the beginning of the thesis knowledge base is placed to support further research. The following part is focused on comparing selected approaches, that are designed for pitch tracking task. Various selected techniques determined for automatic pitch marking task are presented in the major part of the thesis. Some of them were implemented. Comparison of all the available algorithms is included.

Keywords

Pitch mark, automatic pitch marking, glottal closure instant, speech, speech signal, glottal cords, electroglottogram, neural network, backpropagation, fundamental frequency, F_0 contour, autocorrelation, voice classification, spectrum, Fourier transform, speech synthesis, PSOLA.

Obsah

Kapitola 1

Úvod

Lidé mezi sebou odedávna komunikují a tato komunikace se nejen jako celek, ale také jako její dílčí části, vyvíjí a mění. Lidská komunikace se rozpadá na hlavní dvě části, a sice zvukovou a obrazovou. Zpočátku se jednalo o posunky v době archaické společnosti. Řeč se vyvinula až později, ale dnes se jedná o nejrozšířenější formu přímé lidské komunikace.

S pokrokem ve vědě a technice, především na poli výpočetní techniky, se v moderní době spolu s novými možnostmi naskýtá mnoho nápadů, jak lidem ušetřit práci či usnadnit život. Pro předání informace výpočetní technice lidé již několik desetiletí používají polohovací zařízení, klávesnice a další ovládací prvky. Naopak výpočetní technika poskytuje lidem zpětnou vazbu prostřednictvím monitorů a displejů. Jsou zde i samozřejmě jiné multimediální periferie jako mikrofon, reproduktory a další. Zvýšení výpočetního výkonu osobních počítačů, ale také jejich rozšíření do společnosti, přináší možnosti pro vývoj nových aplikací a prostředků, které mohou lidem komunikaci usnadnit. Myšlenka je umožnění přímé komunikace člověka s počítačem způsobem, který je pro člověka nejpřirozenější – řečí.

Realizace této myšlenky by s sebou přinesla nejen usnadnění rozmanitých úkonů na osobních počítačích pro běžné uživatele, ale také dokonce umožnění provádět některé úkony uživatelům handicapovaným (nevidomým, němým, pohybově postiženým). Pro realizaci zmíněné myšlenky je zásadní (kromě případných dalších sounáležitostí) vytvoření dvou systémů. Systému pro rozpoznávání řeči a systému pro syntézu řeči. Počítač by potom byl schopen převádět text na mluvenou řeč (text-to-speech) a naopak mluvenou řeč na text (automatic speech recognition). Vývoj v této oblasti vědy probíhá již řadu let a podílí se na něm škála pracovišť po celém světě. Vzniklo tak několik různých přístupů pro obě dílčí úlohy. Je zřejmé, že každá z těchto úloh vykazuje značnou složitost a skládá se opět z určitých dílčích úkolů.

Motivace a cíle práce

Jednou ze stále určitým způsobem používaných metod v TTS[1] systémech je metoda zvaná TD-PSOLA[2], která byla v historii velmi úspěšná. Tato metoda syntézy pracuje v časové doméně a pro svou činnost potřebuje co nejpřesněji nalezené okamžiky uzavření hlasivek v řečovém signálu. Kvalita syntézy řeči systémů TTS na bázi PSOLA[2] přímo závisí na kvalitě detekce okamžiků uzavření hlasivek. Jedním z hlavních smyslů této práce je porovnání algoritmů pro automatickou detekci hlasivkových pulzů v řečovém signálu.

V současné době nejpřesněji pracující algoritmy detekující okamžiky uzavření hlasivek nepracují pouze s řečovým signálem, ale i se signálem EGG[3]. Řečník, který má za úkol v nahrávacím studiu namluvit určitý řečový korpus, musí často trávit dlouhé hodiny mluvením se snímačem připevněným na krku, což s sebou nese několik nevýhod. Zejména to, že je to nepříjemné, dále také to, že tento snímač může při mluvení do určité míry řečníkovi „překážet“ a třeba i slabě ovlivňovat hlas, ale také jsou zde problémy se správným připevněním na krk řečníka a s měnícími se vlastnostmi signálu tohoto snímače, pokud se jeho umístění změní posunutím. Snahou je tedy v ideálním případě získat algoritmus detekující okamžiky uzavření hlasivek pouze v řečovém signálu tak, aby jeho úspěšnost byla srovnatelná s algoritmy pracujícími i s EGG[3] signálem.

Existuje velké množství korpusů a studiových nahrávek, ke kterým nebyl pořízen při nahrávání také EGG[3] signál. Ve skutečnosti je jich naprostá většina. Získáním algoritmu pracujícího pouze nad řečovým signálem, který by se svou úspěšností detekce vyrovnal algoritmům používajícím i EGG signál, by bylo umožněno použít veškeré kvalitní nahrávky, které byly pořízeny bez EGG signálu v minulosti nebo ke kterým z nějakých důvodů nebude možné EGG signál pořídit.

Stručný popis kapitol práce

1. kapitola – *Úvod* – má za úkol informovat o situaci v tomto oboru a o cílech této práce.
2. kapitola – *Teoretický základ* – shrnuje a vysvětluje pojmy a teorii zasahující do různých směrů oboru této práce. O jednotlivé části této kapitoly se opírají kapitoly následující.
3. kapitola – *Porovnání algoritmů výpočtu základního hlasivkového tónu* – se stará o popis a porovnání různých přístupů realizujících úlohu „detekce základního hlasivkového tónu“.
4. kapitola – *Porovnání algoritmů detekce hlasivkových pulzů* – je klíčovou kapitolou této práce. Zahrnuje v sobě popis vybraných algoritmů určených pro úlohu „automatické detekce hlasivkových pulzů“ a zprostředkovává také jejich detailní porovnání.
5. kapitola – *Závěr* – shrnuje a zhodnocuje získané výsledky a obsahuje náměty na další práci v rámci tématu práce.

[1] Systém text-to-speech pro konverzi textu na mluvenou řeč.

[2] PSOLA je zkratkou metody “Pitch Synchronous Overlap Add”, TD upřesňuje typ na pracující v časové doméně. V podkapitole 2.5 bude podrobněji prezentován princip této metody.

[3] EGG je zkratka signálu ze snímače zvaného „elektroglotograf“, který je konstruován pro připevnění na krk řečníka a jeho úkolem je snímat činnost hlasivek v čase. Jeho funkce je podrobněji vysvětlena v podkapitole 2.2.1.

Kapitola 2

Teoretický základ

V této práci se používá množství teoretických znalostí, odborných názvů a termínů, které jsou spolu s důležitými souvislostmi objasněny v této kapitole. Metody a principy uvedené v příštích kapitolách budou na pojmy v této kapitole odkazovat. V první řadě se jedná o teorii týkající se přirozené produkce řeči a její digitalizaci. Následující část bude vypovídat o některých metodách zpracování řečového signálu. Podrobněji budou vysvětleny způsoby určování základního hlasivkového tónu a v neposlední řadě se zde bude vyskytovat dostatečné množství teorie o syntéze řeči. Na konci této kapitoly bude stručně vysvětlen také princip neuronových sítí, protože je jednou z uvedených metod detekce hlasivkových pulzů využíván.

2.1 Jazyk a řeč

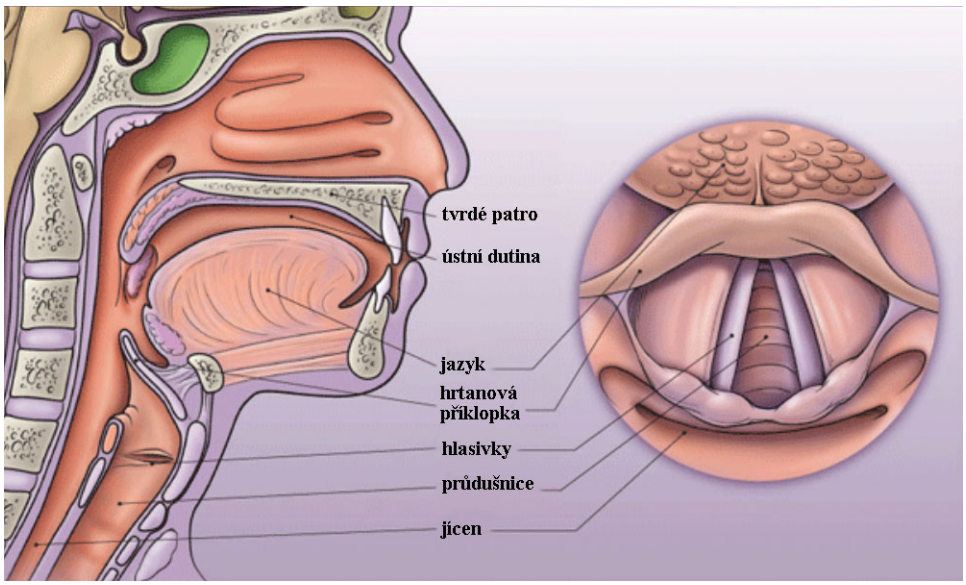
Dávní předkové dnešního člověka neměli schopnost mluvit. Až "Homo Sapiens"v dobách okolo 300 tisíc let před naším letopočtem používal k dorozumívání jednoduchou řeč. Vznikly první jazyky a postupně se vyvíjely až dodnes. Jazyk je pojem, který je úzce spjat s řečí. V současné době existuje velké množství jazyků (přibližně 6 tisíc) a většina z nich má mluvenou i psanou formu. Mluvená forma jazyka je řeč. Jedná se o přímou formu lidské komunikace. Má proto již z principu za úkol přenos nějaké informace od řečníka k posluchači.

2.1.1 Jazyk

Jazyk je obecně velmi složitý komunikační systém, který reprezentuje především lidskou schopnost vyjádřit myšlenku. Jak již bylo uvedeno, většina jazyků se vyskytuje ve dvou podobách a sice v mluvené (řeč) a v psané (písmo). Psaná forma používá zpravidla nějakou definovanou sadu grafických elementů – znaků. Mluvená forma se řídí složitou sadou pravidel pro správnou výslovnost a skládá se z fonetických elementů. Mluvená a psaná forma jazyka je považována za ekvivalentní s tím, že každá s sebou nese jisté výhody i nevýhody oproti té druhé [1].

2.1.2 Produkce řeči

Jak bylo již zmíněno, řeč je mluvená forma jazyka přenášející informaci. Primárním médiem pro přenos řeči je vzduch. Řeč je mechanické vlnění (zvuk) a je vytvářena v tzv. hlasovém traktu lidského těla. Hlasový trakt sestává z dechového, hlasového a artikulačního ústrojí. Dechové



Obrázek 2.1: Hlasový trakt [2]

ústrojí lze v tomto případě chápat jako zdroj energie a jeho součástmi jsou plíce, bránice, dýchací svaly, průdušky a průdušnice. Hlasové ústrojí je ta část, kde se objevuje poprvé zvuk. Hlavním prvkem jsou hlasivky umístěné v nejužší části hrtanu. Posléze se tento zvuk šíří artikulačním ústrojím až ke rtům, odkud je vyzařován dále do okolí. Artikulační ústrojí frekvenčně mění vlastnosti zvuku podle postavení hrtanu, jazyka a zubů v ústní dutině, měkkého patra, rtů a může zde být také přidána šumová složka potřebná pro vyslovení některých souhlásek. Proto může celý hlasový trakt vyprodukovat téměř jakýkoliv zvuk. Činnosti všech spolupracujících orgánů potřebných pro vznik řeči řídí mozek. Na obrázku 2.1 je hlavní část hlasového traktu [1].

2.1.3 Informace zakódovaná v řeči

Bližší informační pohled na řeč napovídá, že existuje několik „vrstev“ řeči, které se navzájem prolínají. Řeč se obecně skládá z promluv, promluvy zase z vět, věty ze slov, slova z morfémů a morfémy z fonémů, které jsou považovány za základní stavební kameny řeči. Každý foném odpovídá charakteristickému postavení všech orgánů hlasového traktu a má proto i specifické frekvenční vlastnosti. Každá řeč je posloupnost fonémů, ale ne každá posloupnost fonémů je řeč.

V řeči existuje akustická vrstva. V té se soustředí na řeč jako na výstup hlasového traktu z pohledu frekvencí a míru jejich zastoupení v tom kterém okamžiku řeči. Řeč může být reprezentována buď samotným časovým průběhem akustického tlaku nebo amplitudově-frekvenčním časovým spektrem (spektrogram) a podobně. Další vrstva je vrstva lingvistická a ta zkoumá řeč a její obsah z jazykového pohledu od fonémů (fonologická úroveň jazyka), jejich skládání do morfémů a slov (morfologická úroveň), dále slovosled slov (syntaktická úroveň), následuje vlastní informace jež byla vyřčena (sémantická úroveň) a konečně co tím bylo ve skutečnosti myšleno (pragmatická úroveň). Lingvistická vrstva respektuje také informaci o gramatice a větné skladbě. Poslední vrstva řeči zkoumá subjektivní informaci o řečníkovi, tedy intonaci, rytmus, barvu hlasu a jeho rozpoložení [3][4].

Představme si konkrétního řečníka jako unikátní osobnost, která svým hlasem pronese konkrétní větu. Potom pronesená věta nese informaci nejen o vysloveném textu, který sám o sobě obsahuje určitá slova skládající se z menších jednotek v konkrétním slovosledu, gramatiku a vlastně i použitý jazyk, ale také informaci o řečníkovi a jeho hlasovém traktu, jeho rozpoložení a několik dalších informací. Na jedné takovéto větě může být zkoumáno: „Kdo ji řekl?“, „Co řekl?“, „V jakém jazyce to řekl?“, „Jak intonoval?“, „V jakém byl rozpoložení?“, ale dokonce i „Co tím myslel?“ atd.

2.1.4 Digitalizace řeči

Pokud v tomto případě uvedeme, že pracujeme s řečí, není to tak úplně pravda. Ve skutečnosti pracujeme s její číslicovou reprezentací. Abychom získali číslicovou reprezentaci řeči, je k tomu potřeba mikrofon a například zvuková karta. Mikrofon zprostředkovává konverzi spojitého průběhu akustického tlaku na spojitý průběh napětí. Tento spojitý průběh napětí je ve zvukové kartě zesílen a digitalizován.

Digitalizace probíhá ve dvou krocích, vzorkování a kvantizace s kódováním. Při vzorkování je vzorkovačem každou vzorkovací periodu odečtena analogová hodnota spojitého průběhu napětí a k ní je přidělena jedna z konečného počtu hodnot. Pro různé použití jsou doporučeny různé standardy. V našem případě jsou používány hlasové nahrávky se vzorkovací frekvencí 16 kHz a s rozlišením 16 bitů. Datový tok je tedy 256 kbps (kilobitů za vteřinu) [1].

2.2 Hlasivky a jejich vlastnosti

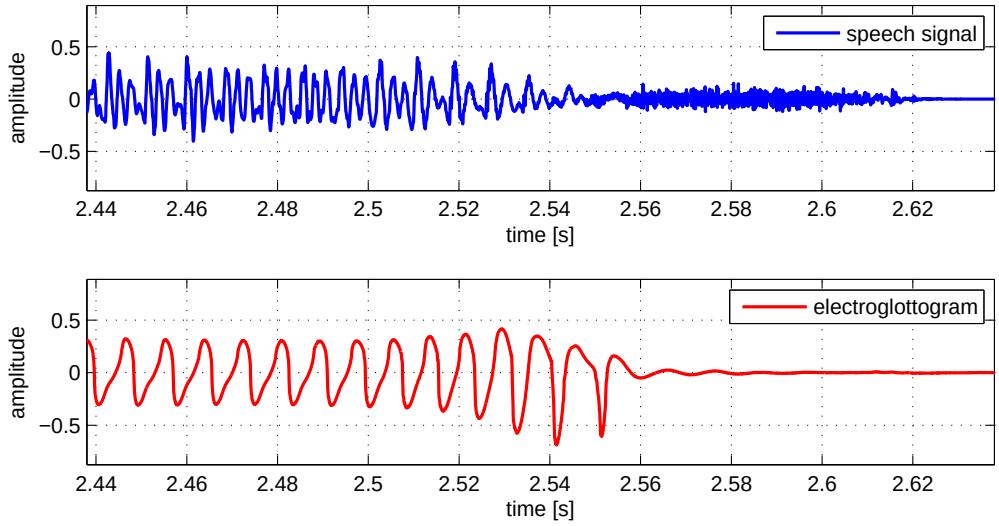
Hlasivky jsou párový sval viz. obrázek 2.1, který je při výdechu možno rozechvět proudem vzduchu díky jejich pružnosti. V místech vzájemného kontaktu jsou tzv. hlasivkové řasy (vazivová tkáň), které tuto pružnost zajišťují. U mužů se jejich délka pohybuje od 18 do 25 milimetrů, u žen je to potom od 14 do 20 milimetrů. Tento parametr spolu s hmotností hlasivek a s tlakem výdechu, ale i s vlastnostmi ostatních orgánů v hrtanu určuje rezonanční frekvenci, na které budou za normálních podmínek hlasivky kmitat. Tato frekvence se nazývá *základní hlasivkový tón* (“fundamental frequency”) a značí se F_0 a vypovídá o tom, kolikrát za vteřinu dojde k jevu nazývanému *okamžik uzavření hlasivek* za normálních podmínek [5].

2.2.1 Činnost hlasivek a elektroglotograf

Činnost hlasivek mimo jiné technologie snímá přístroj zvaný „elektroglotograf“. Jedná se o dvě elektrody připevněné na krk řečníka. Řečník díky tomu při promluvách poskytuje kromě řečového signálu nahrávaného mikrofonem ještě signál snímaný elektroglotografem. Tento signál nazývaný *EGG signál* lze vykreslit do grafu, kterému se pak říká „elektroglotogram“. Příklad takového signálu spolu s korespondujícím řečovým signálem zobrazuje obrázek 2.2. Lokální minima EGG signálu na obrázku 2.2 vypovídají o okamžicích uzavření hlasivek a naopak maxima odpovídají okamžikům úplného otevření hlasivek.

V čase do 2,56 s lze vidět znělý úsek řečového signálu a odpovídající příčinu v podobě chvějících se hlasivek na elektroglotogramu. Od času 2,56 s se hlasivky nechvějí, jedná se o neznělý úsek řeči tvořený čistě artikulačním ústrojím. Konkrétně jde o přechod ze samohlásky „o“ přes

souhlásku „j“ až po souhlásku „s“ představující neznělý úsek. V některých aplikacích se používá derivace EGG signálu, která se označuje jako DEGG.



Obrázek 2.2: Řečový signál a odpovídající elektroglotogram

2.2.2 Hlasivkové pulzy (pitch marky)

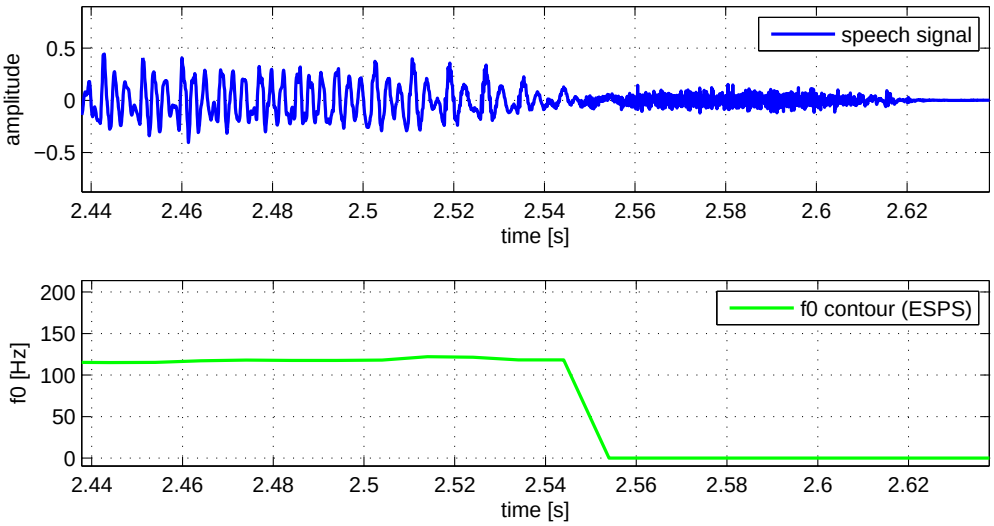
V pravé části obrázku 2.1 jsou vidět hlasivky v otevřené fázi. Pokud se důsledkem kmitání při řeči v určitém okamžiku navzájem dotýkají a je proto uzavřen výdechový proud vzduchu, pak je tento okamžik nazýván *okamžik uzavření hlasivek* (“glottal closure instant”) nebo také *hlasivkový puls* (“pitch mark”). Přesná znalost těchto okamžiků v řečovém signálu je důležitá pro některé metody syntézy řeči, především pro systémy na bázi PSOLA. Jsou zde tři hlavní přístupy, jak pitch marky v řečovém signálu detekovat.

