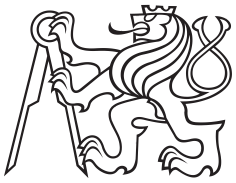
Bakalářská práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra radioelektroniky

Analýza poruch hlasu u Parkinsonovy nemoci pomocí základní hlasivkové frekvence

Vojtěch Illner

Vedoucí: Prof. Ing. Pavel Sovka, CSc.
Obor: Otevřené elektronické systémy
Květen 2018

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Illner** Jméno: **Vojtěch** Osobní číslo: **457432**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra radioelektroniky**

Studijní program: **Otevřené elektronické systémy**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Analýza poruch hlasu u Parkinsonovy nemoci pomocí základní hlasivkové frekvence

Název bakalářské práce anglicky:

Analysis of Voice Disorders in Parkinson's Disease using Pitch Frequency

Pokyny pro vypracování:

1. Seznamte se s možnostmi potenciálního využití hodnocení základní hlasivkové frekvence pro analýzu poruch hlasu u Parkinsonovy nemoci. Na základě literatury nastudujte a vyberte vhodné algoritmy pro detekci základní hlasivkové frekvence u patologické řeči.
2. Vybrané typy detektorů základní hlasivkové frekvence otestujte na vzorku řečových promluv zdravé populace a pacientů s RBD pořízených pomocí smartphone, s přidáním různých typů šumů představujících reálné prostředí.
3. Vyhodnoťte výkon dostupných detektorů základní hlasivkové frekvence pomocí jednoduchých statistických testů.

Seznam doporučené literatury:

1. Tsanas A, Zanartu M, Little MA, McSharry PE. Robust fundamental frequency estimation in sustained vowels: detailed algoritmic comparisons and information fusion with adaptive Kalman filtering. Journal of the Acoustical Society of America 2014; 135: 2885-2901.
2. Yang N, Ba H, Cai W, Demirkol I, Heinzelman W. BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. IEEE/ACM Trancactions on Audio, Speech and Language Processing 2014; 22: 1833-1848.
3. Rusz J, Hlavnicka J, Tykalova T, Buskova J, Ulmanova O, Ruzicka E, Sonka K. Quantitative assessment of motor speech abnormalities in idiopathic REM sleep behaviour disorder. Sleep Medicine 2016;19:141-147.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

prof. Ing. Pavel Sovka, CSc., katedra teorie obvodů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **05.02.2018** Termín odevzdání bakalářské práce: _____

Platnost zadání bakalářské práce: **30.09.2019**

prof. Ing. Pavel Sovka, CSc. podpis vedoucí(ho) práce	podpis vedoucí(ho) ústavu/katedry	prof. Ing. Pavel Ripka, CSc. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

. _____ _____

Datum převzetí zadání Podpis studenta

Poděkování

Poděkování patří především Prof. Ing. Pavlu Sovkovi, CSc., za skvělé vedení této práce, perfektní spolupráci a poskytnuté odborné znalosti. Rád bych také poděkoval Ing. Janu Ruszovi, PhD., za výpomoc a poskytnutí užitečných materiálů. Závěrem děkuji všem, kteří mi motivačně pomáhali při psaní práce, i v celém bakalářském studiu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 23. května 2018

Abstrakt

V této práci se zabýváme hodnocením a testováním estimátorů základní hlasivkové frekvence F_0 na delších plynulých řečových promluvách, nahraných na chytrý telefon. Do nahrávek byl následně přidán šum na určitých hladinách SNR (signal-to-noise ratio) jako simulace reálného prostředí.

Tento přístup byl zvolen z hlediska zamýšlené aplikace, kdy by se pomocí analýzy trendů vývoje F_0 objektivně určovala kvalita řeči a s tím spojený monitoring na přítomnost neurodegenerativních chorob, zvláště Parkinsonovy nemoci. Diagnostika by přitom byla provedena pomocí chytrého telefonu v běžném prostředí.

Z dostupných algoritmů byl po testování vybrán estimátor F_0 SWIPE, jehož účinnost a přesnost vyhovuje vymezeným podmínkám. Pomocí něj je možné F_0 poměrně přesně sledovat, i pro těžké podmínky na nižších hodnotách SNR pro dlouhé řečové monology.

Tento způsob může nalézt své uplatnění v klinické praxi na brzkou diagnostiku Parkinsonovy choroby, sledování efektů případné léčby nebo detailní monitorování progrese nemoci. Veškerá analýza přitom může probíhat pomocí chytrých telefonů pacientů i v reálném prostředí za přítomnosti šumu na pozadí.

Klíčová slova: Parkinsonova nemoc, základní hlasivková frekvence, smartphone

Vedoucí: Prof. Ing. Pavel Sovka, CSc.

Abstract

This thesis is focused on evaluation of the performance of various PDAs (pitch detection algorithms) on recordings of the connected speech obtained via smartphones. An additive noise was added to the records at certain SNR (signal-to-noise ratio) levels as a simulation of everyday environment.

This approach was chosen with respect to a possible application of evaluating the speech quality and associated monitoring of presence of neurodegenerative diseases particularly Parkinson's disease (PD). Analysis of recordings obtained via smartphone in common environment would have potential to provide useful markers for diagnosis of PD.

From tested pitch-trackers, the SWIPE algorithm was found to be most reliable. Using the SWIPE, it is possible to very precisely track fundamental frequency on long speech utterances even in tough conditions at lower SNR levels of long connected speeches.

This approach can have a wide range of use for early diagnosis of Parkinson's disease, evaluation following therapy efficiency and detailed monitoring of the disease progress. All the analysis can be performed via smartphones even in a real environment in the presence of a background noise.

Keywords: Parkinson's disease, vocal chords fundamental frequency (pitch), smartphone

Title translation: Analysis of Voice Disorders in Parkinson's Disease using Pitch Frequency

Obsah

Obrázky

Tabulky

Kapitola 1

Úvod

Neurodegenerativní choroby představují jedny z nejčastějších onemocnění postihujících převážně starší populaci. Jeden z klíčových aspektů zahájení úspěšné léčby nemoci, je její brzká diagnóza, tedy ještě ve stadiu kdy nejsou přítomny zjevné motorické příznaky [13].

Porucha řeči je jeden z nejtypičtějších ukazatelů přítomnosti choroby, detekovatelný již ve velmi brzkých stádiích zhoubného rozvoje [4]. Objektivní sledování řečové kvality v čase může v budoucnosti přispět k brzké diagnostice onemocnění až několik let před vypuknutím prvních motorických příznaků [15].

Objektivnost sledování kvality řeči vyžaduje velký počet promluv, „vzorků“, pro daného pacienta přes určité sledované období. Pořizovaní těchto promluv vyžaduje mj. pravidelné návštěvy lékaře a s tím spojený náročný logistický proces. V současné době, díky nástupu chytrých telefonů a zvýšení výpočetního výkonu, je velmi silný trend tento proces zefektivnit. Pacient by mohl své promluvy nahrávat sám a řečová analýza, spojená s případnou diagnostikou, by následně probíhala na klinickém serveru.

Tento způsob však přináší některé technické problémy. Nahrávání promluvy v ordinaci probíhá v tichu při použití kvalitního mikrofonu. Při nahrávání na chytrý telefon se ocitáme v reálném prostředí, které vnáší problém přítomnosti šumu. Také je využit mikrofon telefonu, který nemusí být zdaleka tak kvalitní.

Studie [20, 21, 22, 23] ukazují, že i při těchto problémech je tento způsob realizovatelný a objektivní, pokud jsou zkoumány ukazatele řečové kvality, které jsou robustní i pro tento případ. Mezi nejvěrohodnější takové ukazatele patří časový průběh základní hlasivkové frekvence F_0 během promluvy, konkrétně její směrodatná odchylka.

Tato práce se zabývá analýzou řečových promluv pořízených pomocí chytrého telefonu za přidání šumů, simulující reálné prostředí. Zkoumaný objektivní parametr kvality řeči je přitom základní hlasivková frekvence F_0.

1.1 Lidská řeč

Řeč je nepřirozenější metodou mezilidské komunikace a dorozumívání. Představuje ale také velmi zajímavý výzkumný objekt. Díky komplexnosti její produkce můžeme řečovou analýzou určovat kvalitu řeči, ale také například odhadovat emocionální rozpoložení řečníka [12].

Analýza kvality promluvy nám přináší přesné a objektivní výsledky, kdy z charakteristik řečových poruch můžeme určovat přítomnost a závažnost neurodegenerativní choroby [5]. V této práci se budeme věnovat základní hlasivkové frekvenci F_0 jako ukazatele kvality promluvy.

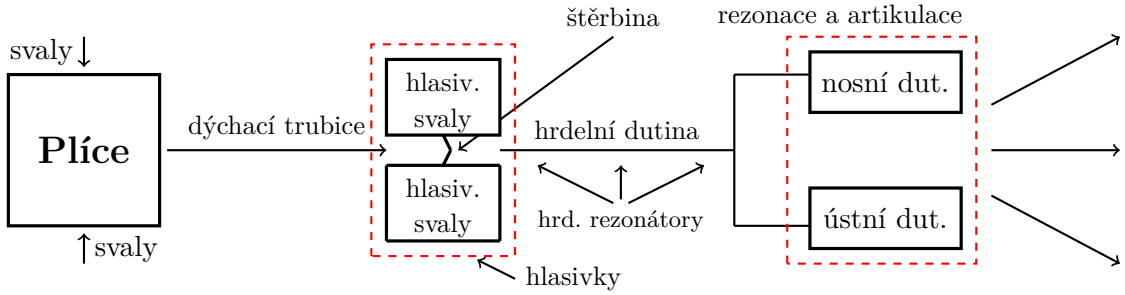
1.1.1 Fyziologie

Řeč je zvuková vlna, vytvořená složitou soustavou vokálního traktu lidského těla. Proces vytváření zvuku je složen ze tří hlavních částí

- Tvorba tlaku vzduchu
- Regulace vibrací
- Kontrola rezonance

Proud vzduchu přichází z plic a prochází *hlasivkovou štěrbinou*, která je obklopena hlasivkovými svaly. Díky průchodu vzduchu štěrbinou dochází k rozkmitání hlasivek, kontrolovaného soustavou menších svalů okolo. Vzduchový proud se tak mění na přerušované pulzy, zvuk připomínající bzučení. Ve vokálním traktu za hlasivkami pak dochází k úzkému zúžení, které z průchodu vzduchu generuje další zdrojové zvuky, šířící se společně se samotným zvukem z hlasivek. V horní části traktu se pak nacházejí *rezonátory* ve formě dutiny hrdelní, nosní a ústní. Ty mění přicházející zvuk na již konkrétnější podobu a také ho zesilují. Výsledný tvar pak utváří *artikulátory*, mezi hlavní patří jazyk, spodní čelist a rty [6]. Schematický nákres procesu tvorby řeči je na obrázku 1.1.

Při produkci neznělých hlásek prochází vzduch hrdelní dutinou s větší rychlostí a v zúženích vznikají vzduchové turbulence, deformující původní tvar. Tento výsledný zvuk již přestává být funkcí frekvence hlasivek a bývá většinou modelován Gaussovským šumem [7].



Obrázek 1.1: Schématický nákres tvorby řeči v lidském těle.

1.1.2 Základní hlasivková frekvence F_0

Průchod proudu vzduchu hlasivkovou štěrbinou způsobuje její rozkmitání, kdy se opakovaně otevírá a zavírá. Tento proces je ovlivňován hlavně samotným proudem vzduchu a hlasivkovými svaly. Působením oscilací vzniká v ideálním případě periodický (impulzový) signál $U(t)$, zobrazený na obrázku 1.2. Pokud bychom zjistili časovou vzdálenost T_0 dvou následujících cyklů, byla by základní frekvence hlasivek rovna inverzi z této hodnoty

$$F_0 = \frac{1}{T_0}. \tag{1.1}$$

Bohužel, díky mnoha fyziologickým aspektům hlasivek a lidského vokálního traktu nemá $U(t)$ nikdy takto ideální tvar. Signál se stává kvaziperiodickým a nestacionárním, měnící svoje vlastnosti a průběh v čase. Díky tomu přestává být jasné, kde přesně brát hodnotu periody T_0. Existuje několik definic F_0 [6] podle způsobu, kde se určuje délka periody, konkrétně její začátek.

Tyto definice by v případě perfektně periodického a stacionárního signálu vedly na stejné výsledky. V reálném případě se však liší a každý estimátor F_0 pracuje s takovou definicí, která nejlépe pasuje na metodu, kterou používá.

Výška hlasu. Psychoakustická [6] veličina, která je přímo spojená se základní frekvencí F_0 je hlasová výška („pitch“). Popisuje jak je lidské zvukové ústrojí schopno vnímat výšku tónu přicházejícího zvuku [6]. Časový průběh F_0 a jeho případné změny člověk vnímá skrze tuto veličinu. Pojem „pitch“ a „fundamental frequency“ bývají často zaměňovány, jedná se ale o rozdílené parametry. Díky vlastnostem sluchového ústrojí je vnímaná výška hlasu nelineární funkcí frekvence.

Způsobů jak změnit měřítko frekvenční závislosti aby vyhovovalo lidskému sluchovému ústrojí, bylo vyvinuto několik. Mezi nejznámější patří *semitóny*, *mel scale* [6] a *bark scale* [24]. V této práci se také budeme zabývat později zavedeným *ERB scale* [9].

Semitóny. Jedním z velmi využívaných postupů, který uplatníme i v této práci, je převedení časového průběhu F_0 během promluvy z Hertzů do semitónů, tedy logaritmické tónové škály. Hlavní důvod je především rozdílnost ve velikosti hlasivkové frekvence mezi pohlavími, kdy mají ženy většinou mnohem vyšší výšku hlasu než nízko posazený mužský hlas a s tím spojené jiné rozdělení F_0 při řeči. Rozdíl také může být v jazycích jednotlivých národů. Převedením do semitónů se rozdělení stávají přibližně stejná [3].

Pokud si označíme s jako poměr dvou frekvencí, vyjadřující *jeden* semitón, bude pro počet semitónů n v rozmezí dvou frekvencí f_1 a f_2, $f_1 > f_2$, platit rovnost

$$\frac{f_1}{f_2} = s^n. \tag{1.2}$$

Z této rovnice si vyjádříme n,

$$n = \frac{\log_{10}\left(\frac{f_1}{f_2}\right)}{\log_{10}(s)}, \tag{1.3}$$

kdy byl zvolen logaritmus o základu 10. s určíme ze znalosti, že semitón je jedna dvanáctina oktávy, která reprezentuje frekvenční poměr 2. Čili $2 = s^{12}$, neboli $s = \sqrt[12]{2}$.

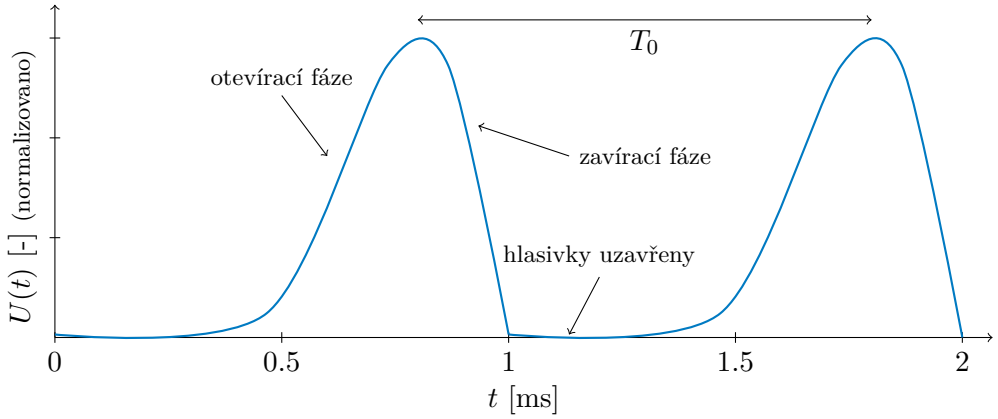
V našem případě nebudeme určovat počet semitónů mezi dvěma frekvencemi, nýbrž převádět hodnotu *jedné* frekvence do semitónové škály. Toto provedeme zvolením f_2 za konstantní. Konkrétní hodnotu určíme dostatečně nízkou abychom nemohli dostávat záporná čísla. Jako ideální taková hodnota bylo zvoleno 60 Hz. Pokud tyto poznatky zkombinujeme, dostáváme finální vzorec pro převod.

$$f^{\text{sem}} = 12\frac{\log_{10}\left(\frac{f^{\text{Hz}}}{60}\right)}{\log_{10}(2)}, \tag{1.4}$$

kde f^{sem} je výsledná frekvence v semitónech a f^{Hz} je původní frekvence v Hertzích.