Prvním je detekovat pitch marky ručně nejlépe za použití řečového signálu i elektroglotogramu, což je samozřejmě velmi zdlouhavé. Tento přístup je však považován za nejpřesnější, potažmo referenční. V této práci bude v následujících kapitolách sada takto získaných pitch marků použita jako referenční.

Druhým přístupem je vyhodnocovat řečový signál i elektroglotogram automaticky a využívat tak obě informace. Tímto přístupem je možné získat výsledky blížící se k ruční detekci, nevýhodou je však potřeba elektroglotogramu. Na pracovišti KKY[1] se tento přístup používá a dosahuje velmi dobrých výsledků.

Třetím a posledním přístupem je analyzovat pouze řečový signál a v něm detekovat pitch marky. Srovnání některých algoritmů pro automatickou detekci pitch marků v řečovém signálu, které různými způsoby řeší tento problém, je předmětem této práce.

[1]Katedra kybernetiky na Fakultě aplikovaných věd, Západočeská univerzita v Plzni



Obrázek 2.3: Řečový signál a odpovídající kontura F_0

2.2.3 Základní hlasivkový tón a jeho kontura

Základní hlasivkový tón charakterizuje řečový signál a velmi úzce souvisí s výškou hlasu. Jeho převrácená hodnota je tzv. základní hlasivková perioda a značí se T_0. Základní hlasivkový tón F_0 se během produkce řeči chová do určité míry dynamicky podle toho, jak řečník intonuje. Je vhodné mluvit o dynamickém rozsahu, který se u mužů pohybuje od 80 do 150 Hz a u žen od 150 do 300 Hz [5].

Základní hlasivkový tón může být pro konkrétní okamžik řečového signálu určen buď analýzou spektra daného okolí signálu nebo analýzou časového úseku řečového signálu. Jsou-li k dispozici pitch marky získané z řečového nebo z EGG signálu, může být jednoduše určena hodnota F_0 převrácením hodnoty časového rozdílu mezi dvěma sousedními pitch marky. Zpravidla však pitch marky k dispozici nejsou a kontura F_0 bývá naopak používána k jejich nalezení. Existují různé přístupy pro získání kontury F_0. Základním přístupem je autokorelační metoda, ale používá se také třeba metoda AMDF (viz kapitola 2.3.3), která pracuje na podobném principu jako korelace.

Kontura F_0 určuje vzájemnou vzdálenost sousedících pitch marků v daném momentu řečového signálu, a naopak vzdálenost sousedících pitch marků v řečovém signálu přímo určuje bod kontury F_0. Je zřejmé, že úloha *automatická detekce pitch marků v řečovém signálu* a úloha *určení kontury F_0 řečového signálu* mají hodně společného a v některých případech jedna využívá výsledky té druhé pro svůj účel a naopak. První zmíněná je však do určité míry obecnější, protože nalezením pitch marků dojde také k získání kontury F_0. Tento výrok naopak neplatí.

Obrázek 2.3 znázorňuje příklad kontury F_0 odpovídající části řečového signálu shodné s tou na obrázku 2.2. Opět je zřejmá souvislost mezi znělostí řečového signálu do času 2,56 s a konturou F_0, jejíž hodnota se pohybuje okolo 120 Hz. Naopak od času 2,56 s je řečový signál neznělý a kontura F_0 tuto skutečnost reflektuje nulovou hodnotou. Je zřejmá i souvislost mezi EGG signálem z obrázku 2.2 a konturou F_0 z obrázku 2.3.

2.3 Zpracování signálu v časové oblasti

Jedná se o operace v časové doméně nad signálem, který nemusí být řečový. Pro zpracování různých signálů v různých odvětvích se ustálilo několik základních a často používaných funkcí. Často je požadavek na tzv. krátkodobou analýzu, což znamená, že výsledná hodnota operace v konkrétním bodě je počítána z nějakého omezeného okolí signálu ke zpracování.

Aby se zamezilo vzniku zkreslení a docílilo požadovaného efektu, aplikuje se váhové okénko na omezené okolí použité k výpočtu. Typy okének jsou různá a technika jejich používání se nazývá okénkování ("windowing"). Většinu krátkodobých funkcí v časové oblasti lze obecně vyjádřit vztahem

$$Q_n = \sum_{k=-\infty}^{\infty} \tau(s(k))w(n-k), \tag{2.1}$$

kde Q_n je krátkodobá charakteristika, $s(k)$ je vzorek signálu v čase k, $\tau(...)$ vyjadřuje transformační funkci a $w(k)$ je okénko, s jehož pomocí se vybírají, případně váží vzorky $s(k)$ [1].

Základním okénkem je pravoúhlé okénko. Toto okénko je definované pro určitý interval jako konstantní funkce s hodnotou rovnou jedné, mimo interval je hodnota rovna nule. Vynásobení signálu po vzorcích takovýmto okénkem znamená výběr vzorků signálu a nemění jejich hodnotu. Tím ale vznikají vyšší harmonické složky kvůli ostrým okrajům, proto je zde snaha je potlačit. Toho docílíme přidělením snižující se váhy směrem k okraji okénka. Tuto vlastnost má například Hammingovo okénko, které se v časové oblasti zpracování signálu používá nejčastěji, a které je v intervalu $n = \langle 0, (L-1)\rangle$, kde L je počet vzorků vybraných okénkem, definováno vztahem

$$w(n) = 0,54 - 0,46cos(2\pi n/(L-1)). \tag{2.2}$$

Kdekoliv mimo interval přiděluje nulovou hodnotu. V některých aplikacích (především filtrace), kde je požadavek na rekonstrukci signálu po aplikované transformaci, se častěji používá Hanningovo okénko (ve vztahu 2.2 by se pouze místo 0,54 a 0,46 objevilo 0,5), které má na rozdíl od Hammingova nulové hodnoty na okrajích [1][6].

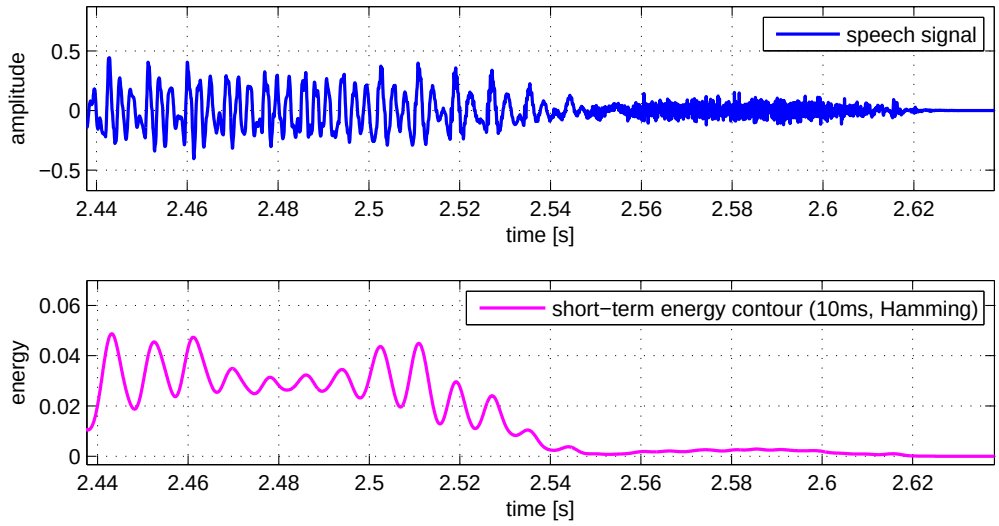
2.3.1 Krátkodobá energie

Funkce zvaná krátkodobá energie (short-term energy) je funkce popisující množství energie v průběhu analyzovaného signálu. Její průběh lze vyjádřit jako

$$E_n = \sum_{k=-\infty}^{\infty} \left[s(k)w(n-k)\right]^2, \tag{2.3}$$

kde důležitou roli hraje opět $w(k)$ jakožto okénko. Je vhodné volit mikrosegment stejně dlouhý jako vhodně zvolené okénko. Pro většinu aplikací se nabízí Hammingovo okénko. Nevýhodou funkce krátkodobé energie je citlivost způsobená kvadrátem. Proto se někdy používá funkce krátkodobá intenzita, kde rozdíl oproti krátkodobé energii spočívá pouze v použití absolutní hodnoty vzorků signálu $s(k)$ a odstranění kvadrátu. Obě tyto funkce poskytují informaci o průměrné hodnotě energie v daném mikrosegmentu signálu [1][6][7].

Pro shodný segment signálu, jako je zobrazen na obrázcích 2.2 a 2.3, byla vyhodnocena funkce krátkodobé energie, jež je znázorněna na obrázku 2.4.



Obrázek 2.4: Řečový signál a odpovídající kontura krátkodobé energie E, 10ms, Hamming

2.3.2 Krátkodobá autokorelační funkce

Obecně korelační funkce vyjadřuje míru podobnosti dvou různých segmentů signálů. Pokud by byl porovnáván segment signálu s odlišným segmentem toho samého signálu, potom se jedná o autokorelační funkci ("autocorrelation function, ACF"). Krátkodobá autokorelační funkce je definována jako

$$R_n(m) = \sum_{k=-\infty}^{\infty} s(k)w(n-k)s(k+m)w(n-k-m), \tag{2.4}$$

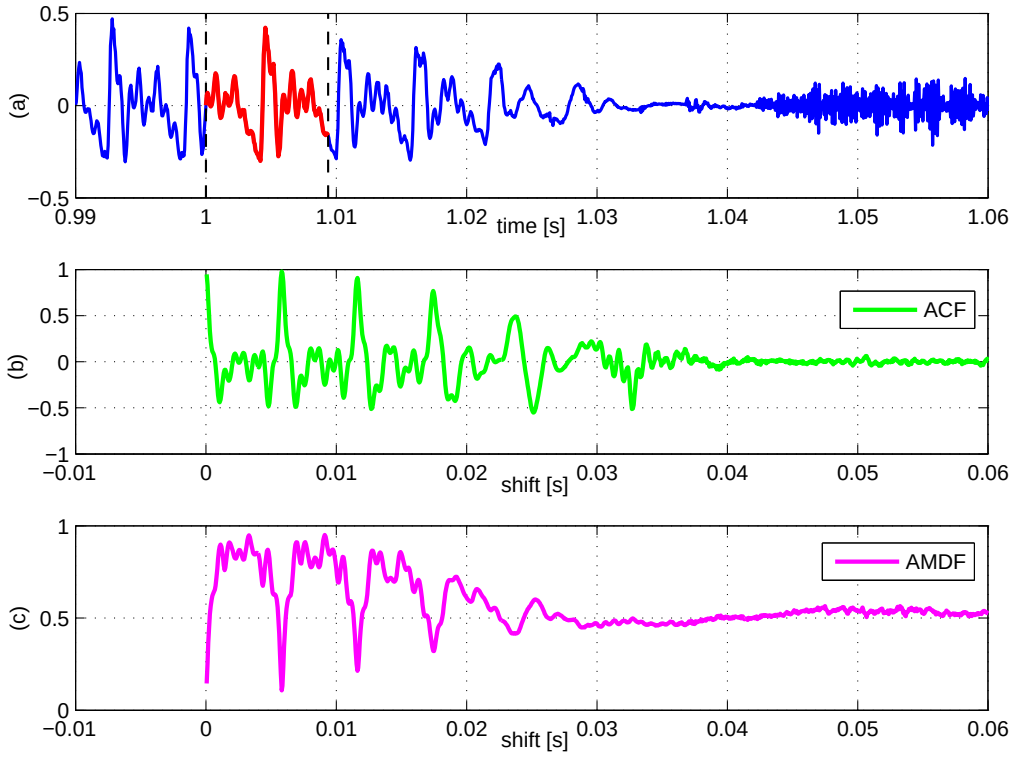
kde $w(k)$ je okénko. Autokorelační funkce vykazuje vlastnosti, které jsou často využívány pro zjištění periodicity signálu. Její lokální maxima se vyskytují v segmentech signálu, které jsou nejvíce korelované s referenčním segmentem. Toho se mimo jiné využívá pro získání kontury základního hlasivkového tónu F_0. Kontura F_0 na obrázku 2.3 je získána pomocí analýzy krátkodobé autokorelační funkce. Obrázek 2.5(b) zobrazuje příklad krátkodobé autokorelační funkce odpovídajícího segmentu signálu [1].

2.3.3 Krátkodobá průměrná rozdílová funkce

Krátkodobá průměrná rozdílová funkce ("average magnitude difference function, AMDF") je určitou obdobou krátkodobé autokorelace. Tato funkce je definována jako

$$D(\tau) = \frac{1}{N-\tau-1} \sum_{k=0}^{N-\tau-1} |s(k) - s(k+\tau)|, \tag{2.5}$$

kde τ vyjadřuje posuv. Je opět možné použít okénko. Tato funkce spočívá ve zvolení segmentu, zvyšování posuvu τ a v postupném vyčíslení průběhu funkce, která vykazuje lokální minima při



Obrázek 2.5: **(a)** řečový signál (modře), výběr segmentu řečového signálu obdélníkovým okénkem (červeně); **(b)** autokorelační funkce (ACF) (zeleně) ; **(c)** krátkodobá průměrná rozdílová funkce (AMDF) (purpurově)

hodnotě posuvu odpovídající periodicitě analyzovaného signálu. Autokorelační funkce vypovídá o periodicitě naopak ve svých maximech. Krátkodobá průměrná rozdílová funkce se proto také používá pro získání kontury F_0 základního hlasivkového tónu. Obrázek 2.5(c) zobrazuje příklad krátkodobé průměrné rozdílové funkce odpovídajícího segmentu signálu [8].

2.3.4 Další metody zpracování v časové oblasti

Často nastávají případy, kdy je vhodné analyzovat derivaci signálu. Ta se nahrazuje diferencí signálu. Pro vyšetření znělosti či neznělosti řečových úseků může být použita krátkodobá funkce středního počtu průchodů signálu nulou, která nabývá velkých hodnot v neznělých úsecích, malých hodnot v úsecích znělých a téměř nulových hodnot v úsecích ticha. Detekce lokálních maxim i minim (anglicky "peak picking") je také častou dílčí úlohou.

Metod zpracování je samozřejmě obrovské množství, od obecných jako jsou výše zmíněné, až po specializované s úzkou možností použití. Do sekce zpracování signálu v časové oblasti spadá v podstatě i celá teorie filtrování, která samozřejmě spadá i do sekce zpracování signálu ve frekvenční oblasti. Dále komprese, interpolace a převzorkování.

2.4 Zpracování signálu ve frekvenční oblasti

Tak, jako se v časové oblasti dobře analyzují některé vlastnosti (podobnost po sobě jdoucích period, vývoj amplitudy, znělost a neznělost řeči apod.) řečového signálu, nebo jakéhokoliv jiného signálu, tak ve frekvenční oblasti je časový signál transformován na spektrum (na nezávisle proměnné ose figuruje místo času frekvence), které poskytuje další možnosti analýzy, které nebyly možné nad signálem v oblasti časové. Je vhodné využívat výhod krátkodobé analýzy, přičemž vývoj spektra napříč celým signálem potom může být znázorněn poskládáním spekter paralelně do grafu, kde jsou amplitudy vyznačeny například pomocí stupňů šedi.

2.4.1 Krátkodobá diskrétní Fourierova transformace

Vypovídací hodnota amplitudového spektra celého signálu není pro většinu aplikací analýzy řeči zajímavá. Využívá se proto opět výhod krátkodobé analýzy, kde dochází k rozdělování analyzovaného signálu na okna, pro která jsou potom vyčíslena spektra. Vyčíslením spekter pro každý vzorek signálu se získají amplitudová spektra, jejichž poskládání vznikne spektrogram signálu.

K výpočtu diskrétní Fourierovy transformace (DFT) i její inverzní podoby se ve většině případů používá algoritmus rychlé Fourierovy transformace (FFT). Ten má však určitá úskalí. Hlavním z nich je požadavek, aby počet vzorků segmentu signálu určeného pro výpočet byl mocninou dvou. Pokud to není možné, používá se nejbližší vyšší mocnina dvou, ale vzniká tím chyba v podobě zkreslení spektra hlavně v místech vyšších frekvencí. Vztah vyjadřující funkci FFT je následující

$$\hat{S}_r(n) = \sum_{m=0}^{R-1} \hat{s}_n(m) e^{-\frac{j2\pi}{R}rm} = \sum_{m=0}^{R-1} \hat{s}_n(m) W_R^{-rm}, \tag{2.6}$$

kde $\hat{S}_r$ je R-bodová transformace v čase n posloupnosti $\hat{s}_n$ analyzovaného signálu. V případě nutnosti použití krátkého okénka nebo v případě požadavku na analýzu definované části spektra je vhodné použití techniky zvané zoom [1].

2.4.2 Krátkodobé kepstrum

Pro některé aplikace se používá tzv. kepstrum (z angl. cepstrum), které v podstatě vypovídá o frekvenčních vlastnostech spektra. Jde o inverzní Fourierovu transformaci logaritmu výkonového spektra neboli kvadrátu spektra, tedy formálně

$$C(\tau) = F^{-1}\left\{log\left|F\left\{\hat{s}_n\right\}\right|^2\right\}, \tag{2.7}$$

kde $\hat{s}_n$ je posloupnost vzorků analyzovaného signálu, F označuje Fourierovu transformaci a F^{-1} označuje inverzní Fourierovu transformaci. Díky vlastnostem výkonového spektra (reálná sudá funkce) může být inverzní Fourierova transformace ve vztahu 2.7 nahrazena přímou Fourierovou transformací [9][10].

Podobně jako u spektrogramu mohou být jednotlivá kepstra poskládána paralelně do grafu, kde například stupně šedi reprezentují amplitudu, čímž vzniká „kepstrogram“. Ten, jak ukazuje článek [10], může být také využit pro získání kontury F_0 základního hlasivkového tónu.

2.4.3 Další metody zpracování ve frekvenční oblasti

V první řadě jde o celou teorii filtrování, především pásmová filtrace. Často je alespoň přibližně známé frekvenční pásmo, které má být analyzováno. Ostatní frekvence jsou tak do určité míry pouze šumem a mohou být odfiltrovány. Dále se jedná o teorii okolo lineární prediktivní analýzy, homomorfního zpracování řeči a dalších přístupů.

2.5 Syntéza řeči

Syntéza řeči je proces umělého vytváření řeči. Zařízení provádějící syntézu řeči se nazývá syntetizér řeči. Je zde požadavek na co největší kvalitu, kde cílem je syntetická řeč k nerozeznání od přirozené řeči. S tím souvisí i ohodnocení kvality syntetické řeči, které se neobejde bez poslechových testů. Syntetizér řeči je systém, který za využití vstupní informace tvoří řečový signál. Tato vstupní informace zpravidla bývá kombinací fonetické a prozodické informace, kde fonetická reprezentuje posloupnost fonémů a prozodická udává informaci o melodii, časování a intenzitě. Systémy konverze textu na řeč (TTS) musí navíc obsahovat mechanismy pro převod holého textu do fonetického popisu a dále pro generování prozodické informace. Existují tři základní přístupy k syntéze řeči, na které se zaměří následující podkapitoly. Artikulační syntéza, formantová syntéza a korpusově orientovaná syntéza, která bude rozebrána podrobněji než ostatní [1]. Artikulační a formantová syntéza byly používány do 90. let minulého století. V současnosti naprostá většina technik syntézy řeči spadá do korpusově orientované.

2.5.1 Artikulační syntéza

Tento přístup vychází přímo z modelování hlasového traktu jako takového, včetně plic, artikulátorů atd. a jejich pohybů. Je nutné matematicky simulovat i výdechový proud plic a na jeho základě pak simulovat činnost modelů traktu. Jedná se tak o nejobecnější metodu syntézy řeči. Složitost tohoto přístupu je však překážkou, která nebyla zcela překonána a systém TTS nebyl zatím pomocí tohoto přístupu realizován [1].

2.5.2 Formantová syntéza

Tento přístup byl dlouhou dobu nejpoužívanějším v tomto oboru. Je založen na teorii zdroje a filtru. V lidském hlasovém traktu jsou zdrojem znělých segmentů řeči hlasivky a zvuk, který se z nich šíří, je potom jen „pasivně“ upraven artikulační částí hlasového traktu. O neznělé zvuky se starají některé artikulátory.