Obtížnost estimace F_0. V průběhu let bylo vyvinuto několik stovek různých estimátorů F_0 (PDAs, pitch determination algorithms) a žádný z nich se nedá použít univerzálně na všechny situace [7]. Estimace F_0 se považuje za jeden z nejtěžších problémů řečové analýzy. Důvodů, proč tomu tak je, je vícero [7]

- Řeč je nestacionární proces. Stav vokálního traktu se může rapidně měnit v čase. F_0 se tak často mění i pro každou periodu hlasivek. V některých případech může hodnota skočit o celou oktávu.
- Je nutné rozlišovat mezi produkcí znělých a neznělých hlásek, kdy je zkoumaný signál deformován vzduchovými turbulencemi. Segmenty řeči se přitom pohybují po celé škále od "úplně znělé"po "úplně neznělé". Čistě znělé úseky přitom mohou trvat jen několik hlasivkových period.



Obrázek 1.2: Průchod proudu vzduchu $U(t)$ hlasivkovou štěrbinou.

- Ve výsledné řeči se často vyskytují subharmonické násobky „pravé“ F_0, některé dokonce s větší amplitudou kvůli rezonanci v dutinách.
- Průchod signálu vokálním traktem může do velké míry pozměnit tvar periodicity.
- Řečový signál má velmi širokou škálu amplitud, od nízkých při pauzách po vysoké pro některé znělé hlásky.
- Šum na pozadí může výrazně ovlivňovat výsledky, jelikož je ho lehké splést za určité segmenty řeči. Obecně je řečový signál extrémně citlivý na jakékoliv deformace.
- Hodnota F_0 se od člověka k člověku může velmi lišit na škále až přes čtyři oktávy (50-800 Hz).

Více jsou estimátory F_0 rozebrány v sekci 1.2.

1.1.3 Parkinsonova nemoc

Parkinsonova nemoc (PN) je v současné době druhým nejčastějším neurodegenerativním onemocněním na světě, po Alzheimerově chorobě [14]. Mechanismus, při kterém odumírají mozkové buňky se šíří od nejhlubších struktur až po vyšší části mozkového kmene. Odumíráním *dopaminergních neuronů* vzniká nedostatek substance *dopaminu* zodpovědného za správnou funkci bazálních ganglií a komunikaci mezi nervovými buňkami [13]. Tím dochází k projevům onemocnění.

Mezi hlavní příznaky nemoci patří svalový třes, snížený rozsah pohybů a jejich zpomalení, neschopnost správně začít pohyb, ztuhlost a postojová

nestabilita [13]. Jedním z motorických symptomů PN pozorovatelných již v preklinické fázi onemocnění je porucha řeči [3]. Díky velké komplexnosti funkcí svalové soustavy při její tvorbě a potřebě velmi jemné motoriky, je tento ukazatel pozorovatelný už v samotných počátcích nemoci. Jeho monitorováním tak může být umožněna případná brzká diagnóza, klíčová pro léčbu.

Poruchy řeči u pacientů s PN. Poruchu řeči u Parkinsonovy choroby, nazývanou *hypokynetická dysartrie*, trpí až 70-90% pacientů. Nemoc ovlivňuje velkou škálu řečových parametrů. Mezi hlavní patří poruchy [16, 17]

- Respirace. Poruchy v dýchacím procesu během řeči vedou ke špatnému frázování a obtížné změně hlasitosti.
- Fonace. Zhoršená schopnost udržet hrtanové svaly ve stabilní poloze působí změny ve frekvenci při produkci znělé hlásky.
- Artikulace. Nedostatečným fungováním vokálního traktu mají pacienti problém vyslovovat některé slabiky, speciálně pokud jsou rychleji za sebou. Dochází ke zpomalování procesu tvorby slabik a řečovému zadrhávání.
- Prozódie. Tímto termínem se souhrnně označují přirozené změny v hlasitosti, tónu, výšce hlasu a rytmu během promluvy. Pacienti s PN mají změny méně četné a málo signifikantní. *Řeč se stává velmi monotónní*, snižuje se schopnost intonace. Při pozorování průběhu F_0 přes celou promluvu byla zjištěna mnohem menší variabilita, než u zdravých lidí.

Všechny tyto poruchy jdou objektivně pozorovat v rámci mnoha řečových ukazatelů [5, 3]. Speciálně snížená intonace, která se dá dobře měřit pomocí směrodatné odchylky F_0, je jádrový příznak poruch řeči již od brzkých stádií PN [4]. Ve studiích [20, 21, 22, 23] bylo ukázáno, že pro analýzu promluv v reálném prostředí za přítomnosti šumu, nahranou mikrofony chytrého telefonu, je přístup hodnocení kvality řeči z hlediska F_0 věrohodný a robustní.

1.2 Metody určování F_0

V průběhu let bylo zavedeno velké množství estimátorů F_0, pracujících s různými metodami.Žádný konkrétní přístup se neukázal jako objektivně nejlepší [7].

Nejčastěji uváděné roztřídění je možno provést podle domény vstupního signálu (řečové promluvy) [6] na estimátory pracující v *časové doméně*, *frekvenční oblasti* nebo kombinující oba přístupy.

Principiální schéma je velmi podobné pro všechny typy metod. Zaznamenávaný signál $x(t)$ je navzorkován frekvencí F_s na diskrétní signál $x[k]$, který je dále segmentován na krátké úseky. Na ty je pak aplikován algoritmus estimátoru, který je možno do jisté míry ovlivnit měnitelnými parametry. Výstupem je odhad $\hat{F}_0$ pro každý časový segment. Výsledkem je pak vektor $\hat{\boldsymbol{F}}_0$ odhadů základní hlasivkové frekvence pro všechny časové segmenty.

V této práci využíváme 3 estimátory základní hlasivkové frekvence, *Praat* [8], *SWIPE* [9] a *BaNa* [2]. V následujících sekcích je jejich principiální popis.

Jelikož tyto algoritmy používají v některých případech stejné nástroje, vysvětlíme princip jejich fungování na začátku, před popisem každého z estimátorů.

1.2.1 Nástroje

Krátkodobá (Short-Time) Fourierova Transformace

Pokud je signál $x[k]$ stacionární, tedy jeho vlastnosti se v čase nemění, definujeme Fourierovu transformaci v diskrétním čase (DtFT) jako

$$X_{\mathrm{DtFT}}(f) = \sum_{k=-\infty}^{\infty} x[k] e^{-\mathrm{j}2\pi f T_\mathrm{s} k}, \tag{1.5}$$

kde T_s je vzorkovací perioda signálu, $T_\mathrm{s} = 1/F_\mathrm{s}$.

Pokud však máme nestacionární signál $x[k]$, nemůžeme tento přístup použít. Signál si tedy rozdělíme do krátkých segmentů, na kterých se chová alespoň přibližně stacionárně. Tento proces realizujeme váhováním signálu vhodně zvoleným oknem $w[k]$, které posouváme po celém časovém průběhu. Takto můžeme definovat tzv. Short-Time Fourierovu tranformaci (StFT) jako

$$X_{\mathrm{StFT}}(f, m) = \sum_{k=-\infty}^{\infty} x[k] w[k-m] e^{-\mathrm{j}2\pi f T_\mathrm{s} k}, \tag{1.6}$$

kde signál $x[k]$ již není stacionární.

Viterbiho algoritmus

Viterbiho algoritmus představuje nástroj na nalezení nejpravěpodobnější *cesty* (posloupnosti) mezi stavy, pokud známe jednotlivé pravděpodobnosti

přechodů [26].

Definujme $r_j[k]$ jako nejlepší „skóre“ (nejvyšší pravděpodobnost) v čase k pro stav S_j ze všech stavů v čase předchozím. Toto skóre určíme rekurzivně přes všechny předchozí stavy jako

$$r_j[k] = \max_{1 \leq i \leq N} \left(r_i[k-1] \cdot \text{Cost}_{ij} \right), \tag{1.7}$$

kde N je počet stavů a Cost_{ij} je „cena“ (pravděpodobnost) přechodu ze stavu S_i do stavu S_j. r_j počátečních stavů musí být známé.

V každém dopředném kroku takto uložíme index i stavu S_i v $r_i[k-1]$, který maximalizuje $r_j[k]$. Příslušné indexy jsou určeny vektorem $\Psi_j[k]$, který tak vlastně představuje „ukazatele“ na nejpravděpodobnější předchozí stav S_i pro S_j.

$$\Psi_j[k] = \arg\max_{1 \leq i \leq N} \left(r_i[k-1] \cdot \text{Cost}_{ij} \right). \tag{1.8}$$

Po určení $\Psi[k]$ pro všechny časy k a stavy je určen nejpravděpodobnější koncový stav, podle jeho hodnoty $r[k]$. Poté nastává *zpětná fáze* algoritmu, kdy postupujeme od posledních hodnot konkrétního $\Psi[k]$ k těm prvním. V každém zpětném kroku zapíšeme aktuální stav a z jeho $\Psi[k]$ určíme nejpravděpodobnější předchozí stav. Takto postupujeme až do časového počátku. Výslednou posloupnost stavů pak označíme jako nejpravděpodobnější možnou a je výstupem algoritmu.

1.2.2 Praat

Praat [8] je jedním z nejrozšířenějších nástrojů na zjišťování časového průběhu F_0 [28]. Jedná se o estimátor pracující v časové doméně, využívající vlastností autokorelační funkce.

Autokorelační funkce. Pokud máme stacionární a *reálný* signál $x[k]$, definujeme jeho autokorelační funkci $r_x[l]$ následovně

$$r_x[l] = \sum_{k=-\infty}^{\infty} x[k]x[k-l]. \tag{1.9}$$

Pro $l = 0$ má $r_x[l]$ globální maximum. Pokud je signál periodický s periodou T_0, bude mít globální maxima i na jejích násobcích.

Pokud signál již přestává být periodický a na násobcích $n \cdot T_0$ již má pouze lokální maxima, stále můžeme periodu určit a tím pádem i základní frekvenci. Takovýto signál může být například součtem periodického signálu a bílého šumu.

Pro klasifikaci násobků T_0 zavedeme veličinu *harmonic strength* R_0. $R_0 \in [0, 1]$ a je rovna hodnotě normalizované autokorelační funkce $r'_x[l]$ v bodě $l_{\max} \neq 0$, pro který je hodnota $r_x[l_{\max}]$ maximální.

Normalizovaná autokorelační funkce má tvar

$$r'_x[l] = \frac{r_x[l]}{r_x[0]}. \tag{1.10}$$

Wienerův–Khinchinův teorém. Výpočet autokorelační funkce je možno provést velmi efektivně, pokud uplatníme Wienerův-Khinchinův teorém. Platí, že autokorelační funkci $r_x[l]$ dostaneme zpětnou Fourierovou transformací spektrální hustoty energie $S(e^{\mathrm{j}2\pi f})$

$$r_x[l] = \int_{(2\pi)} S(e^{\mathrm{j}2\pi f}) e^{\mathrm{j}2\pi l f} \,\mathrm{d}f, \tag{1.11}$$

kde $S(e^{\mathrm{j}2\pi f})$ můžeme spočítat jako kvadrát modulu Fourierovy transformace signálu $x[k]$

$$S(e^{\mathrm{j}2\pi f}) = \left| \sum_{-\infty}^{\infty} x[k] e^{-\mathrm{j}2\pi k f} \right|^2, \tag{1.12}$$

při využití Parsevalovy rovnosti a předpokladu, že signály, se kterými zde pracujeme jsou stacionární v širším smyslu (WSS) a platí vzorkovací teorém. Celý proces by analogicky platil i pro spojitý signál.

Váhování oknem. Pro nestacionární signály je nutné zavést váhování posouvajícím se oknem $w[k]$, které je patřičně ořízne na segment, ve kterém se signál považuje za stacionární. Autoři doporučují použití gaussova okna, které má nejlepší výsledky [8]. Jedná se o okno délky $2T$, kde $w[k]$ je

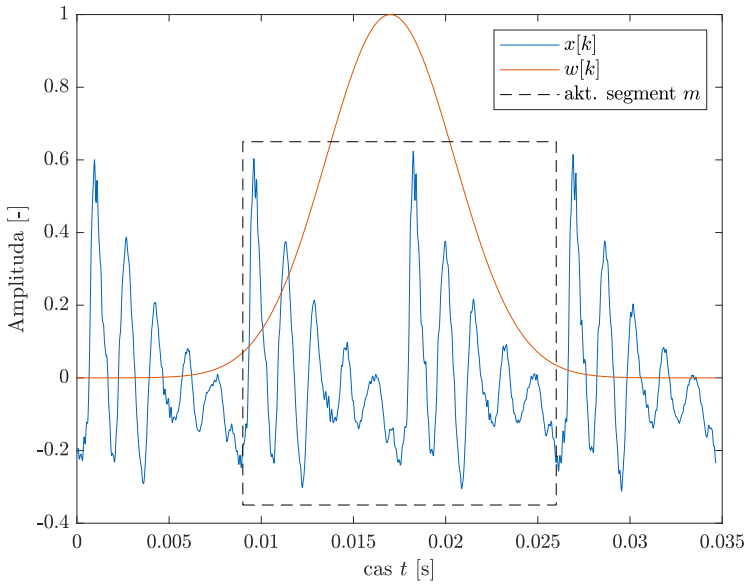
$$w[k] = \begin{cases} \exp\left(-12\left(\frac{k}{T} - \frac{1}{2}\right)^2 - e^{-12}\right) \frac{1}{1-e^{-12}}, & k \in [-\frac{T}{2}, \frac{3}{2}T] \\ 0, \quad \text{jinde} \end{cases}. \tag{1.13}$$

Na obrázku 1.3 je vyobrazen signál $x[k]$ při vyslovení znělé hlásky /a/ spolu s příslušným váhovacím oknem pro daný časový segment. Je vidět, že gaussovo okno lehce přesahuje i do vedlejších segmentů.

Metrika estimátoru. Ukazuje se, že určování odhadu $\hat{F}_0$ z daného segmentu pomocí normalizované autokorelační funkce 1.10 má problém s formantovými frekvencemi, které zaměňuje za fundamentální. Tento problém se však dá úspěšně řešit, pokud autokorelační funkci signálu, pronásobeného $w[k]$ podělíme autokorelační funkcí samotného okna $w[k]$ [8].

Kandidáty $\check{F}_0$ na odhad $\hat{F}_0$ tedy budu hledat jako maxima z následující funkce (metriky)

$$r'[l] = \frac{r'_{xw}[l]}{r'_w[l]}, \tag{1.14}$$



Obrázek 1.3: Průběh signálu $x[k]$ při vyslovení znělé hlásky /a/ spolu s váhovacím oknem, vyřezávající daný segment. Gaussovo okno lehce zasahuje i do vedlejších segmentů.

kde $r'_{xw}[l]$ a $r'_w[l]$ byla získána analogickým postupem podle 1.9 a 1.10 a r_{xw} má tvar

$$r_{xw}[l] = \sum_{k=-\infty}^{\infty} \tilde{x}[k]\tilde{x}[k-l], \tag{1.15}$$

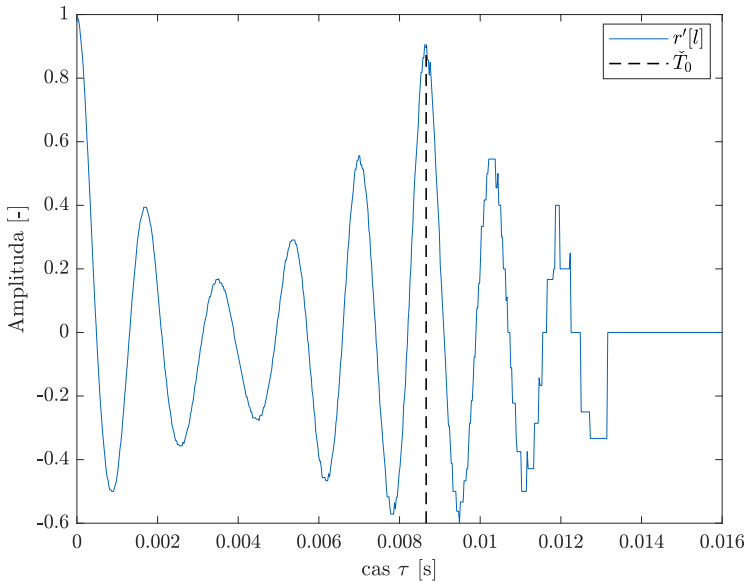
pro

$$\tilde{x}[k] = x[k]w[k]. \tag{1.16}$$

Podoba $r'[l]$ pro stejný signál $x[k]$ jako na obrázku 1.3, tedy znělou hlásku /a/ a gaussova okna pro příslušný segment je na obrázku 1.4. Je viditelně zřetelné lokální maximum, jehož časová pozice nám určí odhad fundamentální periody $\check{T}_0$ z čehož inverzí spočítáme odhad $\check{F}_0$. V každém časovém segmentu takto určíme kandidáty $\check{F}_0$ z pozice maxim funkce 1.14, přičemž kandidáti s hodnotou R_0 pod určitým prahem jsou odmítnuti, stejně jako kandidáti nespadající do námi definovaného rozsahu $[F_{0\min}, F_{0\max}]$.