V teorii zdroje a filtru tak existuje zdroj pulzů (s určitou periodou) odpovídající hlasivkám pro znělé zvuky a zdroj náhodného šumu pro neznělé zvuky. Kombinací těchto dvou zdrojů v čase a aplikací vhodného filtru měnícího své parametry také v čase dochází k syntéze. Filtr bývá realizován sériovým nebo paralelním rezonátorem, kde rezonanční frekvence odpovídají formantům. Jaké vlastnosti rezonátorů pro kterou hlásku nastavit, o to se stará předem vytvořená databáze pravidel, na které je kvalita syntetické řeči silně závislá [1].

2.5.3 Korpusově orientovaná syntéza

V současné době je korpusově orientovaná syntéza nejpoužívanějším přístupem v oboru. Základní předpoklad je, že konečný počet řečových jednotek je schopný produkovat obecnou řeč. Řečovou jednotkou rozumíme například foném, který ale v obecné řeči bývá realizován poměrně širokou škálou variant, kde každá z těchto variant je nazývána realizací řečové jednotky. Proto je nutné před samotnou syntézou vytvořit inventář řečových jednotek nejlépe s dostačujícím počtem realizací každé jednotky. Předpokladem kvalitní syntézy tohoto typu je tedy mimo jiné kvalitní inventář řečových jednotek. Korpusově orientovaná syntéza se však dále dělí na konkatenační syntézu řeči (z angl. concatenative synthesis) a na statistickou parametrickou syntézu řeči (z angl. statistical parametric synthesis).

Konkatenační syntéza zahrnuje celou rodinu přístupů, jejichž společným znakem je konkatenace (řetězení) řečových jednotek. První systémy tohoto typu používaly malé korpusy. Byly založeny například na difónech a korpus byl tvořen třeba jen jedinou realizací každého difónu. Z toho plynula potřeba při řetězení modifikovat mikrosegmenty tak, aby byla prozódie výsledné syntetické promluvy přirozená. Tyto modifikace ve své době velmi dobře zajišťovala metoda PSOLA. Rozvoj v technice umožnil použití stále větších korpusů, začala se tak využívat technika "unit selection". Ta se stará o výběr optimálních jednotek z inventáře. I ve velkém množství realizací jedné konkrétní jednotky není vždy nalezena úplně optimální realizace, proto i v dnešních systémech tohoto typu dochází k jemným modifikacím pomocí podobných technik jako zajišťovala PSOLA.

Statistická parametrická syntéza v dnešní době dosahuje nejlepších výsledků. Je založena také na korpusech, které však při samotné syntéze již používány nejsou. Jádrem této metody jsou skryté Markovovy modely ("hidden Markov models, HMM"), které jsou trénovány pomocí korpusu. Při samotné syntéze pak dochází k řetězení těchto modelů, kterých je pro každou základní jednotku více. I v tomto případě proto dochází k "unit selection", jen se vybírá z modelů jednotek namísto realizací jednotek.

Metoda PSOLA

PSOLA je zkratkou anglického názvu "Pitch Synchronous Overlap Add"a původně vychází z metody OLA ("Overlap Add"), což znamená řetězení sečtením mikrosegmentů s překrytím. PSOLA oproti metodě OLA navíc řetězí mikrosegmenty synchronně se základní hlasivkovou periodou. Tedy jednotlivé mikrosegmenty získané rozdělením původního řečového signálu mají periodu základního hlasivkového tónu a jsou synchronizovány s okamžiky uzavření hlasivek (pitch marky) viz kapitola 2.2.2. Kvalita a konzistence nalezených pitch marků přímo ovlivňuje kvalitu výsledné syntetické řeči. Syntetizéry používající princip PSOLA patří mezi nejúspěšnější metody syntézy řeči 90. let [1].

Obecná funkce metody PSOLA předpokládá původní řečový signál a pitch marky reprezentované časy výskytu uzavření hlasivek v původním signálu. První bod syntézy pomocí PSOLA je *analýza* spočívající v dekompozici řečového signálu na mikrosegmenty podle pitch marků. To probíhá ve znělých úsecích řeči. V neznělých úsecích řeči probíhá dekompozice rovnoměrně. Pro dekompozici se používá Hanningovo okénko [1].

Druhým bodem je potom *modifikace*, při které dochází k potřebným prozodickým či frekvenčním změnám. Provádí se buď operacemi v časové doméně nebo aplikací krátkodobé Fourierovy transformace, jejíž výsledné spektrum může být upraveno podle potřeby [1].

Třetím a posledním bodem je samotná *syntéza*, tedy řetězení řečových mikrosegmentů. K tomu dochází zpravidla pomocí minimalizace nějakého chybového kritéria tak, aby výsledný proud syntetické řeči byl co nejlepší [1].

2.6 Neuronové sítě

Informační obsah této podkapitoly vychází z poznatků získaných na přednáškách předmětu Neuronové sítě [1]. Jedná se o obor zabývající se modelováním umělých neuronových sítí a vychází ze způsobu, jakým se člověk učí nějakou činnost s ohledem na procesy v jeho nervové soustavě. Stavebním kamenem lidského nervového systému je neuron. Jeho modelem je potom perceptron jehož funkce je popsána vztahem

$$y = f(\sum_{i=0}^{n}(w_i x_i + b)), \tag{2.8}$$

kde y je výstup neuronu, x_i je i-tý vstup neuronu a w_i je váha i-tého vstupu, kterou je tento vstup převážen. b je prahová hodnota perceptronu a f označuje obecně nějakou aktivační funkci. Pro diskrétní model perceptronu je to znaménková funkce, pro spojitý model jsou to bipolární či unipolární spojité funkce.

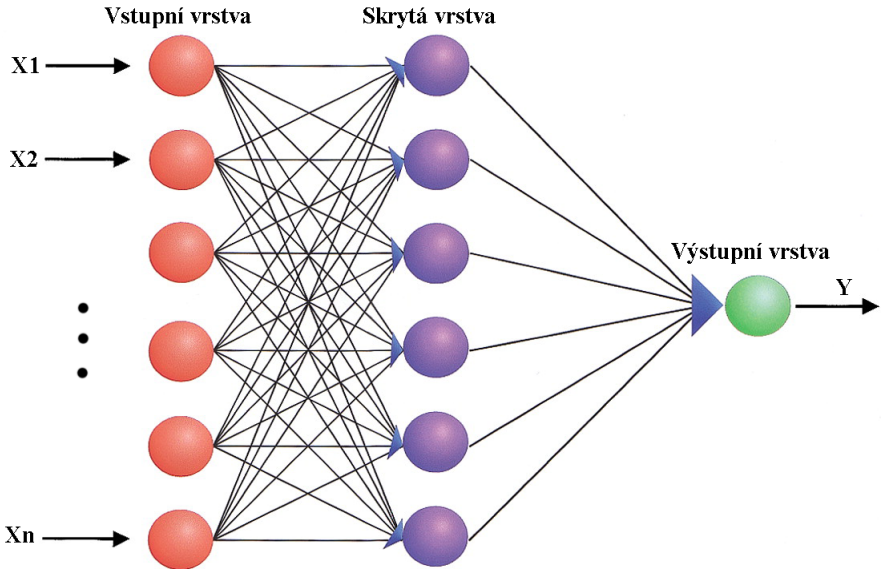
Neuronová síť vzniká spojováním perceptronů do hierarchií, které jsou obecně zpětnovazební nebo přímovazební. Perceptrony se zpravidla ukládají do tzv. vrstev (viz obrázek 2.6), ze kterého je patrné, že se jedná o přímovazební neuronovou síť. Důležité pro činnost sítě jsou aktivační funkce, váhy a prahy. Trénování sítí se provádí pomocí učení s učitelem nebo bez učitele. Obecně díky nějakému požadavku na konkrétní činnost navrhované sítě, dostupnosti a typu dat pro trénování a několika dalších aspektů je možné zvolit topologii sítě, aktivační funkce a způsob (algoritmus) trénování. Neuronové sítě se používají zejména pro úkoly klasifikace ("pattern recognition"), rekonstrukce dat, kódování a komprese dat, shlukování, predikce chování funkce a podobně.

Algoritmus Backpropagation

Jedná se o algoritmus určený pro trénování (učení s učitelem) nelineární vícevrstvé neuronové sítě. Předpokládá dostupnost trénovací množiny. Snahou je minimalizovat odchylku mezi odezvou postupně předkládaných trénovacích dvojic a příslušného požadovaného výstupu sítě. Činnost trénování tohoto typu lze zjednodušeně popsat následujícími čtyřmi kroky:

1. inicializace váhových matic a prahových vektorů
2. výpočet odezvy sítě
3. výpočet odchylky
4. aktualizace vah a prahů na základě zpětného šíření chyby.

Tyto čtyři kroky se opakují po dobu vkládání všech vektorů trénovací množiny, což je doba zvaná trénovací cyklus.



Obrázek 2.6: Příklad 2-vrstvé přímovazební neuronové sítě [12]

Po dokončení trénovacího cyklu se může zhodnotit chyba na testovací množině a je možné trénovat další cyklus. Toto je opakováno, dokud neklesne chyba pod určenou mez nebo dokud není proveden konkrétní počet trénovacích cyklů. Pak je síť považována za natrénovanou a začíná tím pracovní fáze, kdy je síť používána pro úlohu, pro kterou byla navrhnuta.

Kapitola 3

Porovnání algoritmů výpočtu základního hlasivkového tónu

Základní hlasivkový tón neboli F_0 byl popsán v podkapitole 2.2.3. Získání kontury základního hlasivkového tónu je v oblasti zpracování signálu známou úlohou, pro jejíž řešení bylo vyvinuto množství algoritmů využívajících různých přístupů. V této kapitole bude několik vybraných přístupů prezentováno a bude provedeno srovnání jejich přesnosti a vlastností.

Co nejpřesnější detekce základního hlasivkového tónu je potřebná pro množství úloh zabývajících se zpracováním řeči. Například některé algoritmy pro detekci pitch marků vyžadují informaci o kvazi-periodicitě analyzovaného řečového signálu. Dále je požadavek na získání základního hlasivkového tónu pro některé řečové korpusy, kódování signálu, telefonní technologie, ale třeba i pro analýzu a zpracování zvuku hudebních nástrojů (zde se jedná již jen o základní tón, ale princip je stejný).

3.1 Algoritmy pro detekci F_0

Bylo vybráno a získáno několik dostupných algoritmů řešících úlohu získání kontury základního hlasivkového tónu F_0, anglicky "Pitch tracking". V oblasti zpracování řečového signálu je několik nástrojů, které mají řešení této úlohy implementováno. Mezi nejznámější z nich patří programy WaveSurfer[1] a PRAAT[2].

Přístupy pro detekci základního hlasivkového tónu se dělí na detekci v časové oblasti, ve frekvenční oblasti a využívající obě domény. První zmíněné bývají zpravidla výpočetně jednodušší a používají se téměř výhradně tam, kde je požadavek na zpracování v reálném čase. Druhé zmíněné přístupy, tedy ve frekvenční oblasti, využívají analýzy spekter, lineární prediktivní analýzy, analýzy kepster či kepstrálních koeficientů a proto bývá výpočetní náročnost, ale i přesnost, vyšší.

Mezi typické možnosti nastavení algoritmů pro detekci F_0 patří zejména minimální a maximální frekvence, které se mají detekovat. Dále pak například šířka okénka pro analýzu.

[1] WaveSurfer je obecně známý nástroj pro analýzu a úpravy zvuku. Je možné jej stáhnout na stránkách <http://www.speech.kth.se/wavesurfer/>

[2] PRAAT je obecně známý nástroj pro analýzu, modifikace zvuku a množství dalších operací. Je možné jej stáhnout na <http://www.fon.hum.uva.nl/praat/>

3.1.1 WaveSurfer – metoda RAPT

Nástroj WaveSurfer byl vyvinut v institutu Centre for Speech Technology (CTT) ve Stockholmu a je volně dostupný v síti Internet. Tento univerzální program v sobě zahrnuje množství prostředků pro analýzu, úpravy a popis vlastností zvuku. Jedním z těchto elementů je nástroj "Pitch contour", který obsahuje dvě metody detekce kontury F_0. Prvním je metoda RAPT popsaná v článku [14], o které je tato podkapitola. Druhá je pak metoda AMDF, o které bude pojednáno v následující podkapitole 3.1.2.

Metoda RAPT[3] pracuje v časové doméně a je založena na principu krátkodobé autokorelační funkce, jež byla popsána v podkapitole 2.3.2. Výsledkem tohoto algoritmu je vektor hodnot základního hlasivkového tónu v ekvidistantních časových okamžicích. Nulové hodnoty tohoto vektoru vypovídají o neznělosti řečového signálu v tomto mikrosegmentu. Metoda RAPT díky svým vlastnostem navíc klasifikuje zvuk na znělé a neznělé úseky.

3.1.2 WaveSurfer – metoda AMDF

Metoda AMDF je druhou metodou, která může být v nástroji WaveSurfer zvolena pro získání kontury F_0. Tato metoda je založena na principu krátkodobé průměrné rozdílové funkce, jež byla popsána v podkapitole 2.3.3 a jedná se tedy o metodu pracující v časové doméně. Výsledkem tohoto algoritmu je také vektor hodnot základního hlasivkového tónu v ekvidistantních časových okamžicích a nulové hodnoty tohoto vektoru vypovídají o neznělosti řečového signálu v tomto mikrosegmentu stejně jako u metody RAPT. Metoda AMDF také navíc klasifikuje zvuk na znělé a neznělé úseky.

3.1.3 PRAAT – autokorelační metoda

Nástroj PRAAT byl vytvořen na pracovišti Phonetic Sciences na Amsterdamské univerzitě a je volně dostupný v síti Internet. Tento univerzální program v sobě stejně jako WaveSurfer zahrnuje velké množství elementů pro analýzu, úpravy a popis vlastností zvuku. V možnostech analýzy zvuku se nachází sekce nástrojů zvaná "Periodicity", která obsahuje několik metod detekce kontury F_0. Jednou z těchto metod je autokorelační metoda, která je obdobou autokorelační metody viz podkapitola 3.1.1. V tomto případě byla vybrána záměrně právě autokorelační metoda, aby došlo k porovnání přesnosti výsledků stejných přístupů dvou různých nástrojů (PRAAT a WaveSurfer).

Vlastnosti autokorelační metody nástroje PRAAT jsou obdobné jako u metody RAPT nástroje WaveSurfer, tedy analýza probíhá v časové doméně a princip byl popsán v podkapitole 2.3.2. Výsledkem je také vektor hodnot základního hlasivkového tónu v ekvidistantních časových okamžicích. Stejný je i princip obsahu informace o (ne)znělosti segmentů.

3.1.4 T. Ewender – Spojitá kontura F_0

Tato metoda byla prezentována v článku [10] na konferenci Interspeech Thomasem Ewenderem v roce 2009. Tento algoritmus byl vytvořen v Matlabu jako funkce *detect_F0_contour.m.*

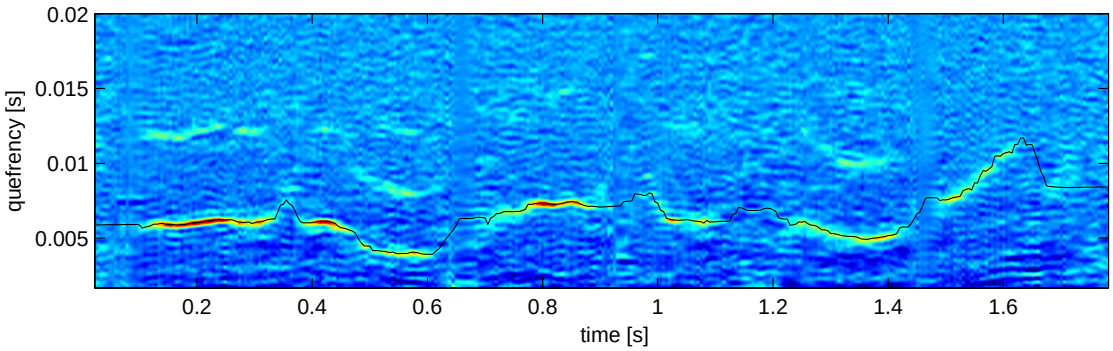
[3]Metoda RAPT pochází z toolboxu ESPS ("Entropic Signal Processing System") a jedná se o systém nástrojů pro analýzu a zpracování zvuku, který je využíván v tomto případě pro získání kontury F_0 v nástroji WaveSurfer. V nástroji WaveSurfer je metoda RAPT označena jako ESPS

Vychází z myšlenky, že člověk je schopen vidět konturu základní hlasivkové periody T_0 v kepstrogramu řečového signálu. Proto se jedná o metodu pracující ve frekvenční doméně. Pro její účel se získává kepstrogram s vysokým rozlišením, který vychází z výpočtu kepster segmentů řečového signálu.

Jakmile je kepstrogram vypočítán a provedeny související operace (eliminace vyšších harmonických složek a interpolace pro zvýšení rozlišení), následuje procedura vycházející z Viterbiova algoritmu. Zjednodušeně se tato procedura zabývá hledáním optimální cesty v kepstrogramu, ve kterém jednotlivé body mající příslušné hodnoty reprezentují stavy.

Na základě přirozených kontur F_0 celkem 17 hodin nahrávek bylo empiricky získáno rozložení sklonu a zakřivení, které bylo proloženo dvourozměrným Gaussovým rozložením. Tento krok v podstatě integruje informaci o fyzikálních vlastnostech řečového traktu a každý bod je tak ohodnocen určitou cenou, na základě které je optimální cesta nalezena. Tato cesta reprezentuje spojitou konturu T_0 a příklad takovéto optimální nalezené cesty v kepstrogramu je znázorněn na obrázku 3.1. Spojitá kontura F_0 se z kontury T_0 získá jednoduchým přepočtem.

U metod RAPT a AMDF je však na rozdíl od této metody „vedlejším produktem“ informace o znělosti. Informace o znělosti či neznělosti segmentů řeči byla autory pro tuto metodu vyřešena doplněním o neuronovou síť klasifikující segmenty řeči do pěti tříd (znělý, neznělý, smíšený, nepravidelný a ticho). Klasifikátor řečového signálu popsaný v článku [10] nebyl v této práci k dispozici. Autoři neuvádějí přesný typ použité neuronové sítě, uvádějí jen, že klasifikátor lze realizovat přímovazební sítí (viz obrázek 2.6), která je i s příslušným trénovacím algoritmem popsána v podkapitole 2.6.



Obrázek 3.1: Příklad nalezené optimální cesty (kontury T_0) v kepstrogramu

3.1.5 GLOAT – SRH metoda

Tato metoda byla prezentována v disertační práci [13] Thomasem Drugmanem. Zkratka GLOAT[4] pochází z “GLOttal Analysis Toolbox”a jedná se o sadu metod pro analýzu činnosti řečového ústrojí z řečových nahrávek. Autorem je Thomas Drugman a zmíněný toolbox byl vytvořen v Matlabu a klíčová funkce pro výpočet kontury F_0 je *SRH_PitchTracking.m*.

Tento přístup je založen na zpracování zbytkového signálu získaného ze signálu řečového. Řečový signál je rozdělen na segmenty. Jsou vypočítána spektra těchto segmentů. Pomocí regrese

[4]GLOAT je sada metod pro analýzu činnosti řečového ústrojí z řečových nahrávek. Je možné jej stáhnout na <http://tcts.fpms.ac.be/~drugman/Toolbox/>

je získána obálka spektra, na kterou je následně aplikována inverzní filtrace a je tak získán signál, ve kterém je eliminován šum a frekvenční složky s vyšší frekvencí, než je základní hlasivkový tón. Jde proto o metodu pracující ve frekvenční doméně.

Dále je získán spektrogram z tohoto zbytkového signálu, který se skládá ze spekter segmentů. Tato spektra mají v neznělých úsecích řeči relativně plochou spektrální obálku a ve znělých úsecích řeči obsahují vrcholek s hodnotou hledaného základního hlasivkového tónu. Ve spektrogramu je pak zřetelná kontura základního hlasivkového tónu F_0. Sečtením vzorků těchto spekter je získána jediná hodnota reprezentující spektrum jednoho segmentu signálu a vzniká tak spojitá funkce "Summation of Residual Harmonics (SRH)". Předpokladem je, že vzorek s vyšší hodnotou pochází ze spektra znělého segmentu signálu obsahující vrcholek. Informace o znělosti či neznělosti segmentů řeči je získána pomocí techniky prahování z SRH signálu. Spojitá kontura F0 je získána nalezením maxima pro každé spektrum. Tato kontura však neodpovídá té z předchozí podkapitoly, jelikož v neznělých intervalech se jedná prakticky pouze o šum.