Vybírání z kandidátů $\check{F}_0$. Po provedení postupu pro všechny časové segmenty v promluvě máme výsledný vektor dvojic hodnot $(\check{F}_{0\,j}[m], R_{0\,j}[m])$, kde m značí index časového segmentu a j index samostatných kandidátů na $\hat{F}_0$ pro každý segment.

Pokud je počet všech segmentů N, definujeme *cestu* $\{p_m, 1 < m < N\}$ jako posloupnost kandidátů $\check{F}_{0\,j}[m]$ pro každý časový segment. Sestavíme funkci $\mathrm{Cost}_{ij}[m]$, která určuje pravděpodobnost přechodu ze stavu $\check{F}_{0\,i}[m-1]$ do stavu $\check{F}_{0\,j}[m]$. Tato funkce závisí na hodnotách R_0 obou těchto stavů a



Obrázek 1.4: Podoba funkce $r'[l]$ pro signál $x[k]$ znělé hlásky /a/ a Gaussovského okna, příslušejícího danému časovému segmentu. Na grafu je znázorněno lokální maximum $r'[l]$ označující kandidáta $\check{T}_0$ resp. $\check{F}_0$ pro daný segment

také na jejich logaritmickém poměru, kdy nechceme velké skoky v časovém průběhu $\hat{F}_0$.

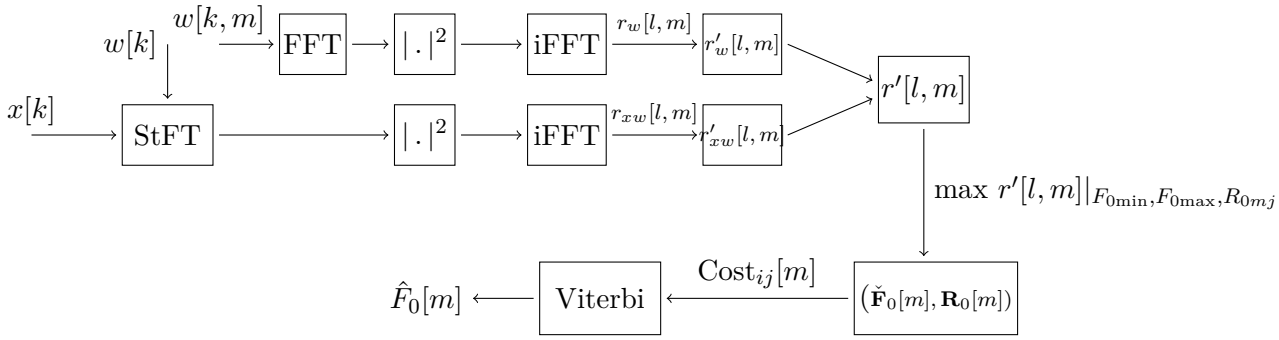
Poté využijeme Viterbiho algoritmus na nalezení nejpravděpodobnější cesty $\{p_m\}$ mezi stavy využitím pravděpodobností přechodů $\text{Cost}_{ij}[m]$. Nalezená posloupnost stavů je pak výstupem celého algoritmu. Schematický princip činnosti estimátoru je na obrázku 1.5.

1.2.3 SWIPE

Sawtooth Waveform inspired Pitch Estimator (SWIPE) [9] pracuje ve spektrální oblasti a snaží se maximalizovat shodu upraveného spektra přijatého signálu s uměle vytvořeným kernelem, přizpůsobeným na očekávané spektrální vlastnosti.

Tvarování spektra signálu. Z přijatého signálu $x[k]$ je vypočítáno jeho spektrum $X(f, m)$ pro daný časový segment m pomocí 1.6. Váhovací okno $w[k]$ je zvoleno Hannovo, tedy

$$w[k] = \frac{1}{T}\left(1 + \cos\left(\frac{2\pi k}{T}\right)\right), \tag{1.17}$$



Obrázek 1.5: Principiální schéma fungování a implementace estimátoru Praat.

pro délku okna T. Hannovo okno bylo zvoleno jako vhodný kompromis mezi jeho spektrálními vlastnostmi, důležitými pro tuto aplikaci a výpočetní náročností.

Ze spektra $X(f, m)$ je spočten jeho modul $|X(f, m)|$. Abychom co nejvíce přizpůsobili estimátor na jeho účely, tj. odhad základní hlasivkové frekvence, přímo spojené s výškou tónu hlasu (viz sekce 1.1.2), je vhodné $|X(f, m)|$ logaritmovat. Tento přístup využívá mnoho dalších metod estimace F_0, jmenujme například Cepstrum [10].

Pro náš případ však tento přístup není vhodný, jelikož modul spektra $|X(f, m)|$ obsahuje některé úseky s velmi malou hodnotou amplitudy. Logaritmus má v těchto úsecích tendenci způsobovat velké „údolí“ ve funkci $\log_{10}|X(f, m)|$, což by nepříjemně ovlivňovalo chod algoritmu a celkové výsledky.

Byl tedy zvolen alternativní přístup, kdy je modul spektra místo logaritmování umocněn na 1/2. Tato volba přinesla nejlepší výsledky oproti jiným možnostem, jako např. umocněním na druhou nebo neupravovat modul spektra vůbec (identita).

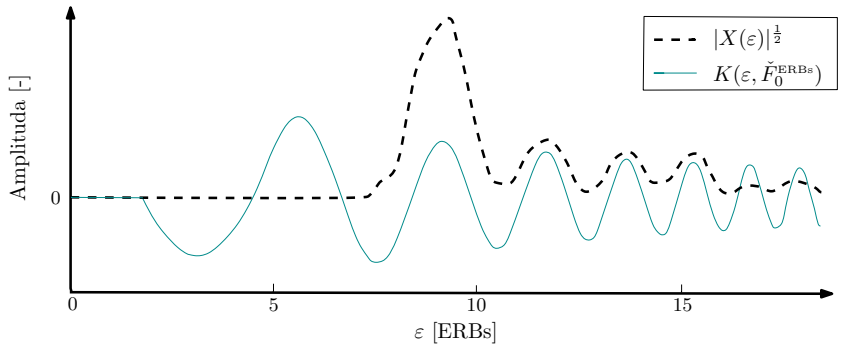
Dalším úpravám bylo podrobeno frekvenční měřítko. Díky mechanismům lidského sluchového ústrojí je vhodné škálu nelineárně změnit pomocí některé z metod, zmíněných v sekci 1.1.2. Autoři jako nejlepší shledali tzv. *ERB scale*, Equivalent Rectangular Bandwidth scale [9], definovaný nelineárním zobrazením ERBs : $f \mapsto \varepsilon$

$$\mathrm{ERBs}(f) = 21.4 \log_{10}\left(1 + \frac{f}{229}\right) \ [\varepsilon], \tag{1.18}$$

kde f je frekvence v Hertzích a ε je frekvence v jednotkách ERBs.

Maximalizace normalizovaného skalárního součinu. Takto upravené spektrum budeme porovnávat s uměle vytvořenou harmonickou funkcí $K(\varepsilon, \check{F}_0^{\mathrm{ERBs}})$ s jasně danou základní frekvencí, tzv. kernelem.

Abychom se zbavili problému subharmonických frekvencí, tedy případu, kdy je za odhad $\hat{F}_0$ zvolen její násobek $n\hat{F}_0, n \in \mathbb{N}$, je kernel váhován klesající



Obrázek 1.6: Spektrum $|X(\varepsilon)|^{1/2}$ signálu $x(t)$ a odpovídající harmonický kernel pro daný časový segment. Ilustrativní obrázek.

funkcí, která bude postupně potlačovat jejich vliv. Ilustrativní obrázek spektra $|X(\varepsilon)|^{1/2}$ a jeho odpovídajícího harmonického kernelu $K(\varepsilon, \check{F}_0^{\mathrm{ERBs}})$ pro daný časový segment je na obrázku 1.6.

Algoritmus estimátoru se bude snažit najít, pomocí maximalizace normalizovaného skalárního součinu, nejlepší shodu mezi spektrem a kernelem pro všechny $\check{F}_0$ jako parametr kernelu. Při nalezení nejlepší možné shody prohlásí parametr kernelu $\check{F}_0$ za odhad estimátoru $\hat{F}_0$ pro daný časový úsek. Prakticky se snažíme o maximalizaci vzájemné energie mezi spektrem $|X(\varepsilon)|^{1/2}$ a kernelem $K(\varepsilon, \check{F}_0^{\mathrm{ERBs}})$

Pro dvě *reálné* skalární funkce $f(x)$ a $g(x)$, které jsou alespoň s kvadrátem integrovatelné, bude mít normalizovaný skalární součin definovaný na intervalu $[a, b]$ tvar

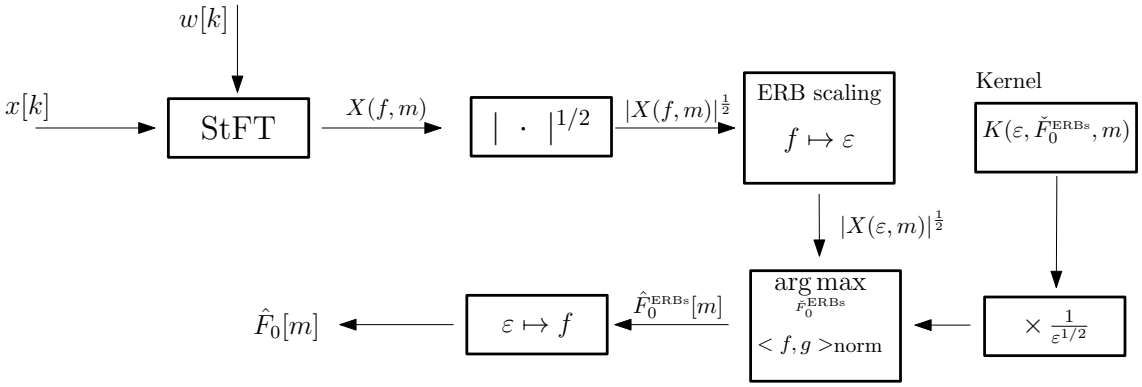
$$< f, g >_{\mathrm{norm}} = \frac{\int_a^b f(x)g(x)\,\mathrm{d}x}{\sqrt{\int_a^b |f(x)|^2\,\mathrm{d}x}\sqrt{\int_a^b |g(x)|^2\,\mathrm{d}x}}. \tag{1.19}$$

Pokud tedy označíme za funkce f a g upravené spektrum a vytvořený kernel (oba splňují předpoklady, za kterých vzorec 1.19 platí), určíme odhad $\hat{F}_0^{\mathrm{ERBs}}$ v jednotkách ERBs pro časový segment m následovně

$$\hat{F}_0^{\mathrm{ERBs}}[m] = \arg\max_{\check{F}_0^{\mathrm{ERBs}}} \frac{\int_0^{\varepsilon_{\max}} \frac{1}{\varepsilon^{1/2}} K(m, \varepsilon, \check{F}_0^{\mathrm{ERBs}})\,|X(m,\varepsilon)|^{\frac{1}{2}}\,\mathrm{d}\varepsilon}{\left(\int_0^{\varepsilon_{\max}} \frac{1}{\varepsilon} K^2(m, \varepsilon, \check{F}_0^{\mathrm{ERBs}})\,\mathrm{d}\varepsilon\right)^{1/2} \left(\int_0^{\varepsilon_{\max}} |X(m,\varepsilon)|\,\mathrm{d}\varepsilon\right)^{1/2}}, \tag{1.20}$$

kde $\varepsilon_{\max}$ určuje maximální frekvenci pro kterou odhad hledáme. Finální výsledek pak dostaneme jen převedením odhadu základní frekvence $\hat{F}_0^{\mathrm{ERBs}}[m]$ v jednotkách ERBs na lineární frekvenční osu, čímž určíme $\hat{F}_0[m]$ nyní v jednotkách Hertz.

V každém časovém segmentu je také určen tzv. „pitch strength“, hodnota, určující jak moc je přijatý signál podobný perfektnímu tónu, tedy o jedné frekvenci. Jeden ze vstupních parametrů estimátoru je práh této hodnoty, kdy



Obrázek 1.7: Principiální schéma fungování estimátoru SWIPE.

jsou odhady $\hat{F}_0$ mající „pitch strength“ pod touto úrovní klasifikovány jako neřečové a nejsou využity. Tato možnost se nám bude velmi hodit v dalších částech, kdy budeme testovat úspěšnost algoritmů za přítomnosti šumu.

Principiální schéma fungování estimátoru SWIPE je na obrázku 1.7.

1.2.4 BaNa

Jedná se o relativně nedávno navrhnutý estimátor [2] pracující ve spektrální oblasti využívající poměry harmonických násobků základní frekvence ve spektru.

Poměry spektrálních vrcholů. Ze vstupního signálu $x[k]$ je pomocí Short-Time Fourierovy tranformace vypočítáno jeho spektrum, ze kterého nás zajímá modul $|X(f,m)|$. V každém časovém segmentu m jsou vypočítány pozice prvních p spektrálních vrcholů $F_i'[m], i = 0, 1, \ldots, p-1$, jejichž odpovídající magnitudy $|X'(F_i, m)|$ přesahují hodnotu určitého prahového parametru pTHR. Důvod výběru pouze prvních p vrcholů je ten, že harmonické násobky F_0 bývají více rozpoznatelné na nižších frekvencích [2].

Z určených pozic $F_i'[m]$ vypočítáme všechny vzájemné poměry

$$R_{ij}[m] = \frac{F_j'[m]}{F_i'[m]}, \quad i < j, \quad \begin{matrix} i = 0, 1, \ldots p-2 \\ j = 1, 2, \ldots p-1 \end{matrix} . \tag{1.21}$$

Tato metoda spočívá v tom, že poměrně přesně víme, jaký bude poměr dvou harmonických násobků F_0. Pokud některý z $R_{ij}[m]$ vyhovuje tomuto poměru v určitých mezích věrohodnosti, dokážeme určit z jakých násobků F_0 jsme ho určovali. Kandidát na odhad $\hat{F}_0[m]$, $\check{F}_0[m]$ pak bude

$$\check{F}_{0,ij}[m] = \frac{F_j'[m]}{j+1} \quad \text{pokud} \quad R_{ij}[m] \in [R^{\mathrm{exp}} - \xi, R^{\mathrm{exp}} + \xi], \tag{1.22}$$

kde R^{exp} značí některou z hodnot očekávaného poměru dvou harmonických násobků F_0 a ξ určuje meze věrohodnosti okolo R^{exp}, jelikož $R_{ij}[m]$ nebude mít nikdy přesně stejnou hodnotu.

Příklad: máme zjištěné pozice dvou spektrálních vrcholů F'_1 a F'_3 v konkrétním časovém segmentu. Zjistíme, že jejich poměr $R_{13} = F'_3/F'_1$ padne do rozsahu $R^{\text{exp}} \pm \xi$, pro určitý známý poměr R^{exp}, který by násobky F_0 měly mít. Výsledný kandidát na odhad z poměru těchto dvou vrcholů tedy bude $\check{F}_{0,13} = F'_3/4$.

Povšimneme si, že v každém časovém segmentu takto nenavrhujeme první spektrální vrchol $F'_0[m]$. Přidáme ho tedy jako $\check{F}'_0[m]$ k seznamu kandidátů určených z jednotlivých poměrů. Dále pro zvýšení univerzality a dosažení lepších výsledků [2] přidáme dalšího kandidáta $\check{F}^{\text{ceps}}[m]$, získaného Cepstrální metodou [10].