3.1.6 Referenční kontura F_0

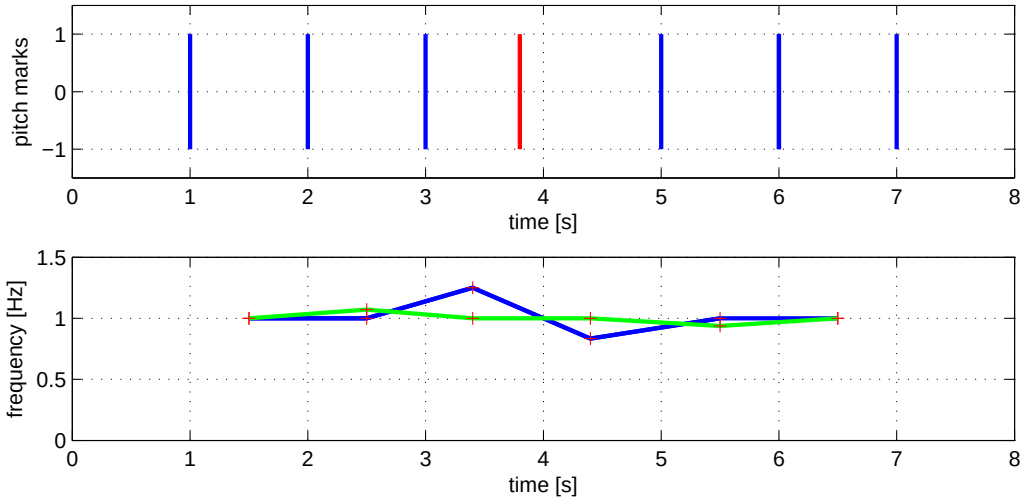
Zhodnocení výsledků některé z výše prezentovaných metod výpočtu kontury F_0 vyžaduje nějaká referenční data považovaná za správná. V této fázi se přímo nabízelo použít řečové nahrávky a k nim dostupné ručně určené pitch marky, protože vzdálenost dvou sousedních pitch marků odpovídá základní hlasivkové periodě T_0 a její převrácená je pak základní hlasivkový tón.

Data pro tento účel byla poskytnuta pracovištěm KKY ve složení celkem 83 (viz Korpus83 v příloze A) několikavteřinových nahrávek ve formátu jednokanálových 16-bitových souborů typu ".wav"a jim odpovídajících 83 souborů obsahující časy ručně určených pitch marků. Celkem se jedná o více než 11 minut záznamu obsahujících více než 50 tisíc ručně určených pitch marků. Nahrávky obsahují promluvy v češtině, slovenštině, angličtině, němčině a francouzštině od řečníků obou pohlaví. Vyskytují se zde také nahrávky s veselým intonačním zabarvením a dále 3 nahrávky, které obsahují problematické nepravidelné úseky.

Zbývalo zvolit způsob, kterým se z pitch marků získá kontura základního hlasivkového tónu F_0. Pouhým převedením vzdálenosti mezi sousedními pitch marky vznikala kontura, která byla místy značně lokálně zvlněná. To je dáno určitou nepravidelností period mezi jednotlivými pitch marky, zvláště pak v případech, kdy se střídala kratší perioda s periodou delší. Nabízí se možnost lokálního průměrování, které v této práci bylo použito.

Dostatečně hladká kontura byla docílena průměrováním přes 3 lokální periody, jehož výsledek je zobrazen na obrázku 3.2. Zde jsou v horním grafu znázorněny pitch marky ekvidistantně rozmístěné (modře), jeden z nich je o 20% lokální periody posunut (červeně). To má v dolním grafu za následek modře znázorněný průběh F_0. Při použití průměrování přes 3 periody je zeleně zobrazený výsledný průběh F_0 dostatečně hladký, taková kontura bude nadále v této práci chápána jako referenční.

V prostředí Matlab byla vytvořena funkce *Reference.m*, která při zavolání načte soubor s ručně určenými pitch marky, provede uvedené průměrování přes 3 lokální periody a jeho výstupem je pak vektor obsahující konturu F_0 a odpovídající časový vektor. V posledním výstupním vektoru jsou uloženy pitch marky načtené ze souboru.



Obrázek 3.2: Demonstrace výhodnosti průměrování přes 3 periody

3.2 Použité způsoby porovnání kontur F_0

V podkapitole 3.1 bylo popsáno celkem pět zvolených přístupů pro získání kontury F_0 a také způsob vytváření referenční kontury F_0. Otázkou v této chvíli bylo, jakým způsobem porovnat konturu vždy jedné z metod uvedených v podkapitolách 3.1.1 až 3.1.5 s konturou referenční. Po pečlivém zvážení všech požadavků na výsledky porovnání byly vybrány 2 hlavní způsoby porovnání. Prvním je porovnání 2 průběhů ve smyslu RMSE ("Root Mean Square Error") a druhým je porovnání pomocí vzájemné korelace.

3.2.1 Porovnání dvou průběhů F_0 ve smyslu RMSE

RMSE je často používaný a uznávaný statistický způsob porovnání odpovídajících prvků dvou množin. V tomto případě jsou množiny reprezentovány průběhy základních hlasivkových tónů, tedy posloupnostmi hodnot. Aby mělo porovnání smysl, musí dojít k porovnání vždy hodnoty jednoho průběhu F_0 s průběhem druhého, referenčního průběhu ve stejném čase. Hodnota testovaného a hodnota referenčního průběhu musí tedy příslušet stejnému časovému okamžiku řečového signálu. Obecně tomu tak není, proto byla použita lineární interpolace pro získání hodnot testovaného průběhu v časech, ve kterých je definován referenční průběh F_0, což je vždy v polovině periody ohraničené dvěma sousedícími pitch marky(znázorněno červenými body na obrázku 3.2). Kritérium RMSE je definováno vztahem

$$RMSE(R,T) = \sqrt{E((R-T)^2)} = \sqrt{\frac{\sum_{i=1}^{n}(r_i - t_i)^2}{n}}, \tag{3.1}$$

kde E zastupuje střední hodnotu a dále R značí vektor referenčních hodnot r_i a T vektor testovaných hodnot t_i, v obou případech délky n. V této fázi by již mohlo dojít k samotnému porovnání ve smyslu RMSE. Ještě před tím však bylo učiněno rozhodnutí ignorovat 3 hodnoty na začátcích a koncích znělých intervalů řečového signálu. To je proto, že referenční pitch marky jsou ručně

klasifikovány s použitím nejen řečového signálu, ale i EGG signálu viz podkapitola 2.2.1. Díky tomu jsou referenční pitch marky klasifikovány na začátcích a koncích znělých intervalů řečového signálu i tam, kde v řečovém signálu je téměř nepatrná energie a běžný algoritmus pracující jen na základě řečového signálu by v těchto místech pitch marky ani konturu F_0 nedetekoval. Nyní již mohlo dojít k porovnávání ve smyslu RMSE, jehož výsledky jsou obsaženy v podkapitole 3.3.

3.2.2 Porovnání dvou průběhů F_0 pomocí vzájemné korelace

Korelace je další kritérium často používané pro vzájemné porovnání dvou signálů. Toto kritérium bylo použito obdobně jako RMSE v předchozí podkapitole, tedy došlo k porovnání pouze v bodech, ve kterých je definován referenční průběh základního hlasivkového tónu F_0. Lineární interpolace popsaná taktéž v předchozí kapitole byla již provedena pro výpočet RMSE a mohla tak být použita i v případě korelace. Stejně tak byly ignorovány vždy 3 pitch marky na začátcích a koncích znělých intervalů řečového signálu a obě kritéria byla tedy vyčíslena pro stejnou sadu hodnot. Použit byl Pearsonův typ vzájemné korelace definovaný vztahem

$$COR(R,T) = \frac{E(RT) - E(R)E(T)}{\sqrt{E(R^2) - E^2(R)}\sqrt{E(T^2) - E^2(T)}}, \tag{3.2}$$

kde E zastupuje střední hodnotu a dále R značí vektor referenčních hodnot r_i a T vektor testovaných hodnot t_i, v obou případech délky n. Byla použita Matlabovská funkce *corr* a výsledné hodnoty jsou prezentovány v podkapitole 3.3.

3.3 Výsledky porovnání

Pro možnost porovnání byla v prostředí Matlab vytvořena funkce *Comparison.m*, která v prvním kroku načítá soubory obsahující vektory průběhů základních hlasivkových tónů F_0 metod Wavewurfer(RAPT), WaveSurfer(AMDF) a PRAAT(autocorrelation), posléze se volá funkce *Reference.m* zmíněná v podkapitole 3.1.6, která vrací vektor s referenční konturou F_0. Dále se volá Ewenderova funkce *detect_F0_contour.m* a Drugmanova funkce *SRH_PitchTracking.m* pro výpočet průběhů a výstupy těchto dvou funkcí jsou taktéž vektory obsahující kontury F_0.

Ve druhém kroku dochází k lineární interpolaci těchto celkem 5 vektorů tak, aby byla získána co nejpřesnější hodnota v bodech, ve kterých je definována referenční kontura.

Třetím krokem je již výpočet parametrů RMSE a COR prezentovaných v podkapitole 3.2 a tyto parametry jsou zároveň i výstupními proměnnými této funkce.

Pro porovnání bylo vybráno 20 (viz Korpus20 v příloze A) nahrávek z dostupných dat tak, aby byly obsaženi všichni řečníci i všechny jazyky. V prostředí Matlab byl vytvořen také jednoduchý skript *Comparison_all.m*, který postupně volá výše zmíněnou funkci *Comparison.m* vždy s jednou z vybraných nahrávek, a sjednocuje výsledky všech těchto 20 vybraných nahrávek. Dosažené výsledky jsou zaznamenány v tabulce 3.1.

Na první pohled vypadají výsledky nejlépe pro Ewenderovu metodu získání spojité kontury F_0 nalezením optimální cesty v kepstrogramu, avšak při zamyšlení vzniká otázka, zda není svojí spojitostí v čase celé nahrávky tato metoda zvýhodněna. A skutečně tomu tak je, protože všechny ostatní metody na svých výsledných konturách mají aplikované rozhodnutí o (ne)znělosti. Existují proto takové segmenty řeči, kde všechny ostatní metody chybně klasifikují neznělý úsek na rozdíl od Ewenderovy metody a tím se hodnoty obou kritérií těchto metod zhoršují. Nastal proto

Kritérium	**RAPT**	**AMDF**	**EWENDER**	**PRAAT**	**GLOAT**
RMSE [Hz]	20,68	41,93	8,08	22,33	24,32
COR	0,771	0,615	0,854	0,733	0,699

Tabulka 3.1: Výsledky kritérií RMSE a COR pěti vybraných metod

požadavek na aplikaci rozhodnutí o (ne)znělosti na konturu F_0 Ewenderovy metody, aby nebyla tato metoda zvýhodněna. Klasifikátor realizovaný neuronovou sítí nebyl nebyl k dispozici, proto musela být aplikována informace o (ne)znělosti z jiné metody. Vzhledem k tomu, že hodnoty kritérií RMSE a COR určitým způsobem souvisí i s přesností klasifikace na znělé a neznělé intervaly, došlo k aplikaci informace o (ne)znělosti z metody RAPT, protože dosažené výsledky zobrazené v tabulce 3.1 ukazují, že právě tato metoda je v obou kritériích druhá nejlepší hned po Ewenderově metodě. Po aplikaci informace o (ne)znělosti na Ewenderovu metodu pro stejná data, ze kterých vychází výsledky tabulky 3.1, byly dosaženy výsledky Ewenderovy metody viz následující tabulka 3.2.

Kritérium	**EWENDER**
RMSE [Hz]	15,39
COR	0,783

Tabulka 3.2: Nezvýhodněné výsledky kritérií RMSE a COR Ewenderovy metody

Porovnáním výsledků z tabulky 3.2 s výsledky ostatních metod v tabulce 3.1, vychází Ewenderova metoda stále v obou ohledech jako nejpřesnější a tento výrok je závěrem porovnání uvedeného v této kapitole. Z toho také plyne, že pokud to bude možné, bude v následující kapitole používána tato metoda získávání kontury F_0 v pitch markovacích algoritmech požadujících tuto informaci při své činnosti. Je však nutné dodat, že pro svou funkci potřebuje také externě dodanou informaci o (ne)znělosti získanou z metody RAPT nástroje Wavesurfer.

Kapitola 4

Porovnání algoritmů detekce hlasivkových pulzů

Hlasivkové pulzy nebo též pitch marky jsou popsány podkapitole 2.2.2. Úloha automatické detekce pitch marků v řečových a EGG signálech je v oboru syntézy řeči důležitou úlohou. Bylo již vyvinuto množství algoritmů, které se různými přístupy snaží dosáhnout co možná nejlepších výsledků detekce. V této kapitole bude několik vybraných přístupů prezentováno a bude provedeno srovnání jejich přesnosti a vlastností.

Požadavek na co nejpřesnější detekci okamžiků uzavření hlasivek pouze z řečového signálu je zřejmý z faktu, že konkatenační typy TTS systémů, zejména těch na bázi PSOLA, vyžadují co nejpřesněji detekované pitch marky pro svou činnost. Princip jejich činnosti spočívá v řetězení mikrosegmentů získaných z řečového signálu „rozsekáním“ právě na základě informace o lokaci pitch marků. V současné době nejpřesněji detekující algoritmy pracující s řečovým signálem současně používají i EGG signál. Snahou je získat algoritmus detekující pitch marky pouze z řečového signálu tak, aby se úspěšnost detekce vyrovnala algoritmům používajícím kromě řečového i EGG signál. Důvodem je to, že pořizování EGG signálu se provádí snímačem elektroglotograf připevněným na krk řečníka, ten je po delší době nahrávání řečníkovi nepříjemný, může částečně ovlivňovat hlas a v průběhu nahrávání se může posunout, čímž se mění vlastnosti signálu, který je jeho výstupem. Navíc by korpusy, již zaznamenané bez EGG signálu, mohly být použity pro účel syntézy řeči.

Dalšími úlohami požadujícími detekci pitch marků v řečovém signálu je kromě syntézy řeči například změna řečníka či hlasu, ale také klasifikace řečníka (muž, žena či dítě). V neposlední řadě se jedná o úlohu tzv. pitch-synchronní extrakce příznaků v úloze rozpoznávání řeči.

4.1 Algoritmy detekce hlasivkových pulzů

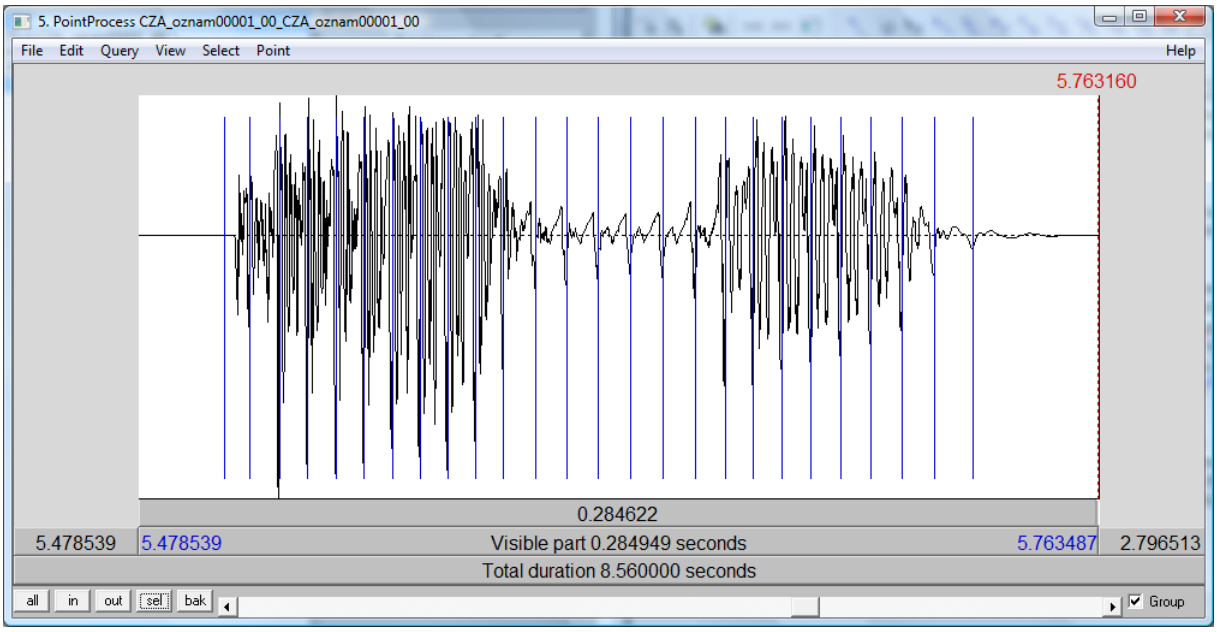
Bylo vybráno a získáno několik dostupných algoritmů řešících úlohu detekce pitch marků v řečovém signálu. Některé byly implementovány. Již zmíněné prostředí PRAAT obsahuje i nástroje pro detekci okamžiků uzavření hlasivek neboli pitch marků v řečovém signálu. Existuje zde dokonce více volitelných možností, jak pitch marky detekovat. Dalším volně dostupným nástrojem zprostředkovávajícím řešení této úlohy je zmíněný toolbox GLOAT, jehož výsledky budou v této kapitole vyhodnoceny. Kromě těchto volně dostupných nástrojů zde budou uvedeny implementované metody. Všechny vyhodnocované algoritmy budou podrobně popsány dále.

4.1.1 PRAAT – Sound&Pitch: To PointProcess(peaks) (*PRAAT-AC*, *PRAAT-CC*)

Nástroj PRAAT byl vytvořen na pracovišti Phonetic Sciences na Amsterdamské univerzitě a je volně dostupný v síti Internet. Tento fonetický program v sobě zahrnuje velké množství elementů pro analýzu, úpravy a popis vlastností zvuku. V možnostech analýzy zvuku se nachází sekce nástrojů zvaná "Periodicity", která obsahuje několik metod detekce kontury F_0.

Jednou z těchto metod je autokorelační metoda "AC" ("auto-correlation") vycházející z hledání maxim autokorelační funkce popsané v kapitole 2.3.2. Specifikum této implementace je, že délka okénka pro konstrukci autokorelační funkce bývá alespoň dvojnásobkem předpokládané nejdelší hlasivkové periody v signálu. Druhou z těchto metod detekce kontury F_0 je korelační metoda "CC" ("cross-correlation"), která funguje v podstatě stejně jako autokorelační metoda a vychází i ze stejné definice viz vztah 2.4. Rozdíl je ve zvolené délce okénka, která se pohybuje pouze okolo průměrné hlasivkové periody v celé nahrávce řečového signálu nebo v určitém okolí analyzovaného segmentu. Analýza obou těchto způsobů probíhá tedy v časové doméně a výsledkem obou je také objekt "Pitch". Obě tyto 2 metody detekce kontury F_0 mají své výhody i nevýhody a vykazují zvláště v problematických segmentech řečového signálu rozdílné výsledky [14].

Jádro metody detekce pitch marků implementované v prostředí PRAAT pro svůj výpočet používá objekt "Sound", tedy objekt obsahující načtenou nahrávku analyzovaného zvuku. Kromě objektu "Sound" používá také objekt "Pitch", který obsahuje konturu F_0 získanou pomocí metody "AC" nebo "CC". Celý název této metody je proto označen jako "Sound&Pitch: To PointProcess(peaks)". Nadále bude tato metoda nazývána *PRAAT-AC* a *PRAAT-CC* podle toho, pomocí jaké z metod (AC nebo CC) byl získán objekt "Pitch". Nastavení nástroje To PointProcess(peaks) je volba, zda detekovat pitch marky v maximech nebo v minimech řečového signálu a výsledný objekt se nazývá "PointProcess".



Obrázek 4.1: Příklad zobrazení nahrávky s lokalizovanými pitch marky v prostředí PRAAT

Výsledky metod *PRAAT-AC* a *PRAAT-CC* budou uvedeny na konci kapitoly. Zahrnutí těchto metod do srovnání bylo inspirováno v [15].

4.1.2 GLOAT – Detekce událostí v řeči využívající reziduální excitaci a průměrovaný signál (*SEDREAMS*)

Metoda detekce událostí v řeči využívající reziduální excitaci a průměrovaný signál (z angl. speech event detection using the residual excitation and a mean-based signal, zkráceně *SEDREAMS*) byla prezentována v disertační práci [13] Thomasem Drugmanem. Zkratka GLOAT pochází z "GLOttal Analysis Toolbox"a jedná se o sadu metod pro analýzu činnosti řečového ústrojí z řečových nahrávek. Zmíněný toolbox byl vytvořen v prostředí Matlab a klíčová funkce pro úlohu detekce pitch marků v řečovém signálu je *SEDREAMS_GCIDetection.m*. Metoda bude nadále označována jako *SEDREAMS*. Z vlastností této metody vyplývá, že dochází k detekci pitch marků napříč celou nahrávkou nehledě na (ne)znělost segmentů řečového signálu. Pro vyřazení pitch marků, které byly detekovány v neznělých segmentech se navíc používá informace o (ne)znělosti získaná z funkce *SRH_PitchTracking.m*, která je taktéž nástrojem toolboxu GLOAT, a která je blíže popsána v podkapitole 3.1.5.

Princip metody *SEDREAMS* spočívá v několika krocích. V tom prvním se počítá průměrovaný signál $y(n)$ podle vztahu

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m), \tag{4.1}$$

kde $w(m)$ je okénko délky $2N + 1$ a $s(n)$ je původní řečový signál. Výsledný průměrovaný signál má určité vlastnosti, ze kterých metoda *SEDREAMS* těží. Hlavní takovou vlastností je, že se průměrovaný signál blíží signálu sinus s variabilní frekvencí podle základního hlasivkového tónu analyzované řečové nahrávky. Výhodou průměrovaného signálu je také jeho spojitost na celém intervalu. Vždy jedna perioda znělého segmentu řečového signálu odpovídá jedné periodě průměrovaného signálu. „Naneštěstí" zde existuje také fázový posuv, vždy mezi dvěma příslušícími periodami řečového a průměrovaného signálu, a tento fázový posuv se obecně v čase mění. Nelze tedy tvrdit, že pitch marky vždy souvisí například s minimem tohoto průměrovaného signálu. Drugman na tento problém poukázal a vyřešil ho empiricky. Pomocí dostatečného množství dat od různých řečníků vytvořil určitý histogram pitch marků v relativním vztahu k fázi příslušné periody průměrovaného signálu a z výsledku stanovil, že pitch marky se nacházejí obecně od 50 do 85% lokální periody uvažované od maxima k sousednímu maximu pro řečové signály se zápornou polaritou. Pro signály s polaritou kladnou je lokální perioda uvažována vždy mezi dvěma sousedními lokálními minimy. Požadavek na znalost polarity je vyřešen použitím funkce *OMPD_PolarityDetection.m* pro detekci polarity řečového signálu. Tím byl dokončen první krok, který definuje intervaly, kde jsou očekávány pitch marky.

V kroku druhém přichází řada na reziduální signál lineární predikce (z angl. linear prediction (LP) residual signal). Zjednodušeně se jedná o část signálu, která je odstraněna inverzní filtrací řečového signálu. Inverzní filtrací je myšlen proces inverzní k procesu vytváření řeči a jedná se tím pádem o získávání excitace z řečového signálu. Důležitá vlastnost takto vzniklého signálu je, že se vyskytují lokální maxima v těch místech, kde se vyskytuje energetická změna původního řečového signálu, která odpovídá okamžiku uzavření nebo otevření hlasivek. V předchozím odstavci byl

popsán první krok, ze kterého vyplývaly intervaly, ve kterých jsou očekávány pitch marky. V závěru tohoto druhého kroku dochází ke kombinaci informací z obou kroků, hledají se tedy maxima popsaného signálu v intervalech očekávajících pitch marky. Lokální maxima odpovídající hlasivkovým otevřením jsou eliminovány, jelikož se vyskytují zpravidla mimo zmíněné intervaly.

Posledním krokem je vyloučení těch pitch marků, které jsou mimo znělé intervaly. Informace o znělosti byla získána pomocí Matlabovské funkce *SRH_PitchTracking.m*. Výsledek je znázorněn na obrázku 4.2.

Popis obrázku 4.2

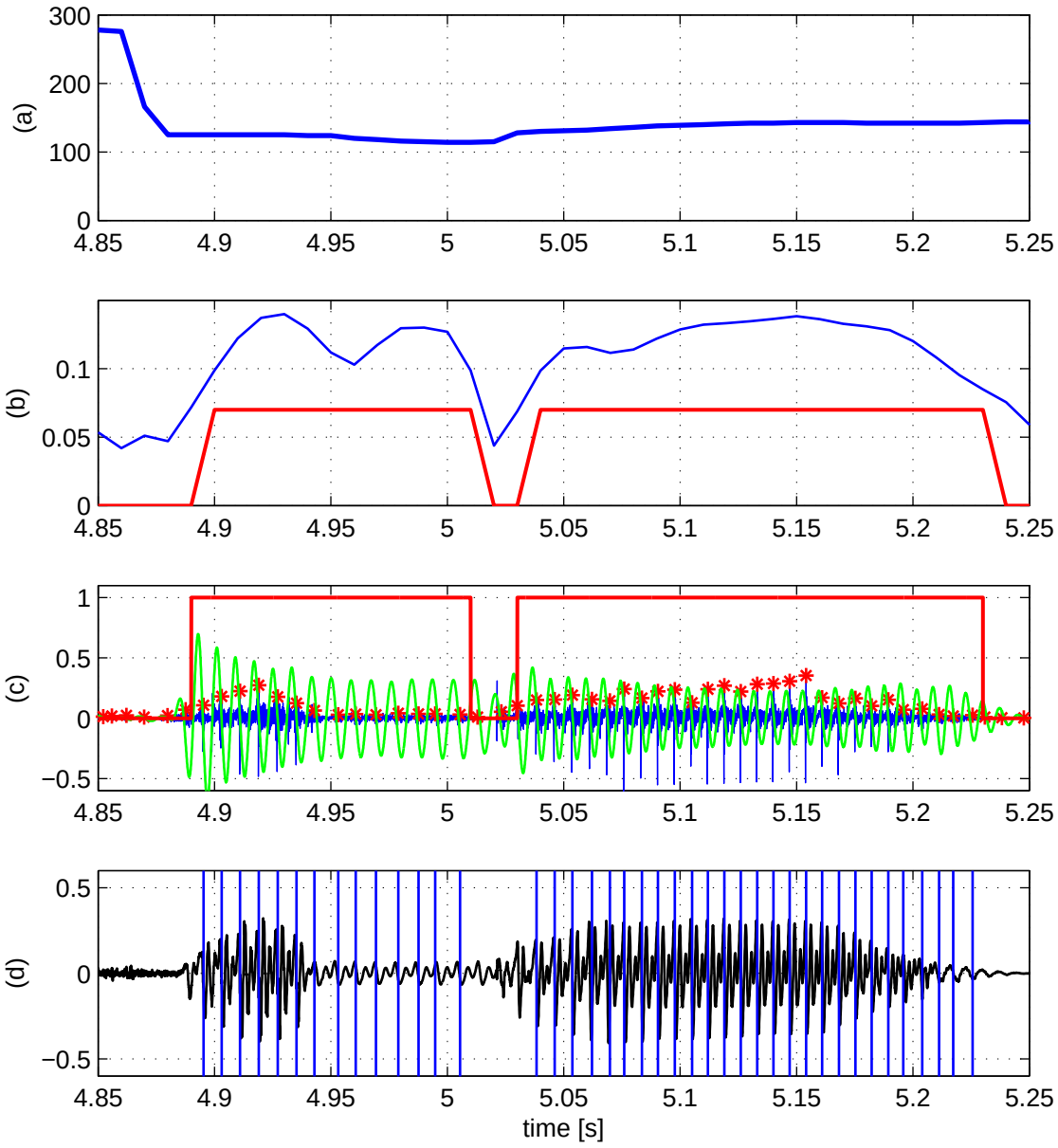
(a) – Na tomto grafu se vyskytuje kontura F_0 základního hlasivkového tónu. V levé části je zde patrný interval, kde hodnoty jsou nepřirozeně vysoké. To je charakteristické pro takto získávanou konturu, jedná se však o neznělý interval.

(b) – Zde je znázorněna SRH funkce modře, tato funkce byla vysvětlena v podkapitole 3.1.5. Prahováním této funkce byla získána červeně vyznačená funkce, která vyjadřuje rozhodnutí o znělosti/neznělosti analyzovaného řečového signálu.

(c) – V tomto grafu je vyznačen LP zbytkový signál modře. Ten se vyznačuje lokálními extrémy (maximy), které odpovídají pozicím pitch marků. Lokální maxima, která byla klasifikována jako kandidáti na pitch marky, jsou označeny červenými body. Zeleně je označen průměrovaný signál, který byl vypočítán ze signálu řečového. Červeně je vyznačeno rozhodnutí o (ne)znělosti analyzovaného řečového signálu po korekci.

(d) – Černě je vykreslený řečový signál, modré jsou odpovídající detekované pitch marky, které vznikly z kandidátů v grafu (c) výběrem těch, které se nacházejí ve znělé části promluvy.

Na nezávisle proměnných osách všech čtyř podgrafů je čas v sekundách. Je zde také pozorovatelný určitý typ chyby, jelikož v podgrafu (b) SRH funkce klesne po 5. sekundě pod hodnotu prahu. Tím se zde objeví neznělý úsek, ale z řečového signálu v podgrafu (d) je jasně patrné, že zde chybí 3 pitch marky. S určováním znělosti/neznělosti se tato metoda ukázala jako problematická, ve výsledcích ke konci této kapitoly bude předvedeno alternativní řešení.



Obrázek 4.2: **(a)** kontura F_0 [Hz]; **(b)** SRH funkce (modře) a klasifikace (ne)znělosti (červeně); **(c)** LP reziduální signál (modře), označení pitch marků (červené body), průměrovaný signál (zeleně) a klasifikace (ne)znělosti (červeně); **(d)** řečový signál (černě) a označení pitch marků jen ve znělých intervalech (modře)

4.1.3 Ewender – Přesná detekce pitch marků pro úpravy prozódie řečových segmentů (*EWM*)

Metoda přesné detekce pitch marků pro úpravy prozódie řečových segmentů (z angl. accurate pitch marking for prosodic modification of speech segments) je popsána v článku [7] a byla prezentována Thomasem Ewenderem na konferenci Interspeech v roce 2010. Metoda byla zvolena k implementaci v prostředí Matlab, bude dále označována jako metoda *EWM*. Myšlenka tohoto algoritmu vychází z průběhů dvou funkcí. První z nich je nazývána spojitý průběh krátkodobé energie (z angl. continous short-term energy countour). Lokální maxima této funkce vyjadřují lokální zvýšení energie, o kterých se předpokládá, že vznikla následkem uzavření hlasivek. Druhou funkcí je „základní vlna“ (z angl. fundamental wave), která pochází z nízkofrekvenční složky řečového signálu, který obsahuje „nejčistší“ informaci o činnosti hlasivek. První z těchto funkcí je používána jako primární, druhá doplňuje informaci v segmentech signálu, kde primární průběh krátkodobé energie nevykazuje úspěšnou detekci maxim. Z kombinace obou funkcí je možné úspěšně získat umístění pitch marků v řečovém signálu.

Výpočet spojitého průběhu krátkodobé energie vychází z řečového signálu a princip byl nastíněn v kapitole 2.3.1. V tomto případě je zde navíc normalizace funkce podle velikosti okénka. Dochází totiž k výpočtu pro každý vzorek signálu s variabilní velikostí okénka měnící se podle hodnoty základního hlasivkového tónu v příslušném bodě. Bez normalizace by pro delší okénka při výpočtu průběhu energie vycházely vyšší hodnoty v rámci jedné nahrávky a nemuselo by pak dojít k úspěšné detekci. Je zde tedy požadavek na Hammingovo okénko o měnící se (celočíselné) délce $T_0 = \frac{1}{F_0}$ pro každý vzorek, kde F_0 je základní hlasivkový tón v konkrétním vzorku řečového signálu. Základní hlasivkový tón F_0 je pro každou nahrávku získáván pomocí metody *detect_F0_contour.m* prezentované v podkapitole 3.1.4. Tento vyšel jako nejpřesnější z porovnání uvedeného v kapitole 3, ve které také bylo zmíněno, že jako doplňková informace je potřebné rozhodnutí o (ne)znělosti segmentů. Zde se liší postup prezentovaný v článku [7] od implementace, protože v tomto případě na rozdíl od prezentované metody není k dispozici neuronová síť pro klasifikaci segmentů na znělé a neznělé. Chybějící informace bude v případě této práce získávána z metody RAPT programu WaveSurfer.

Kontura základního hlasivkového tónu je počítána s krokem 5 ms, a hodnota tak není dostupná pro každý vzorek. Proto je nutná interpolace, čímž získáme hodnotu F_0 pro každý vzorek. V tomto případě byla použita lineární interpolace. Spojitý průběh krátkodobé energie pak může být stanoven pro každý (n-tý) vzorek pomocí vztahu

$$E(k) = \frac{\sum_n \left[s(n)w(n-k)\right]^2}{\sum_n w(n)^2}, \tag{4.2}$$

kde $j = 1, 2, ..., T_0$, s představuje řečový signál a w Hammingovo okénko délky T_0 pro každý vzorek. Druhou funkcí která je v této metodě potřebná je fundamentální vlna. Ta je rovněž vypočítávána pro každý vzorek signálu a proto vyžaduje hodnotu T_0 pro každý vzorek, ale tento výpočet byl již obstarán pro výpočet průběhu krátkodobé energie. Její hodnota pro každý vzorek se získává konvolucí centrované části signálu o délce T_0 s Hammingovým oknem o stejné délce. Jedná se vlastně o dolnopropustní filtraci řečového signálu se zlomovou frekvencí rovnou základnímu hlasivkovému tónu. Dojde proto k odfiltrování složek s frekvencí vyšší, než je základní hlasivkový tón. Maxima této základní vlny však není možné použít přímo jako hlasivkové pulzy. Podobný problém řešil také Drugman ve své metodě prezentované v podkapitole 4.1.2.

Zde je možné si všimnout určité analogie s průměrovaným signálem definovaným vztahem 4.1. Způsob výpočtu průměrovaného signálu a fundamentální vlny je různý. Amplituda se však liší o více než 2 řády. A je zde i rozdíl ve tvaru funkcí. Společný je však fázový posuv a průchody nulou se přesně kryjí. Oba průběhy byly pro zajímavost normalizovány maximální absolutní hodnotou a došlo k porovnání pomocí vzájemné (Pearsonovy) korelace na několika nahrávkách se střední hodnotou výsledných korelací 0,973.

Jsou tedy definovány 2 funkce, průběh krátkodobé energie a fundamentální vlna. Ani jedna sama o sobě nemá předpoklady na úspěšnou detekci pitch marků, jelikož obě dvě se potýkají s problémy. Stručně se jedná o zdvojení frekvence (z angl. pitch doubling), se kterou jsou pitch marky detekovány, u téměř sinusových signálů v podání průběhu krátkodobé energie. Dalším problémem je efekt jitter, anebo v případě smíšených zvuků se jedná o nízké falešné vrcholky také průběhu krátkodobé energie. V případě fundamentální vlny je největší a zároveň jedinou překážkou měnící se fázový posun v čase.

Myšlenkou je, že pokud jsou vrcholky průběhu krátkodobé energie podle určitých kritérií kvalitní, tak jsou přímo určeny jako pitch marky. Pokud kvalitní nejsou, přichází na řadu fundamentální vlna, která krátkodobě nijak výrazně svůj fázový posuv nemění, a proto se použije na doplnění menšího množství pitch marků tak, že se volí stejná fáze, v které byl lokalizován nejbližší pitch mark z průběhu energie.

Zmíněná kritéria jsou popsány několika pravidly v článku [7], které byly integrovány do algoritmu. Celý algoritmus byl implementován jako funkce *findGCI.m*, kde vstupem je název souboru k analýze a název souboru s konturou F_0 z metody RAPT, avšak pouze kvůli informaci o (ne)znělosti. Výstupem je pak vektor časů reprezentujících nalezené pitch marky v řečovém signálu. Výsledek práce této funkce je znázorněn na obrázku 4.3.

Popis obrázku 4.3

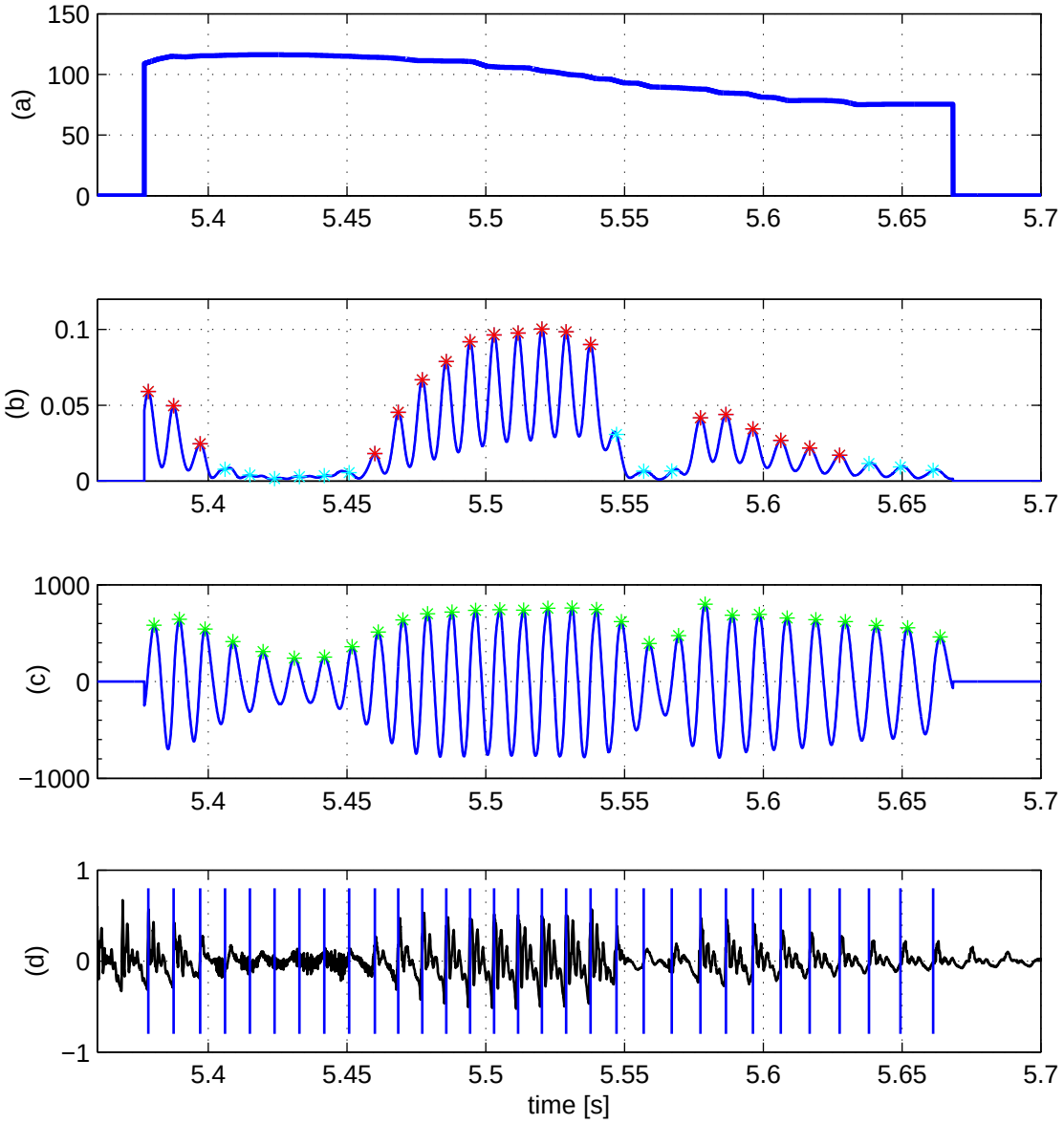
(a) – Na tomto grafu se vyskytuje kontura F_0 základního hlasivkového tónu, která vznikla kombinací Ewenderovy funkce *detect_F0_contour.m* a kontury F_0 metody RAPT, ze které byla extrahována pouze informace o (ne)znělosti.

(b) – Zde je znázorněn spojitý průběh krátkodobé energie modře, nalezená maxima tohoto průběhu odpovídající pitch markům jsou zaznamenány pomocí červených bodů, azurově jsou vyznačeny ty body, které byly doplněny na základě fundamentální vlny a lokálních hodnot kontury F_0.

(c) – V tomto grafu je modře vyznačena fundamentální vlna a její lokalizovaná maxima zeleně potřebná pro určování lokální fáze vůči průběhu krátkodobé energie.

(d) – Černě je vykreslený řečový signál, modré jsou odpovídající detekované pitch marky.