Výběr odhadu základní frekvence z kandidátů. K každém časovém segmentu tak určuji výsledný odhad $\hat{F}_0[m]$ z množiny kandidátů

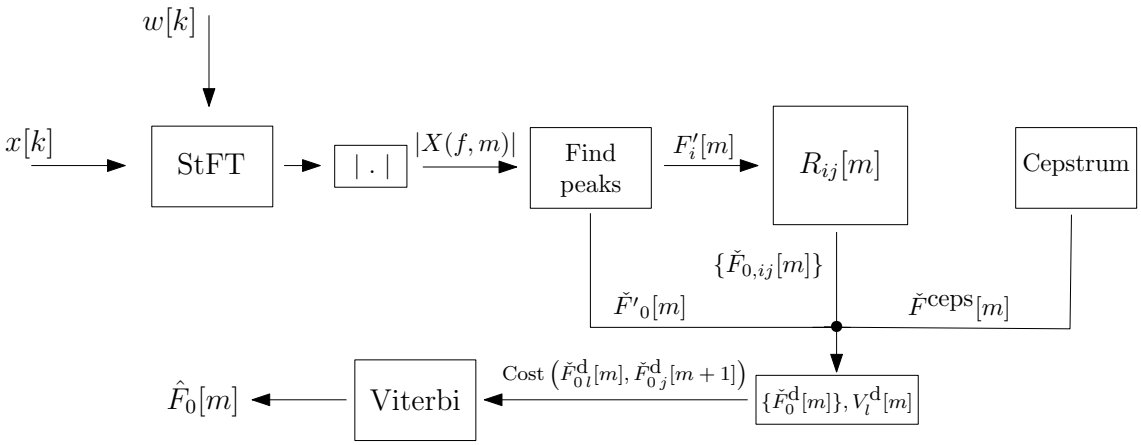
$$\{\check{F}_0[m]\} = \left\{\check{F}^{\text{ceps}}[m], \check{F}'_0[m], \{\check{F}_{0,ij}[m]\}\right\}, \tag{1.23}$$

tedy kandidát získaný Cepstrální metodou, pozice prvního spektrálního vrcholu a kandidáti určení z jednotlivých poměrů.

Pro každý časový úsek určíme „významné“ kandidáty $\{\check{F}^{\text{d}}_0[m]\} \in \{\check{F}_0[m]\}$ následujícím procesem. Ti kandidáti, kteří se od sebe liší jen o ε [Hz] jsou označeni jako jedna množina „blízkých“ kandidátů. Z každé takové množiny vybereme takového kandidáta, který má počet „blízkých“ největší. Dostáváme tak množinu „významných“ kandidátů $\{\check{F}^{\text{d}}_0[m]\}$ spolu s odpovídající množinou věrohodnostního skóre $V^{\text{d}}_l, l = 1, \dots L$, která značí počet „blízkých“ kandidátů každého l-tého „významného“ kandidáta pro L jako počet „významných“ kandidátů. Výsledný odhad pro daný časový úsek tedy určujeme z množiny $\{\check{F}^{\text{d}}_0[m]\}$.

Na vybrání správných kandidátů si pro každý segment m zavedeme funkci pravděpodobnosti přechodů $\text{Cost}\left(\check{F}^{\text{d}}_{0\,l}[m], \check{F}^{\text{d}}_{0\,j}[m-1]\right)$. Tato funkce bere v potaz věrohodnostní skóre $V^{\text{d}}_l[m-1]$, $V^{\text{d}}_j[m-1]$ a také velikost skoku odhadu $\check{F}^{\text{d}}_{0\,l}[m-1]$ do $\check{F}^{\text{d}}_{0\,j}[m]$. Jelikož by nemělo docházet k velkým skokům v F_0 mezi sousedícími segmenty, budou upřednostňováni kandidáti s malým vzájemným rozdílem. Ten je díky lepší kompatibilitě s lidským zvukovým vnímáním vyjádřen ve stupnici *mel*, viz sekce 1.1.2.

Podobně jako u estimátoru Praat, nyní hledáme nejpravděpodobnější *cestu*, (posloupnost) mezi stavy za použití informace v $\text{Cost}\left(\check{F}^{\text{d}}_{0\,l}[m], \check{F}^{\text{d}}_{0\,j}[m-1]\right)$. Tento proces opět provedeme Viterbiho algoritmem. Výstupem bude časová posloupnost nejpravděpodobnějších kandidátů $\hat{F}_0$. Schematický obrázek, popisující principiální fungování estimátoru BaNa je na obrázku 1.8.



Obrázek 1.8: Schéma popisující principiální fungování estimátoru BaNa. Blok označený jako $R_{ij}[m]$ určuje kandidáta na $\hat{F}_0[m]$, $\check{F}_{0,ij}[m]$ podle rovnice 1.22.

Kapitola 2

Testování estimátorů

2.1 Výběr estimátorů

Jak již bylo řečeno, existuje velké množství metod zjišťování F_0. Žádná není dokonale univerzální pro všechny typy aplikací. Existuje několik studií porovnávající některé vybrané estimátory na různých případech. V této práci jsme se především inspirovali článkem [1], porovnávající 10 nejvíce známých estimátorů, používaných v dnešní době.

Článek [1] však pracuje s krátkými řečovými záznamy *pouze znělých hlásek*. Tedy velmi rozdílný případ oproti našemu, kdy analyzujeme delší plnohodnotnou řečovou promluvu, navíc v reálném prostředí. Výsledky se tím pádem mohou velice lišit.

Dále jsme brali v potaz dosažitelnou implementaci, také detailní dokumentaci a výpočetní náročnost. Ve výsledku byly vybrány na testování tyto estimátory

- **Praat**, [8]
  Jedná se o standardní nástroj na analýzu řeči v klinické praxi. Často jsou k němu porovnávány výsledky získané z jiných metod.
- **SWIPE**, [9]
  Osvědčil se, podle [1], jako velmi efektivní metoda odhadu F_0. Také je výpočetně velice nenáročný.

- **BaNa**, [2]
  Vybrán jako zástupce algoritmů speciálně cílených na dobrou šumovou odolnost a použití v chytrých telefonech.

Jejich principiální fungování bylo naznačeno v 1.2.2, respektive 1.2.3 a 1.2.4. Krátkému testování byl podroben také estimátor YIN [11], ten ale již na začátku dával velmi špatné výsledky v porovnání s ostatními a proto byl vyřazen.

Jednotlivé algoritmy byly testovány podle následující metodiky. Nejprve jsme vyhodnotili robustnost na čistých nahrávkách bez šumu a podle toho zvolili vhodné vstupní parametry. Dále bylo provedeno zhodnocení šumové odolnosti při umělém přidání různých typů šumů do původních čistých nahrávek. V návaznosti na to jsme prozkoumali časový průběh odhadu $\hat{F}_0$ se zaměřením na problémy, na kterých estimátory mohou selhávat. Také jsme se soustředili na charakteristiky časového průběhu a s tím spojenou možnou filtraci. V poslední testovací části pak byly zkoumány reálné nahrávky, pořízené v rušném prostředí, kde jsme se snažili odhadnout úroveň přítomnosti šumu.

Nejprve však musíme uvést podle jakých kritérií budeme estimátory hodnotit.

2.2 Kritéria testování

Jak již bylo řečeno v úvodu, naše cílová zkoumaná hodnota bude směrodatná odchylka z časového průběhu F_0. Vzhledem ke zmíněným rozdílům v rozdělení F_0 pro odlišná pohlaví a různý věk, převedeme si hodnoty do semitónové škály, popsané v sekci 1.1.2, konkrétně využijeme vztah 1.4. Jelikož dopředu neznáme, jaké rozdělení hodnot bude průběh mít, pracujeme s výběrovou směrodatnou odchylkou. Zkoumaná veličina v rámci jedné promluvy tedy bude rovna

$$\hat{\sigma}_{F_0}^{\text{sem}} = \sqrt{\frac{1}{N-1}\sum_{m=0}^{N-1}\left(\hat{F}_0^{\text{sem}}[m] - \hat{\mu}\right)^2} \quad [\text{sem}], \tag{2.1}$$

kde m značí daný časový úsek, N je počet časových úseků, $\hat{F}_0^{\text{sem}}[m]$ je odhad F_0 estimátoru pro daný čas. segment převedený do semitónů podle 1.4 a $\hat{\mu}$ je odhad střední hodnoty jako aritmetický průměr hodnot celého průběhu,

$$\hat{\mu} = \frac{1}{N}\sum_{m=0}^{N-1}\hat{F}_0^{\text{sem}}[m] \quad [\text{sem}]. \tag{2.2}$$

RMSE. Jedno z používaných kritérií při hodnocení estimátorů je RMSE (Root Mean Square Error) $\sigma_{F_0}^{\text{sem}}$, definované jako

$$\text{RMSE}_{\sigma_{F_0}^{\text{sem}}} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\sigma_{F_0}^{\text{sem}}[i] - \hat{\sigma}_{F_0}^{\text{sem}}[i]\right)^2} \quad [\text{sem}], \tag{2.3}$$

kde $\sigma_{F_0}^{\text{sem}}[i]$ je skutečná hodnota odchylky pro i-tou promluvu, $\hat{\sigma}_{F_0}^{\text{sem}}[i]$ je její odhad z patřičného estimátoru a M je počet analyzovaných řečových promluv.

Toto kritérium bylo zvoleno jako standardní používané při estimaci F_0 [1]. Mezi další používaná kritéria patří MAE (*mean absolute error*) a MRE (*mean relative error*).