Na nezávisle proměnných osách všech čtyř podgrafů je čas v sekundách. Tato metoda má určité výhody i nevýhody. Vzhledem k tomu, že vychází z energie definované vztahem 4.2, tak sice nepotřebuje, ale vlastně ani neumí využít informaci o polaritě řečového signálu. Někdy se stává, že menší část řečového signálu má polaritu opačnou, než celá nahrávka. V takovém případě dochází k detekci pitch marků nekonzistentně. Tento problém byl částečně vyřešen později.



Obrázek 4.3: **(a)** kontura F_0 [Hz]; **(b)** Průběh krátkodobé energie (modře), body nalezené z průběhu krátkodobé energie (červeně) a body doplněné z fundamentální vlny a kontury F_0 (azurově); **(c)** Fundamentální vlna (modře) a lokalizovaná lokální maxima (zeleně) **(d)** řečový signál (černě) a označení pitch marků (modře)

4.1.4 Neuronová síť pro detekci pitch marků (*NNPM*)

Metoda neuronová síť pro detekci pitch marků (z angl. pitch detection with a neural-net classifier) popsaná v článku [16] byla prezentována na konferenci Signal Processing již v roce 1991. Implementace není volně dostupná a proto byla tato metoda zvolena k implementaci v prostředí Matlab. Metoda bude nadále označována jako *NNPM*. Jedná se o detekci pitch marků v řečovém signálu s použitím neuronové sítě. Autoři této metody prezentovali dva přístupy k realizaci. V prvním přístupu neuronová síť používá určitý počet vzorků signálu okolo klasifikovaného vrcholku jako vstup. Ve druhém přístupu neuronová síť používá jako vstup dále uvedený soubor lokálních vlastností signálu. Autoři empiricky zjišťují optimální počty parametrů pro oba přístupy a v závěrečné fázi článku hodnotí jejich úspěšnosti. Druhý přístup využívající soubor lokálních vlastností vykazuje chybu 2 % což je oproti 2,5 % prvního přístupu lepší výsledek. Tato práce se proto bude zabývat pouze druhým přístupem, tedy realizací neuronové sítě vycházející ze souboru lokálních vlastností.

Kvůli velké komplexitě řečového signálu byl pro její snížení použit low-pass filtr se zlomovou frekvencí 700 Hz. Znamená to, že vysokofrekvenční složky řečového signálu jsou eliminovány a z řečového signálu zůstane pouze hladká vlna. Byl použit filtr s nulovým fázovým posuvem tak, aby nedošlo ke zkreslení poloh maxim a minim. Dále se pracuje pouze s odfiltrovaným signálem, ve kterém jsou nalezeny průchody nulou a mezi každými dvěma průchody pak maximum nebo minimum. Předpokladem je, že pitch marky, původně korespondující například k nějakému lokálnímu minimu řečového signálu, budou po filtraci konzistentně náležet odpovídajícím nejbližším minimům odfiltrovaného signálu a nedojde tak ke zkreslení poloh pitch marků.

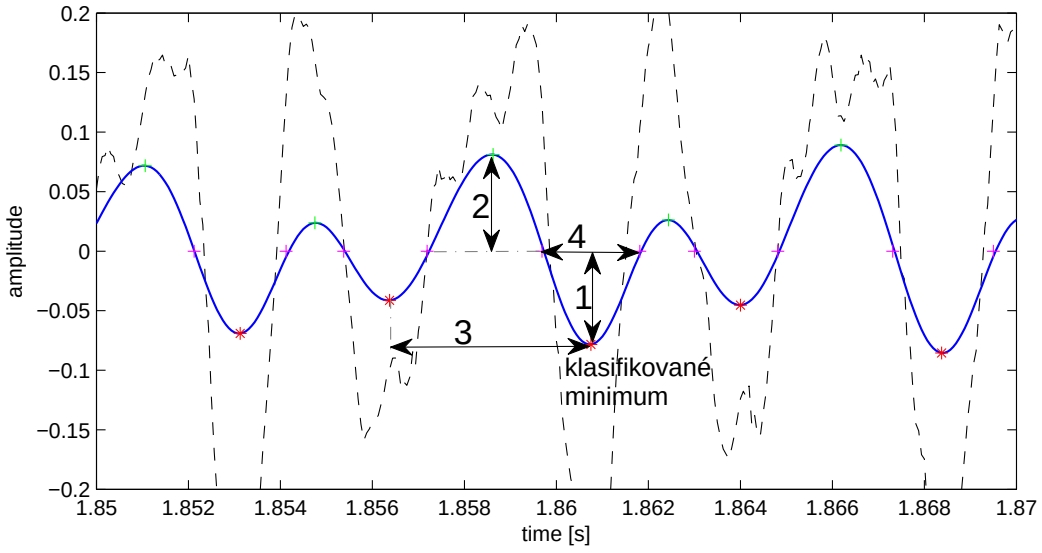
Data pro účel realizace této sítě byla poskytnuta pracovištěm KKY ve složení celkem 83 (viz Korpus83 v příloze A) několikavteřinových nahrávek ve formátu jednokanálových 16-bitových souborů typu ".wav" a jim odpovídajících 83 souborů obsahující časy ručně určených pitch marků. Celkem se jedná o více než 11 minut záznamu obsahujících více než 50 tisíc ručně určených pitch marků. Nahrávky obsahují promluvy v češtině, slovenštině, angličtině, němčině a francouzštině od řečníků obou pohlaví. Vyskytují se zde také nahrávky s veselým intonačním zabarvením a dále 3 nahrávky, které obsahují problematické nepravidelné úseky. Z těchto dat pak také vychází i celkové množství vrcholků ke klasifikaci, respektive vstupních vektorů datové množiny 84 634. Všechny nahrávky měly v našem případě zápornou polaritu ve smyslu, že pitch marky odpovídaly vždy lokálním minimům.

Soubor lokálních parametrů

Autoři zmíněného článku dosáhli nejlepších výsledků použitím celkem pěti typů deskriptorů. Jedná se o amplitudu minima, amplitudu maxima, časový rozdíl, šířku pulzu a korelaci. Ke každému minimu, které má být klasifikováno, zda se jedná nebo nejedná o pitch mark, tak potřebujeme určitý počet těchto vlastností na každou stranu. Tento počet autoři stanovili na $n = 4$, parametry jsou tedy následující:

- Amplituda minima – první z deskriptorů, demonstrováno šipkou 1 v obrázku 4.4. Hodnota je normalizována minimální hodnotou ve vzdálenosti 250 ms na obě strany. Pro $n = 4$ jsou potřebné 4 hodnoty doleva, 4 hodnoty doprava a hodnota klasifikovaného minima, celkem tedy 9 parametrů.

- Amplituda maxima – druhý deskriptor, demonstrováno šipkou 2 v obrázku 4.4. Hodnota je normalizována maximální hodnotou ve vzdálenosti 250 ms na obě strany. Pro $n = 4$ jsou potřebné 4 hodnoty doleva, 4 hodnoty doprava, celkem tedy 8 parametrů.
- Časový rozdíl – třetí z deskriptorů, demonstrováno šipkou 3 v obrázku 4.4. Hodnota je normalizována hodnotou 20 ms. Pro $n = 4$ jsou potřebné 4 hodnoty doleva, 4 hodnoty doprava, celkem tedy 8 parametrů.
- Šířka pulzu – čtvrtý deskriptor, demonstrováno šipkou 4 v obrázku 4.4. Hodnota je normalizována hodnotou 2 ms. Pro $n = 4$ jsou potřebné 4 hodnoty doleva, 4 hodnoty doprava a hodnota klasifikovaného minima, celkem tedy 9 parametrů.
- Korelace – poslední deskriptor, pro $n = 4$ byly použity hodnoty negativní korelace (max korelovanost odpovídá nule, minimální jedné) údolí klasifikovaného minima se 4 údolími vlevo a 4 údolími vpravo, celkem 8 parametrů. Dohromady pro $n = 4$ se jedná o 42 parametrech příslušících jednomu minimu ke klasifikaci.

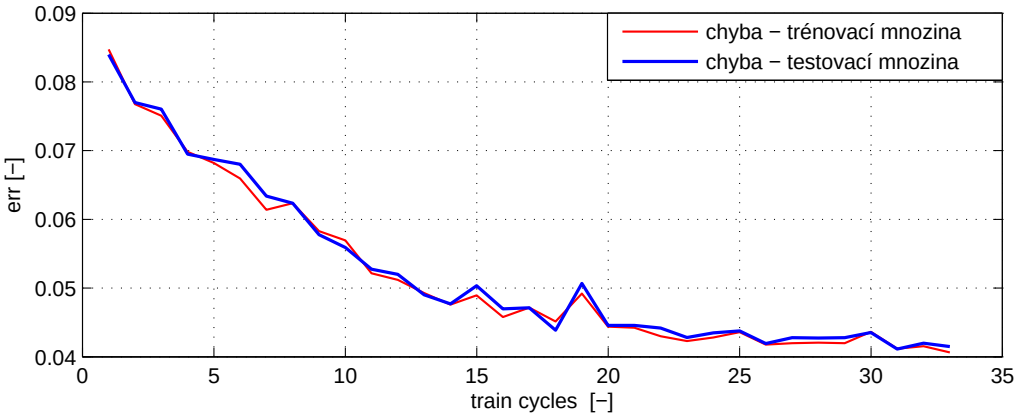


Obrázek 4.4: Ukázka struktury deskriptorů, filtrovaný signál (modře), řečový signál (černě).

Fáze trénování

V prostředí Matlab byla vytvořena funkce *NN_PM.m*, jejímiž vstupy jsou cesty ke složce se zvukovými soubory a ke složce s referenčními pitch marky. Byl zvolen stejný typ sítě, tedy dvouvrstvá nelineární síť, i stejný trénovací algoritmus ("backpropagation"– viz podkapitola 2.6). Autoři pro $n = 4$ neuvádějí počet neuronů ve skryté vrstvě, v tomto případě byl tedy stanoven na 20. Algoritmus umožňuje změnu neuronů ve skryté vrstvě změnou jedné konstanty, stejně tak lze měnit n udávající počet parametrů změnou jiné konstanty. Dále je zde konstanta ovlivňující poměr rozdělení datového setu na trénovací a testovací. Ve skryté vrstvě byly použity nelinerární aktivační funkce, a sice bipolární spojité. Ve vrstvě výstupní byly použity aktivační funkce lineární.

Funkce *NN_PM.m* tedy načte postupně všechny nahrávky, odfiltruje je, vyhledá průchody nulou, maxima a minima, ke každému minimu vypočte vektor 42 parametrů. Pro vytvoření datového setu je zaměstnán skript *create_set.m*, který z 83 nahrávek vytvoří matici obsahující 84 634 vektorů po 42 parametrech. Podle referenčních pitch marků byly vytvořeny vektory, které ke každému vstupu definují požadovaný výstup, tedy zda se jedná nebo nejedná o pitch mark. Následuje rozdělení na trénovací a testovací množinu a samotné trénování, které trvá, dokud neklesne chyba pod nastavitelnou mez nebo dokud se nevyčerpá nastavitelný povolený počet trénovacích cyklů. Bylo dosaženo 33 trénovacích cyklů, během kterých se chyby na trénovací a testovací množině chovaly jak ukazuje obrázek 4.5.

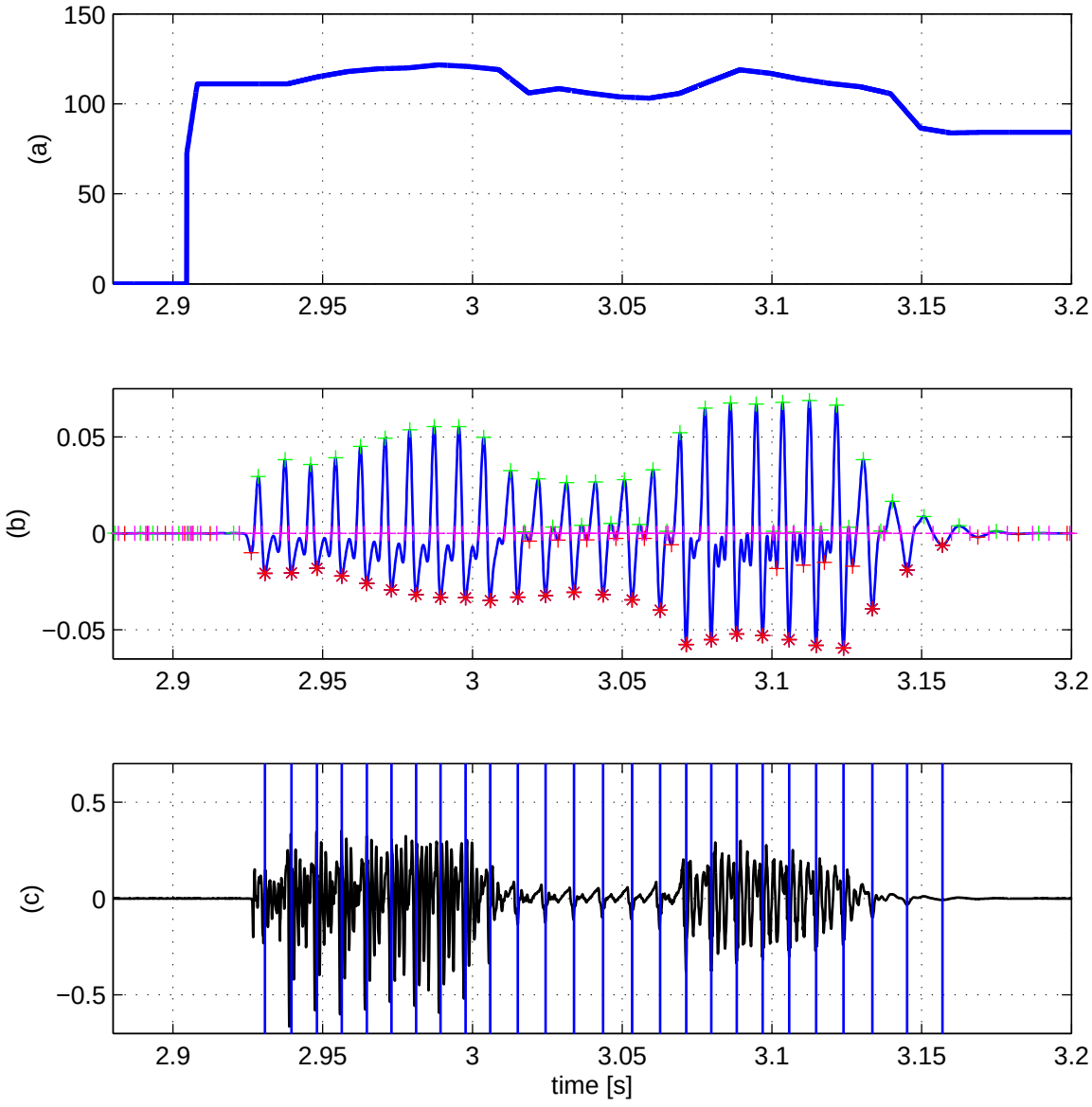


Obrázek 4.5: Průběh chyby na trénovací (červeně) a testovací (chyby) množině během trénování

Fáze pracovní

Po natrénování byly uloženy 4 matice reprezentující natrénovanou neuronovou síť do souboru *NN33.mat* a byla vytvořena funkce *find_PM.m*, jejímž vstupem je zvukový soubor. Tato funkce vytvoří z nahrávky příslušný počet 42-parametrových vektorů ke klasifikaci, kterých je tolik, kolik je nalezeno minim ve filtrovaném signálu. Vektor po vektoru je klasifikován a výsledkem jsou časy těch minim, které byly vyhodnoceny jako pitch mark. Později byla zakomponována informace o (ne)znělosti metody RAPT a tím mohou být eliminovány pitch marky, které se nacházejí v neznělých částech řečového signálu. Znělé intervaly však byly o 30 ms rozšířeny, aby se zmenšil podíl případné chybovosti kontury F_0. Ukázka činnosti tohoto algoritmu je zobrazena na grafech obrázku 4.6.

Na nezávisle proměnných osách všech tří podgrafů je čas v sekundách. Tato funkce však ke své činnosti potřebuje informaci o polaritě řečového signálu. Nepřesností této metody je zejména posuv pitch marků vlivem dolno-propustní filtrace. Tento problém byl částečně vyřešen dále v této kapitole.



Obrázek 4.6: **(a)** kontura F_0 [Hz]; **(b)** Průběh low-pass zero-phase (700Hz) filtrovaného signálu (modře), nalezená lokální maxima (zeleně), průchody nulou (purpurově), minima ke klasifikaci (červené „+“), minima klasifikované jako pitch marky (červené „*“); **(c)** řečový signál (černě) a označení pitch marků (modře)

4.1.5 KKY – Více-stupňový algoritmus (*MPA*)

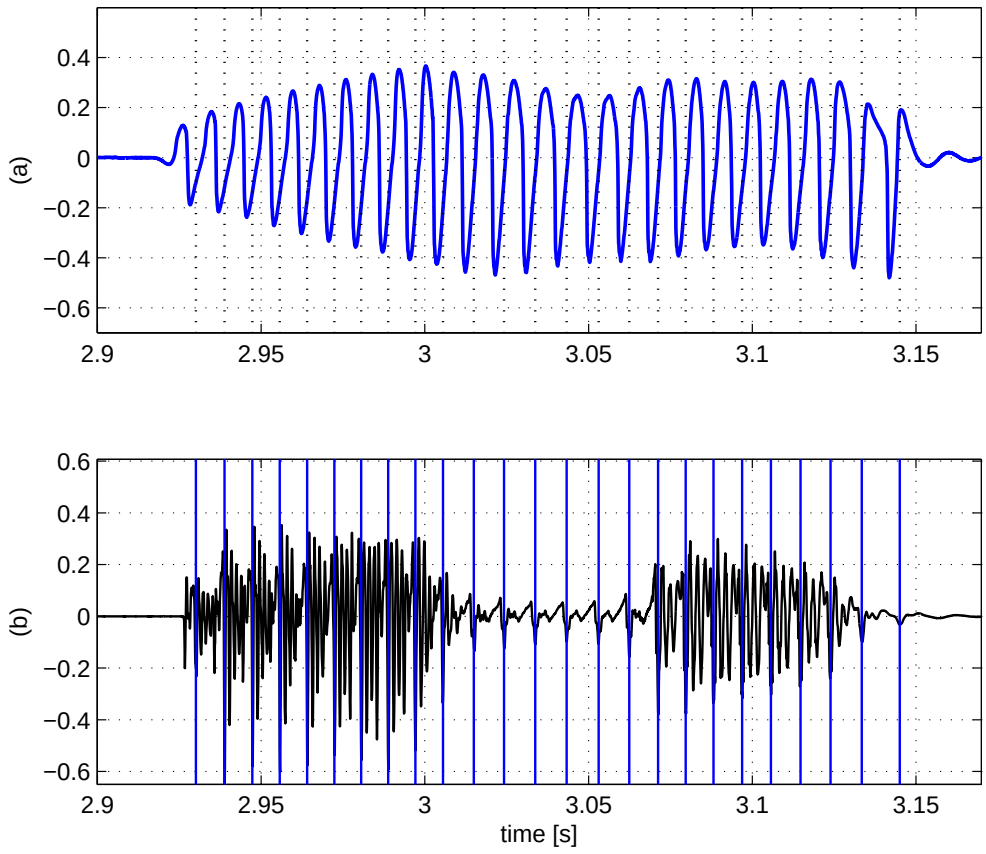
Více-stupňový algoritmus (z angl. multi-phase algorithm, zkráceně *MPA*) popsaný v článku [15] byl prezentován v roce 2011. Algoritmus *MPA* byl vyvinut na pracovišti Katedry kybernetiky Fakulty aplikovaných věd v Plzni. Vychází kromě řečového signálu hlavně ze signálu EGG (viz podkapitola 2.2.1) a v článku, kde byl prezentován, vykazuje velmi dobré výsledky. Cílem této práce je srovnání algoritmů pro detekci pitch marků. Důležité je právě srovnání s algoritmem *MPA*, který bude stručně popsán v této podkapitole. *MPA* totiž jako jediný používá doplňující informaci o činnosti hlasivek, proto jsou jeho výsledky zpravidla lepší.

Před procedurami algoritmu *MPA* je analyzován řečový signál ve smyslu detekce polarity tohoto signálu. Byl pozorován velký rozptyl ve výsledcích pitch markovacích algoritmů a autoři odhalili, že velký podíl na tomto rozptylu má právě žádná nebo nepřesná detekce polarity analyzovaného signálu. Procedura klasifikující signál podle polarity spočívá v porovnání počtu lokálních maxim a minim řečového signálu v pořadí od nejexcentričtějších dokud nedojde k poklesu energie zbývajícího signálu pod určitou mez. Jakmile je nějaké lokální minimum nebo maximum zahrnuto do porovnání, žádné další maximum ani minimum z okolí $\frac{4}{3} \times T_0$ nebude zahrnuto. Je-li nalezeno více maxim, výsledný signál má pozitivní polaritu. Pokud převažují minima, jedná se pak o negativní polaritu. Pokud navíc převažuje energie ve stejné polorovině signálu, rozhodnutí je definitivní. Pokud tato energie s rozhodnutím nekoresponduje, rozhodnutí je označeno za „nejisté“.

V následující sekci článku [15] je popsána procedura “BaseLine Algorithm (BLA)”, kde dochází k vytvoření diferenčního signálu DEGG ze signálu EGG. Dále dochází k prahování a vyhlazování vzniklého signálu, který vykazuje ostrá údolí v místech, kde pravděpodobně došlo k uzavření hlasivek. Následuje určitý “postprocessing”mající za úkol eliminovat pitch marky v neznělých oblastech signálu a naopak chybně neklasifikované pitch marky ve znělých částech signálu doplnit. V konečné fázi tohoto algoritmu dochází k synchronizaci se signálem řečovým. Autoři podrobně popsali nedostatky tohoto algoritmu a vyplývající potřebu navrhnout algoritmus robustnější.

MPA – struktura

První dva bloky samotného *MPA* se zabývají segmentací signálu podle (ne)znělosti z EGG i řečového signálu a detekcí kontury F_0 z EGG signálu. Tyto „dvě nově vzniklé“ informace jsou využity dále v modifikovaném BLA, který místo globálního prahu využívá právě konturu F_0 a pomocí informace o neznělosti eliminuje pitch marky v neznělých částech promluvy. Tímto vzniká hlavní část sady kandidátů na pitch marky, která je doplněna o kandidáty nalezené pomocí metody “Simple Pitch Marking Method (SPM)”popsané v článku [17], což je procedura starající se o hledání možných pitch marků z řečového signálu tak, že mezi kandidáty řadí i často přítomné pulzy vedlejší, které mají alespoň 80% amplitudy hlavních pulzů. Všechny kandidáty na pitch marky ohodnocuje další část *MPA* podle jejich amplitudy, pozice v řečovém signálu, ale také podle jejich následovníka. Ze sady takto ohodnocených kandidátů může být vybrána optimální sekvence pitch marků. Pozice takových pitch marků jsou dále vyhlazeny, čímž vzniká konečná sekvence pitch marků detekovaných pomocí *MPA*. V samém závěru ještě dochází k ohodnocení nalezené sekvence pitch marků, což s sebou nese výhodu ve formě informace o tom, s jak velkou mírou jistoty byl ten který pitch mark detekován. Ukázka detekce přístupu implementovaného pomocí *MPA* je znázorněna na obrázku 4.7 spolu s výchozím EGG signálem.



Obrázek 4.7: **(a)** EGG signál zobrazuje činnost hlasivek (modře), detekované pitch marky reflektované z podgrafu (b) (černě); **(b)** řečový signál (černě) a označení pitch marků (modře).

4.1.6 Další potenciální metody

Metod určených pro řešení úlohy automatické detekce hlasivkových pulzů v řečovém signálu existuje samozřejmě více, než je popsaných v této práci. Pokud by tato práce měla být rozšířena zahrnutím dalších metod, v první řadě by se zde objevila metoda "A Two-Phase Pitch Marking Method for TD-PSOLA Synthesis"prezentovaná článkem [17], která se částečně stala inspirací pro metodu *MPA* popsanou v podkapitole 4.1.5.

Dalšími potenciálními metodami pro možné budoucí zahrnutí do obou porovnání této práce jsou metody "Noise robust F_0 determination and epoch-marking algorithms"popsané v článku [18]. Jsou zde obsaženy klasifikátory znělosti realizované neuronovými sítěmi, což by mohlo být zajímavé porovnání například s neuronovou sítí prezentovanou Ewenderem v [10] nebo s klasifikací znělosti pomocí metody RAPT.

4.2 Použitý způsob porovnání dvou sekvencí pitch marků

V této podkapitole bude popsána ověřená metoda sloužící pro porovnání dvou sekvencí pitch marků, kdy jedna z nich je považována za vztažnou neboli referenční. Jedná se o jednoduchý skript nazvaný *StatisticsPM*. Skript byl poskytnut k použití vedoucím této práce.

StatisticsPM porovnává dvě sekvence pitch marků reprezentované pomocí časů, ve kterých byly detekovány. Požadovaný formát těchto souborů je definován tak, že na každém řádku se nachází čas pitch marku reprezentovaný desetinným číslem a mezerou oddělený typ pitch marku (V – voiced, U – unvoiced, T – transition). Je zde požadavek, aby souvislé znělé úseky řeči byly ohraničeny pitch marky typu „T“ a sice v obou sekvencích. Tedy například $T\,V\,V\,V\,T\,T\,V\,V\,T$ (bez časů), což znázorňuje začátek první znělé části, ve které jsou 3 pitch marky a následuje konec první znělé části a začátek druhé znělé části čítající 2 pitch marky, následuje ukončující poslední pitch mark „T“.

Při volání skriptu jsou vyžadovány celkem tři parametry v tomto pořadí: název referenčního souboru s pitch marky, mezerou oddělený název testovaného souboru s pitch marky a opět mezerou oddělené číslo znamenající maximální povolenou (nepenalizovanou) vzdálenostní *odchylku*. Skript využívá techniku transformace testované sekvence na referenční sekvenci pouze za pomoci 3 základních operací:

- Posun pitch marku – posune pitch mark o potřebnou hodnotu. Posunutí v lokální periodě procentuálně menší, než bylo vstupem v parametru *odchylka*, není penalizováno. Posunutí o vzdálenost větší zvýší počet N_S posunutých pitch marků.
- Vymazání pitch marku – vymaže pitch mark, jež by v daném místě neměl být. Vymazání zvýší počet N_D vymazaných pitch marků.
- Vložení pitch marku – vloží pitch mark, jež v daném místě chybí. Vložení zvýší počet N_I vložených pitch marků.

Algoritmus pro srovnání sekvencí hledá co nejkratší posloupnost takových operací, aby byla testovací sekvence pitch marků převedena na referenční. Výsledná přesnost testované sekvence vzhledem k referenční je pak definována vztahem

$$Acc = \frac{N - N_S - N_D - N_I}{N} \cdot 100 \quad [\%] \tag{4.3}$$

$$N = max(N_t, N_r) - N_? \tag{4.4}$$

kde N_t je počet testovaných pitch marků, N_r je počet pitch marků referenčních a $N_?$ je počet ignorovaných pitch marků v referenční sekvenci. Ignorování některých pitch marků vychází ze skutečnosti, že v některých případech si ani člověk, expert, vyhodnocující pitch marky ručně nebyl jistý, zda se jedná o pitch mark, či nikoliv. Příklad volání skriptu StatisticsPM v konzoli: *./StatisticsPM CZA_oznam00001_00_ref.pm CZA_oznam00001_00_GCI.pm 10*

Parametr *Odchylka* je v tomto případě 10%, a vzhledem k výsledkům v článku [19] vypovídajícím o netečnosti kvality řeči při posunutí pitch marků o 10%, bude používána tato hodnota ve všech porovnáních této práce.

4.3 Výsledky porovnání

Před tím, než mohlo dojít k porovnávání, musely být soubory s referenčními pitch marky převedeny do požadovaného formátu. Dále musely být upraveny funkce tak, aby byl jejich výstup ukládaný do souboru příslušného formátu. Byla proto vytvořena funkce *TT.m*, která toto zajišťuje. Ta pro svou činnost potřebuje znát název souboru s referenčními pitch marky, aby mohlo dojít k vložení hraničních pitch marků typu „T“. To vyžaduje skript StatisticsPM navrhnutý pro porovnání. Pro porovnání byl zvolen Korpus20 viz příloha A.

4.3.1 Prvotní porovnání

Jako první bylo převedeno všech 20 souborů s ručně detekovanými pitch marky do požadovaného formátu. Dále byly spuštěny algoritmy metod *EWM* (podkapitola 4.1.3), *NNPM* (podkapitola 4.1.4) a *SEDREAMS* (podkapitola 4.1.2) pro všech 20 nahrávek s ukládáním do souborů. První výsledky byly poněkud horší, než se očekávalo. Tabulka 4.1 zobrazuje výsledky získané těmito 3 metodami pomocí skriptu *StatisticsPM*.

Metoda	N_t	$N_t - N_?$	N_r	N_D	N_I	N_S	Acc [%]
EWM	10687	10299	10833	549	501	2423	**66,75**
NNPM	10837	10449	10833	390	254	2532	**69,60**
SEDREAMS	11062	10674	10833	1050	645	5280	**34,65**

Tabulka 4.1: Výsledky prvního testu tří vybraných metod

Na výsledcích tohoto prvního testu je zřejmá souvislost ve sloupci N_r, což je v pořádku, protože pro testování byly použity stejné sady referenčních pitch marků. Tyto sady tedy dohromady obsahují 10 833 pitch marků. Výsledky sloupců N_D, N_I a N_S přímo ukazují, kolik pitch marků bylo chybně umístěno pomocí testované metody a o jaký typ chyby se jedná. N_D tedy znamená počet operací “delete”a vyjadřuje tím počet pitch marků, které byly „navíc“. N_I je počet operací “insert”a znamená počet pitch marků, které „chyběly“. N_S je počet operací “shift”a znamená počet pitch marků, které byly „posunuty“ o více než 10% lokální periody.

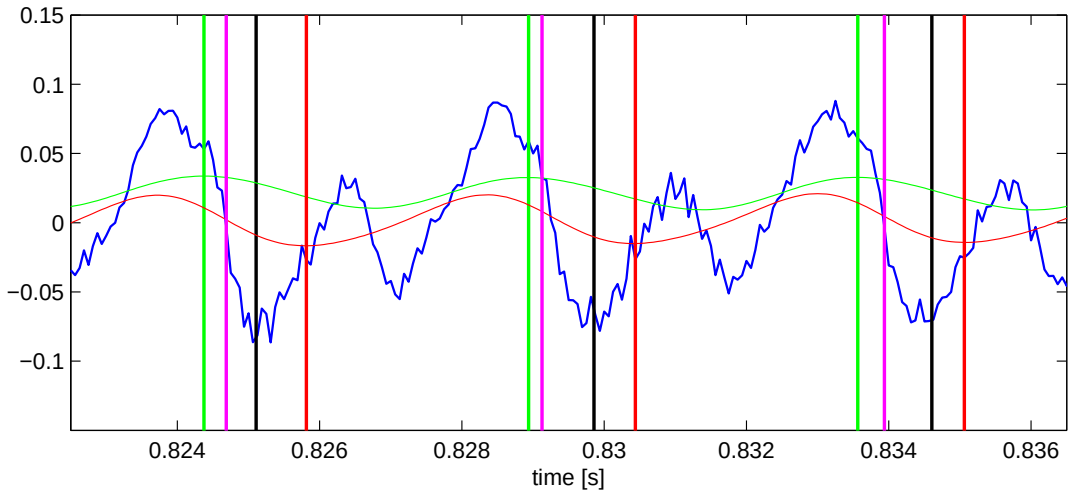
Na výsledcích tohoto testu jsou ve všech třech, ale zvláště v případě metody *SEDREAMS*, zřetelné enormní hodnoty posunů ve sloupci N_S. Pokud se má testovat 10674 pitch marků vzhledem k 10833 referenčním pitch markům, a 5280 jich je posunutých, tak „je něco špatně“. Způsob, jakým jednotlivé metody umisťují pitch marky, je specifický pro každou metodu. Například metoda *NNPM* pitch mark z principu nemůže umístit jinam, než do minima filtrovaného signálu. Takže pokud nekoresponduje minimum filtrovaného signálu s lokálním minimem řečového signálu, téměř určitě nastane chyba typu “shift”.

Navíc v případě metody *SEDREAMS* bylo zjištěno, že má problémy s určováním znělosti. Konkrétně klasifikuje intervaly řeči obsahující ticho jako znělé, což se odráží v počtu N_D. Bylo odhaleno, že v nahrávkách se vyskytuje šumová složka 50 Hz, pravděpodobně indukovaná ze sítě. Tento šum má však odstup více než 40dB k signálu desetiprocentní možné amplitudy. Tento problém byl vyřešen použitím externí informace o (ne)znělosti, jedná se opět o nejlépe vycházející metodu RAPT. Chyby metody *SEDREAMS* ve sloupci N_D tabulky 4.2 se snížily téměř na polovinu. Nadále bude tato metoda používat externí informaci o (ne)znělosti, jelikož se výsledek, byť pouze o 4% zlepšil.

Metoda	N_t	$N_t - N_?$	N_r	N_D	N_I	N_S	Acc [%]
SEDREAMS	10836	10448	10833	579	371	5445	**38,79**

Tabulka 4.2: Výsledky metody *SEDREAMS* s externí informací o (ne)znělosti

Nutností se stalo analyzovat a vyřešit problém s posuny pitch marků všech 3 zmíněných metod. Situaci vystihuje obrázek 4.8.



Obrázek 4.8: Ukázka lokalizace pitch marků 3 výše zmíněných metod, řečový signál (modře), ruční pitch marky (černě), křivka energie a pitch marky v jejích maximech – *EWM* (zeleně), filtrovaný průběh signálu a pitch marky v jeho minimech – *NNPM* (červeně), pitch marky metody *SEDREAMS* (purpurově).

Je zde zobrazen vybraný úsek řečového signálu spolu s ručními pitch marky a s pitch marky metod uvedených v tabulce 4.1. Na první pohled je jasný problém, a sice automatické pitch marky jsou posunuty od ručně určených. Uplatňuje se tím operace "shift"při vyhodnocení přesnosti, a tím vyhodnocená přesnost algoritmů klesá. Okamžik uzavření hlasivek by však měl být pro signály s negativní polaritou v lokálním minimu řečového signálu viz ruční pitch marky.

Bylo empiricky zjištěno, že naprostá většina posunutých automaticky detekovaných pitch marků má určitou tendenci ohledně směru posunu s ohledem na konkrétní metodu. Byla proto navržena jednoduchá procedura, která byla integrována jako poslední „post-processingový" stupeň do tří metod z tabulky 4.1. Jedná se o hledání lokálního minima v lokální periodě řečového a lehce filtrovaného řečového signálu pomocí okénka. Bylo použito Hanningovo okénko. Tato procedura byla vždy nastavena pro vlastnosti posunů konkrétní metody. Například metoda *NNPM* z principu posouvá téměř výhradně doprava. Lokální perioda okolo pitch marku, který se má posunout do lokálního minima řečového signálu, byla převážena doleva posunutým okénkem. Posunutí bylo na základě empirických pokusů nastaveno na 5%, čímž bylo zvýhodněno posunutí doleva. V případě metod *EWM* a *SEDREAMS* bylo nastavení stejné s rozdílem, že bylo zvýhodněno posunutí doprava vyplívající z vlastností těchto dvou metod.

4.3.2 Algoritmus *MPA*

V samotném závěru vyhodnocování výsledků jednotlivých algoritmů zahrnutých do porovnání v kapitole 4 této práce, byla objevena určitá souvislost mezi dvěma signály. Jedná se o *měřený hlasivkový signál EGG* využívaný algoritmem *MPA* a *fundamentální vlnu* prezentovanou Thomasem Ewenderem v článku [7]. Teoreticky nejsou její vlastnosti vhodné, protože její fáze s řečovým signálem se mění v čase. Algoritmus *MPA* však posouvá kandidáty do lokálních minim (pro zápornou polaritu) řečového signálu, což by mohlo problém částečně vyřešit.

Pro simulaci EGG signálu fundamentální vlnou byla zpočátku použita fundamentální vlna stejná jako v případě metody *EWM*. Tato vlna byla generována pomocí spojité kontury F_0 popsané v podkapitole 3.1.4 a byla nulována v neznělých úsecích podle informace o (ne)znělosti získané metodou RAPT, což s sebou nese i nepřesnosti této metody. Tento proces využívající externí informaci o (ne)znělosti byl implementován funkcí *Fundamental.m*. Základní vlna generovaná funkcí *Fundamental.m* bude nadále označována jako *FW1*. Příklad fundamentální vlny zobrazuje podgraf (c) v obrázku 4.3, kde je zároveň zřetelný i případ chyby metody RAPT. V samém začátku podgrafu (a) stejného obrázku je vidět neznělá klasifikace, ale v podgrafu (d) jsou jasně vidět pitch marky. Navíc tím mohou vzniknout nespojitosti, které mohou být problémem zvláště v případech, kdy se počítá diference takového signálu s nespojitostí. To je přesně případ algoritmu *MPA*. Vznikl proto požadavek na generování fundamentální vlny spojité na celém signálu. K tomu byla opět použita spojitá kontura F_0 popsaná v podkapitole 3.1.4, ale již bez informace o (ne)znělosti signálu. Tento proces vytváření spojité fundamentální vlny byl implementován funkcí *Fwave.m*. Spojitá základní vlna generovaná funkcí *Fwave.m* bude nadále označována jako *FW2*.

Pro porovnání bylo použito 20 nahrávek (viz Korpus20 v příloze A), ke kterým byly vygenerovány oba „typy“ fundamentálních vln. Dosažené výsledky jsou obsaženy v tabulce 4.3.

Metoda	N_t	$N_t - N_?$	N_r	N_D	N_I	N_S	Acc [%]
MPA + EGG	11001	10613	10833	293	50	346	**93,51**
MPA + FW1	10678	10301	10611	294	116	348	**92,64**
MPA + FW2	11209	10821	10833	519	70	360	**91,23**

Tabulka 4.3: Porovnání výsledků algoritmu *MPA* spolu s hlasivkovým signálem a simulovanými hlasivkovými signály fundamentální vlnou.

Zde je důležité zdůraznit, že počet referenčních pitch marků metody *MPA + FW1* je nižší. Je tomu tak proto, že nebyla úspěšně vygenerována posloupnost pitch marků algoritmem *MPA* pro jednu z nahrávek. To je pravděpodobně z důvodu větší nespojitosti fundamentální vlny použité místo signálu EGG, který je přirozeně spojitý. O trochu vyšší přesnost této metody je vykoupena nižší robustností, která je zajištěna v metodě *MPA + FW2*.

V rámci budoucí možné práce v tomto směru by mohlo být experimentováno i s použitím průměrovaného signálu definovaného vztahem 4.1 na místo EGG signálu. Způsob výpočtu průměrovaného signálu a fundamentální vlny je různý a je zde proto i rozdíl ve tvaru funkcí. Společný je však fázový posuv a průchody nulou se přesně kryjí. Oba průběhy byly pro zajímavost normalizovány maximální absolutní hodnotou a došlo k porovnání pomocí vzájemné (Pearsonovy) korelace na několika nahrávkách se střední hodnotou výsledných korelací 0,973.

4.3.3 Konečné porovnání

V této fázi již nebyla žádná metoda z podkapitoly 4.3.1 znevýhodněna a mohlo dojít k porovnání. Bylo použito 20 nahrávek (viz Korpus20 v příloze A). Do tohoto porovnání jsou zahrnuty výsledky metod *EWM*, *NNPM*, *SEDREAMS*, *MPA + EGG*, *MPA + FW1*, *MPA + FW2*, *PRAAT-AC* a *PRAAT-CC*. Dosažené výsledky jsou zaznamenány do tabulky 4.4.

Metoda	N_t	$N_t - N_?$	N_r	N_D	N_I	N_S	Acc [%]
MPA + EGG	11001	10613	10833	293	50	346	**93,51**
MPA + FW1	10678	10301	10611	294	116	348	**92,64**
NNPM	10837	10449	10833	390	254	254	**91,41**
MPA + FW2	11209	10821	10833	519	70	360	**91,23**
PRAAT-CC	11268	10880	10833	667	114	317	**89,91**
PRAAT-AC	11174	10786	10833	642	157	311	**89,71**
SEDREAMS	10836	10448	10833	569	359	395	**87,34**
EWM	10687	10299	10833	524	503	370	**86,63**

Tabulka 4.4: Porovnání všech dosažených výsledků seřazených od nejlepšího.

Nahrávky použité pro srovnání obsahovaly i tři známé problematické nahrávky a také dvě vesele mluvené nahrávky. Nejednalo se proto o jednoduchý úkol ani pro algoritmy nástroje PRAAT, který je často považován za referenční na poli analýzy signálu. Zhodnocení:

- *MPA + EGG* – podle očekávání dopadla tato metoda nejlépe, která má jako jediná více informace v podobě EGG signálu.
- *MPA + FW1* – velice překvapivý výsledek. Informace v podobě EGG signálu je nahrazena simulací pomocí fundamentální vlny. V tomto případě se však bohužel nejedná o plně robustní verzi, protože fundamentální vlna obsahuje nespojitosti.
- *NNPM* – velmi dobrý výsledek. V tomto případě je však nutné přiznat, že 20 nahrávek použitých pro vyhodnocení těchto výsledků bylo ze $\frac{3}{4}$ zainteresovaných do trénování této neuronové sítě. Další vyhodnocení výsledků bude uvedeno v příští podkapitole.
- *MPA + FW2* – Výsledek ztrácející pouze 2,27 %. V tomto případě jde již o plně robustní verzi. Jedná se o prokazatelně nejlepší dosažený výsledek bez EGG signálu této práce, jelikož metoda *NNPM* je ovlivněna trénovacími daty ve vyhodnocení a metoda *MPA + FW1* není robustní kvůli nespojitostem, které se v této verzi nevyskytují.
- *PRAAT* – stabilně dobrý výsledek v případech obou metod získání kontury F_0. Metoda *PRAAT-CC* vykazuje 0,2 % lepší výsledek než PRAAT-AC.
- *SEDREAMS* – následuje tato metoda, která celkově na testovaných nahrávkách neměla oslnivé výsledky. Bez použití externí informace o (ne)znělosti by byl výsledek s největší pravděpodobností horší.

- *EWM* – na pomyslném „chvostu“ tohoto srovnání. To je možné zdůvodnit hlavně tím, že předpoklad této metody se ve skutečnosti neuplatňuje do takové míry, jak se původně myslelo. Jedná se o předpoklad, že pitch marky se nacházejí v lokálních maximech průběhu krátkodobé energie. Tento průběh vykazuje různé speciální případy v závislosti na tvaru řečového signálu, takže nelze vždy zcela robustně detekovat pitch marky. Avšak ani tento výsledek není nijak zvláště špatný, jelikož z 20 nahrávek jich je 5 „náročnějších“.

4.3.4 Dodatečné porovnání metody *NNPM*

V předchozí podkapitole bylo přiznáno zvýhodnění neuronové sítě (metody *NNPM*) v rámci porovnání. Data, která byla použita pro srovnání, byla používána také k trénování této sítě. Výsledek metody *NNPM* je v tomto případě vhodné označit za zaujatý. Není tedy k dispozici zcela věrohodný výsledek přesnosti metody *NNPM*. Všech 83 nahrávek (viz Korpus83 v příloze A) bylo použito k trénování neuronové sítě. Pouze k těmto byly k dispozici ručně detekované pitch marky. Byl proto proveden test na 20 nahrávkách, které nebyly použity pro trénování neuronové sítě. Jako referenční budou použity pitch marky nalezené metodou *MPA + EGG*, která vychází i z hlasivkového signálu a jejíž výsledky se v předchozí podkapitole podle očekávání ukázaly jako nejlepší. Výsledky dodatečného porovnání jsou zaznamenány v tabulce 4.5.

Metoda	N_t	$N_t - N_?$	N_r	N_D	N_I	N_S	Acc [%]
NNPM	12074	12074	11745	491	162	302	**92,09**
PRAAT-CC	12380	12380	11745	855	220	352	**88,47**
SEDREAMS	11883	11883	11745	683	545	723	**83,58**
EWM	12055	11785	11745	482	442	1051	**83,24**

Tabulka 4.5: Výsledky *NNPM* a dalších 3 vybraných metod vzhledem k výsledkům metody *MPA + EGG* na datech, které nebyly použity pro trénování sítě použité v metodě *NNPM*.

Tyto výsledky byly získány nikoliv vzhledem k ručně detekovaným pitch markům, ale vzhledem k pitch markům detekovaným algoritmem *MPA + EGG*. Mají proto do určité míry jen orientační charakter. Výsledky v sobě totiž zahrnují určitou informaci o tom, jak moc jsou si podobné vlastnosti uvedených pitch markovacích algoritmů s vlastnostmi algoritmu *MPA + EGG*. Ten však v minulé podkapitole prokázal nejlepší výsledky. Dosažený výsledek metody *NNPM* proto funkčnost neuronové sítě potvrdil. Ve prospěch věrohodnosti výsledků metody *NNPM* je také obrázek 4.5, který vypovídá o podobnosti chyby neuronové sítě na testovacích a trénovacích datech. To znamená, že neuronová síť metody *NNPM* neklasifikuje výrazně hůře na neznámých datech. Zdá se, že natrénování neuronové sítě je v pořádku a síť neztratila schopnost „zobecnění“.

Kapitola 5

Závěr

Téma této práce bylo stručně nastíněno v kapitole 1. Byly zde uvedeny motivace a cíle této práce, ale také stručný popis jednotlivých kapitol.

V kapitole 2 bylo vysvětleno množství odborných názvů, termínů, příslušné teorie, ale také metod zpracování, které jsou používány v následujících kapitolách 3 a 4, které v případě potřeby odkazují na příslušný termín uvedený v teoretickém základu práce. V první řadě se jedná o teorii týkající se přirozené produkce řeči a její digitalizaci. Následující část vypovídá o metodách zpracování řečového signálu. Podrobněji byly vysvětleny způsoby určování kontury F_0 a v neposlední řadě se zde vyskytuje dostatečné množství teorie o syntéze řeči. Na konci této kapitoly byl stručně vysvětlen princip neuronových sítí.

V kapitole 3 bylo popsáno několik vybraných přístupů pro výpočet základního hlasivkového tónu a bylo zde provedeno porovnání jejich přesnosti a vlastností pomocí uvedených postupů. V závěru kapitoly 3 byly prezentovány dosažené výsledky.

Kapitola 4 se zabývala porovnáním algoritmů provádějících detekci hlasivkových pulzů. Několik vybraných přístupů bylo prezentováno, některé byly implementovány, a bylo provedeno srovnání jejich přesnosti a vlastností. Dosažené výsledky tohoto srovnání byly uvedeny v závěru kapitoly 4.

Hlavním cílem této práce bylo porovnání algoritmů pro automatickou detekci hlasivkových pulzů v řečovém signálu. Úloha detekce hlasivkových pulzů, nebo též pitch marků, je vyžadována řadou metod a přístupů týkajících se syntézy řeči a automatického rozpoznávání řeči. Existuje množství algoritmů, které se touto úlohou zabývají. Nejlepší výsledky vykazují ty algoritmy, které využívají kromě řečového signálu i hlasivkový EGG signál. Snahou při plnění této práce bylo nalézt způsob, jak se s úspěšností detekce co nejvíce přiblížit pomyslnému „stropu“ v podobě algoritmu *MPA* (viz kapitola 4.1.5), který EGG signál využívá. Pokud by se podařilo dosáhnout jeho úspěšnosti detekce bez použití EGG signálu, odpadla by tím nutnost používat elektroglotograf při nahrávání korpusů v budoucnu a bylo by tím také umožněno použití korpusů nahraných bez EGG signálu v minulosti.

Některé algoritmy, které nepracují s EGG signálem, vyžadují alespoň informaci o průběhu základního hlasivkového tónu. Ukázalo se, že jejich úspěšnost je na kvalitě takového průběhu silně závislá. Proto bylo v této práci zahrnuto také srovnání dostupných nástrojů, které řeší úlohu detekce základního hlasivkového tónu.

Pro porovnání algoritmů výpočtu základního hlasivkového tónu byly vytvořeny referenční kontury z ručně detekovaných pitch marků. Dosažené výsledky kapitoly 3 určují metodu spojité kontury F_0 prezentované v článku [10] jako nejlepší. Vzhledem k tomu, že je spojitá, musela

být zahrnuta informace o (ne)znělosti z jiné metody. V tomto případě byla použita informace o (ne)znělosti z metody RAPT, protože její výsledky byly v obou kritériích na druhém místě. Dosažený výsledek Ewenderovy spojité kontury byl v případě kritéria RMSE o 25,6 % nižší než vykazovala metoda RAPT. V případě korelace byla Ewenderova spojitá kontura o 1,2 % lepší než druhá nejlepší RAPT.

Pro porovnání algoritmů detekce pitch marků v řečových signálech byly ručně detekované pitch marky použity jako referenční. Mezi algoritmy pro detekci pitch marků v řečových signálech se podle očekávání ukázal jako nejlepší algoritmus *MPA* (pracující s EGG signálem) s úspěšností 93,51 %. Byl proveden experiment, kdy byl nahrazen EGG signál v algoritmu *MPA* fundamentální vlnou prezentovanou v článku [7]. Spojitá fundamentální vlna (FW2) byla použita místo EGG signálu v algoritmu *MPA* s výsledkem jen o 2,27 % horším než s originálním EGG signálem. Neuronová síť pro detekci pitch marků (*NNPM*) měla sice přibližně o dvě desetiny procenta lepší výsledky, ale data použitá pro určení těchto výsledků byla částečně použita pro její trénování. Na nezaujatých datech však metoda *NNPM* prokázala svou velmi dobrou přesnost detekce (viz kapitola 4.3.4). O 3,6 % horší výsledek, než algoritmus *MPA*, vykazuje lepší z metod nástroje PRAAT.

Pokud z nějakého důvodu není k dispozici EGG signál ke zvuku, ve kterém mají být nalezeny pitch marky, nejlepší výsledek se podle dosažených přesností získá simulací EGG signálu fundamentální vlnou nebo použitím neuronové sítě. Pokud EGG k dispozici je, není důvod jej nevyužít a nejlepší výsledek bude dosažen algoritmem *MPA*.

Náměty na další práci

Kapitola 3 této práce pojednává o porovnání algoritmů výpočtu základního hlasivkového tónu. Algoritmy jsou porovnány ve smyslu RMSE a pomocí vzájemné korelace ve znělých úsecích řeči. Mohl by navíc být navržen způsob porovnání korektnosti klasifikace (ne)znělosti, tedy jak přesně jsou detekovány začátky a konce znělých intervalů řečového signálu. Navrženým způsobem by pak mohly být všechny algoritmy porovnány. Do tohoto porovnání by mohly být zahrnuty i klasifikátory znělosti/neznělosti realizované neuronovými sítěmi v článcích [7] a [18]. Získáním dostatečně přesné klasifikace (ne)znělosti by mohla být tato informace integrována do fundamentální vlny *FW2*. Tím by velice pravděpodobně došlo ke zvýšení přesnosti metody *MPA + FW2* využívající právě signálu *FW2* namísto hlasivkového EGG signálu.

Dalším námětem by mohlo být provedení testů algoritmu *MPA* také s průměrovaným signálem uvedeným v podkapitole 4.1.2 místo EGG signálu. Průměrovaný signál vykazuje podobné vlastnosti jako fundamentální vlna. Alternativně by mohly být zvoleny nebo navrženy další metody filtrace. Tato cesta simulace EGG signálu se ukázala jako poměrně úspěšná.

V podkapitole 4.1.6 jsou zmíněny další dvě metody detekce pitch marků v řečových signálech, které by mohly být zahrnuty do porovnání kapitoly 4. Tyto dvě metody byly prezentovány v článcích [17] a [18].

Neuronová síť v metodě *NNPM* (viz podkapitola 4.1.4) funguje na základě dolnopropustní filtrace se zlomovou frekvencí 700 Hz. Základní hlasivkový tón jednoho řečníka se mění v průběhu jedné promluvy. Navíc různí řečníci mají různý hlasivkový tón. Zde by mohla být místo uvedené filtrace použita filtrace měnící vlastnosti podle základního hlasivkového tónu.

Literatura

[1] PSUTKA, J. – MÜLLER, L. – MATOUŠEK, J. – RADOVÁ, V. *Mluvíme s počítačem česky (Talking with Computer in Czech)*. Praha : Nakladatelství Academia, 2006.

[2] Hlasivky. *Domovské stránky uživatelů systému Orion* [online]. Pavlína Heiderová 2011. [citováno: březen 2012]. Dostupné na: <http://home.zcu.cz/~heidpa/semestralka/Hlasivky.gif>

[3] PSUTKA, J. *Komunikace s počítačem mluvenou řečí*. Praha : Nakladatelství Academia, 1995.

[4] HORÁK, A. Roviny analýzy jazyka, Fonetika. *Úvod do počítačové lingvistiky* [online]. 2012. [citováno: březen 2012]. Dostupné na: <http://nlp.fi.muni.cz/poc_lingv/pl02.pdf>

[5] ŠLAPÁK, I. – JANEČEK, D. – LAVIČKA, L. *Základy otorinolaryngologie a foniatrie* [online]. 2009. [citováno: březen 2012]. Dostupné na: <http://is.muni.cz/elportal/estud/pedf/js09/orl/web/doc/zaklady-orl-a-foniatrie.pdf>

[6] Signal Processing Toolbox documentation. *Hamming window* a *Hanning window*. [online]. [citováno: duben 2012]. Dostupné na: <http://www.mathworks.com/help/toolbox/signal/ref/hamming.html> a <http://www.mathworks.com/help/toolbox/signal/ref/hann.html>

[7] EWENDER, T. – PFISTER, B. Accurate Pitch Marking for Prosodic Modification of Speech Segments. In: *Proceedings of Interspeech 2010*. Makuhari, Chiba, Japan, 2010, pp. 178-181.

[8] GHULAM, M. Extended Average Magnitude Difference Function Based Pitch Detection. *The International Arab Journal of Information Technology*. Vol. 8, No. 2, April 2011, pp. 197-203.

[9] ŠMÍD, R. Metody analýzy signálu. *Diagnostické systémy* [online]. 2008. [citováno: duben 2012]. Dostupné na: <http://measure.feld.cvut.cz/usr/staff/smid/lectures/sigproc07.pdf>

[10] EWENDER, T. – HOFFMAN, S. – PFISTER, B. Nearly Perfect Detection of Continous F_0 Contour and Frame Classification for TTS Synthesis. In: *Proceedings of Interspeech 2009*. Brighton, UK, 2009, pp. 100-103.

[11] RADOVÁ, V. Nelineární sítě. *Neuronové sítě – KKY/NEU*. 2011/2012. Západočeská univerzita v Plzni. Fakulta aplikovaných věd. Katedra kybernetiky, 2011.

[12] Forex-neural. *BJF Trading Group Inc.* [online]. BJF Trading Group Inc. 2010. [citováno: březen 2012]. Dostupné na: <http://iticsoftware.com/image/data/forex-neural.jpg>

[13] DRUGMAN, T. *Advances in Glottal Analysis and its Applications*. PhD Thesis. University of Mons, Faculty of Engineering, Mons, 2011.

[14] TALKIN, D. A Robust Algorithm for Pitch Tracking (RAPT). In: KLEIJN, W. B. – Paliwal, K. K., eds. *Speech Coding and Synthesis*. New York: Elsevier Science B.V., 1995, chapter 14, pp. 495-518.

[15] LEGÁT, M. – MATOUŠEK, J. – TIHELKA, D. On the Detection of Pitch Marks Using a Robust Multi-Phase Algorithm. *Speech Communication*. 2011, Vol. 53, No. 4, pp. 552-566.

[16] BARNARD, E. – COLE, R. A. – VEA, M. P. – ALLEVA, F. A. Pitch detection with a neural-net classifier. *IEEE Transactions on Signal Processing*. February 1991, Vol. 39, No. 2, pp. 298-307.

[17] CHENG-YUAN, L. – JYH-SHING, R. J. A Two-Phase Pitch Marking Method for TD-PSOLA Synthesis. In: *Proceedings of Interspeech 2004*. Jeju, Korea, 2004, pp. 1189-1192.

[18] KOTNIK, B. – HÖGE, H. – KAČIČ, Z. Noise robust F_0 determination and epoch-marking algorithms. *Signal Processing 89 (2009)*. April 2009, pp. 2555-2569.

[19] MOULINES, E. – CHARPENTIER, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*. December 1990, Vol. 9(5-6), pp. 453-467.

Příloha A

Korpus83, Korpus20

Korpus 83 studiových nahrávek, které byly spolu s korespondujícími stejnojmennými soubory obsahujícími ručně detekované pitch marky poskytnuty pracovištěm KKY pro tuto práci.

CZA – čeština, muž	*CZF – čeština, žena*	*CZM – čeština, muž*
CZA_oznam00001_00.wav	**CZF_Sentence05025_000.wav**	**CZM_Sentence10000.wav**
CZA_oznam00002_00.wav	CZF_Sentence05026_002.wav	CZM_Sentence10001.wav
CZA_oznam00003_00.wav	CZF_Sentence05027_000.wav	CZM_Sentence10002.wav
CZA_oznam00004_00.wav	CZF_Sentence05028_000.wav	CZM_Sentence10003.wav
CZA_oznam00005_00.wav	CZF_Sentence05029_000.wav	CZM_Sentence10004.wav
CZA_oznam09996_00.wav	**CZF_sentence0006.wav**	**CZM_sentence00006.wav**
CZA_oznam09997_00.wav	CZF_sentence0007.wav	CZM_sentence00007.wav
CZA_oznam09998_00.wav	CZF_sentence0008.wav	CZM_sentence00008.wav
CZA_oznam09999_00.wav	CZF_sentence0009.wav	CZM_sentence00009.wav
CZA_oznam10000_00.wav	CZF_sentence0010.wav	CZM_sentence00010.wav
EN – angličtina, muž	*FR – francouzština, žena*	*DE – němčina, muž*
EN_Sentence0001.wav	**FR_Sentence0001.wav**	**DE_Sentence0001.wav**
EN_Sentence0002.wav	FR_Sentence0002.wav	DE_Sentence0002.wav
EN_Sentence0003.wav	FR_Sentence0003.wav	DE_Sentence0003.wav
EN_Sentence0004.wav	FR_Sentence0004.wav	DE_Sentence0004.wav
EN_Sentence0005.wav	FR_Sentence0005.wav	DE_Sentence0005.wav
EN_Sentence6697.wav	**FR_Sentence7982.wa**v	**DE_Sentence5554.wav**
EN_Sentence6698.wav	FR_Sentence7983.wav	DE_Sentence5555.wav
EN_Sentence6699.wav	FR_Sentence7984.wav	DE_Sentence5556.wav
EN_Sentence6700.wav	FR_Sentence7985.wav	DE_Sentence5557.wav
EN_Sentence6701.wav	FR_Sentence7986.wav	DE_Sentence5558.wav
SK – slovenština, žena	*HA – „happy“, čeština, žena*	*TT – „problematic“, čeština, žena*
SK_sentence0001.wav	**HA_Sentence00003_000.wav**	**TT_Sentence02590_000.wav**
SK_sentence0002.wav	HA_Sentence00007_000.wav	**TT_Sentence03251_000.wav**
SK_sentence0003.wav	HA_Sentence00020_000.wav	**TT_Sentence04059_000.wav**
SK_sentence0004.wav	HA_Sentence00021_000.wav	
SK_sentence0005.wav	HA_Sentence00022_000.wav	
SK_sentence7008.wav	**HA_Sentence00037_000.wav**	
SK_sentence7009.wav	HA_Sentence00047_000.wav	
SK_sentence7010.wav	HA_Sentence00050_000.wav	
SK_sentence7011.wav	HA_Sentence00062_000.wav	
SK_sentence7012.wav	HA_Sentence00063_000.wav	

Tučně vyznačené nahrávky byly vybrány pro vyhodnocení některých výsledků. Dále budou označovány jako **„Korpus20“**. Všechny nahrávky jsou ve formátu mono 16-bit ".wav"souborů. Celkem se jedná o více než 11 minut záznamu s více než 50 tisíci ručně určených pitch marků. Nahrávky typu "happy"a "problematic"znamenají těžší podmínky pro analyzující algoritmy.

Příloha B

KorpusNN

Korpus 20 studiových nahrávek různých od těch v příloze A, ke kterým byly algoritmem *MPA + EGG* vyhodnoceny pitch marky, které byly uloženy ve formě stejnojmenných souborů. Všechny nahrávky jsou v češtině.

Řečník AJ – muž	*Řečník JS – muž*	*Řečník KI – žena*	*Řečník MR – muž*
oznam00001_00.wav	oznam00002_00.wav	oznam00003_00.wav	oznam00004_00.wav
oznam00005_00.wav	oznam00006_00.wav	oznam00007_00.wav	oznam00008_00.wav
oznam00009_00.wav	oznam00010_00.wav	oznam00011_00.wav	oznam00012_00.wav
oznam00013_00.wav	oznam00014_00.wav	oznam00015_00.wav	oznam00016_00.wav
oznam00017_00.wav	oznam00018_00.wav	oznam00019_00.wav	oznam00020_00.wav