Důvod, proč můžeme využít výše zmíněnou definici 2.3 je, že jsme schopni určit referenční hodnotu, tzv. *Gold standard*, kterou simulujeme skutečnou hodnotu $\sigma_{F_0}^{\text{sem}}$. Viz další sekce 2.2.1.

Spearmanův korelační koeficient. Další kritérium, které budeme používat, opět využívá možnosti Gold standardu. V našem případě volíme kritérium *výběrového Spearmanova korelačního koeficientu* r jako metriku míry nelineární korelace mezi dvěma náhodnými veličinami. $r \in [-1, 1]$, kde 1 značí úplnou korelaci, 0 žádnou a -1 úplnou negativní korelaci. Výběrový Spearmanův koeficient r pro veličiny $\sigma_{F_0}^{\text{sem}}$ a $\hat{\sigma}_{F_0}^{\text{sem}}$ je definovaný [25] jako

$$r = \frac{\sum_{i=1}^{M}\left(\text{r}_i(x) - \overline{\text{r}(x)}\right)\left(\text{r}_i(y) - \overline{\text{r}(y)}\right)}{\sqrt{\sum_{i=1}^{M}\left(\text{r}_i(x) - \overline{\text{r}(x)}\right)^2}\sqrt{\sum_{i=1}^{M}\left(\text{r}_i(y) - \overline{\text{r}(y)}\right)^2}}, \tag{2.4}$$

kde i je index promluv a M jejich počet, x značí vektor odhadů $\hat{\sigma}_{F_0}^{\text{sem}}$ pro M promluv, analogicky y, reprezentuje referenční hodnoty Gold standardu $\sigma_{F_0}^{\text{sem}}$. Operátor $\text{r}(\cdot)$ se označuje jako *rank* a určuje *pořadí* prvků v rámci výběru. Například $\text{r}(2.0, 1.5, 3.1, 0.8) = (3, 2, 4, 1)$. Symbol r_i pak značí i-tý prvek tohoto pořadí a symbol $\overline{\text{r}(\cdot)}$ je odhad střední hodnoty celé posloupnosti, určený analogicky podle vztahu 2.2.

Vzorec 2.4 platí obecně. Pokud však budu mít posloupnost pořadí $\text{r}(\cdot)$ takovou, že každý prvek je v ní zastoupen právě jednou (tedy původní výběr měl všechny hodnoty různé), situace se velmi zjednoduší [25] na vztah

$$r = 1 - \frac{6\sum_{i=1}^{M} d_i^2}{M(M^2-1)}, \tag{2.5}$$

kde d_i je rozdíl v pořadí pro i-tý prvek v prvním a druhém výběru. Tento vztah zde budeme používat primárně, jelikož vypočítaný odhad je určen na 4 řády přesnosti a v poměrně malém počtu nahrávek ($M = 60$) je pravděpodobnost přesné shody u dvou promluv velmi malá.

Důvod použití právě Spearmanova korelačního koeficientu oproti jiným, například *Pearsonova*, nebo *Kendallova*, je ten, že má věrohodnější výsledky pro náhodné veličiny, které nemají normální rozdělení a dává menší váhu extrémním „outlierům“, hodnotám velmi rozdílných od střední hodnoty. Což jak uvidíme později, nám vyhovuje.

2.2.1 Gold standard

Referenční hodnota $\sigma_{F_0}^{\text{sem}}$, kterou prohlásíme za tzv. *Gold standard*, nám poskytne nástroj k vyhodnocení kvality estimátorů, kdy pak budeme schopni určit patřičný RMSE pro daný estimátor z definice 2.3, spolu se Spermanovým korelačním koeficientem r.

Gold standard byl určen pomocí estimátoru Praat, používaného běžně v klinické praxi. Bylo manipulováno s povoleným rozsahem $[F_{0\,\text{min}}, F_{0\,\text{max}}]$ tak, aby byl zachován co nejpřesnější odhad F_0 (při standardním nastavení ostatních parametrů) a u každého signálu byla provedena vizuální i poslechová kontrola, že je F_0 zachycena dobře. Snažili jsme se zejména vyhnout efektu „pitch halving“ a „pitch doubling“, který by významně zkreslil výsledky. Z něj pak byla pomocí vzorce 2.1 vypočtena reference $\sigma_{F_0}^{\text{sem}}$.

2.3 Vyhodnocení robustnosti na čistých nahrávkách

V této části se budeme zabývat testováním estimátorů na čistých nahrávkách bez šumu a vyladěním vstupních parametrů.

Databáze promluv. Nahrávky byly pořízeny na chytrý telefon Sony Xperia Z1 v tichém prostředí. Vzorkovací frekvence F_s byla nastavena na

$$F_s = 44\,100\ [\text{Hz}], \tag{2.6}$$

nejvyšší, kterou přístroj umožňoval.

Nahrávání probíhalo v letech 2015 - 2017 a studie byla schválena etickou komisí Všeobecné fakultní nemocnice v Praze. Databáze obsahuje záznamy monologu 30ti pacientů (26 mužů, 4 ženy) s nově diagnostikovanou Parkinsonovou nemocí, určenou z kritérií Parkinson‘s disease society [18]. Základní klinické charakteristiky této skupiny jsou uvedeny v tabulce 2.1. Jako kontrolní skupina, 30 zdravých lidí (26 můžu a 4 ženy) s průměrným věkem 65

Skupina PN	
Průměrný věk [roky]	62.3 (σ 11.3)
Průměrná doba trvání symptomů [roky]	1.9 (σ 1.3)
Průměrná hodnota Hoehn & Yahr škály hodnocení	2.1 (σ 0.4)
Průměrná hodnota MDS-UPDRS III škály hodnocení, celkově	29.4 (σ 12.8)
Průměrná hodnota MDS-UPDRS III škály hodnocení, „speech item“	0.6 (σ 0.6)
Počet pozitivních případů PN v rodině (%)	20

Tabulka 2.1: Základní klinické charakteristiky skupiny pacientů s nově diagnostikovanou Parkinsonovou chorobou. σ značí směrodatnou odchylku a MDS-UPDRS je škála hodnocení nemoci, *Movement Disorder Society – Unified Parkinson's Disease Rating Scale*.

let a směrodatnou odchylkou σ 10 let a jejich promluvy byly také zahrnuty do studie.

Všichni pacienti s PN se stali součástí studie při první návštěvě lékaře a byl zaznamenán jejich monolog ještě před tím, než byla zahájena symptomatická léčba. Žádný pacient neabsolvoval v minulosti žádnou z jejích forem. Diagnóza byla provedena neurologickými experty a pacienti byli ohodnoceni škálou MDS-UPDRS III (*Movement Disorder Society – Unified Parkinson's Disease Rating Scale* [19]), značící závažnost motorických symptomů na hodnotách mezi 0 (žádné příznaky) a 132 (velmi závažné příznaky). Škála také obsahuje hodnocení pro kvalitu řeči (tzv. „speech item“) v rozsahu 0 (normální řeč) a 4 (nesrozumitelnou).

Vstupní parametry. Parametry stejné pro všechny tři estimátory byly následující

$$\begin{aligned} F_{0\text{min}} &= 50 \text{ [Hz]}, \\ F_{0\text{max}} &= 500 \text{ [Hz]}, \\ \text{timestep} &= 0.01 \text{ [s]}, \end{aligned} \tag{2.7}$$

kde velikost timestep značí velikost časového segmentu a $F_{0\text{min}}, F_{0\text{max}}$ rozmezí ve kterém hledám odhad $\hat{F}_0$. Takovéto rozmezí je již dostatečně univerzální pro všechny typy lidí, zároveň však není tak široké aby mohlo obsahovat i případné chyby velmi vzdálené od správné hodnoty. Timestep byl zvolen jako ideální kompromis mezi přesností a výpočetní náročností, kdy při vyšších hodnotách nebudeme mít průběh dostatečně jemný a při nižších máme velký počet vzorků.

Vstupní parametry vlastní danému estimátoru byly většinou vzaty takové, které doporučují autoři, jelikož pro čistou promluvu se opravdu ukazovaly jako ideální.

2.3.1 Výsledky

Na obrázku 2.1 jsou vyobrazeny odhady estimátorů $\hat{\sigma}_{F_0}^{\text{sem}}$ oproti Gold standardu $\sigma_{F_0}^{\text{sem}}$ pro 60 promluv. Plné body označené jako HC jsou promluvy zdravých lidí, PN jsou pořízené u pacientů s Parkinsonovou chorobou. Každý bod na grafu je tak reprezentován svými souřadnicemi $[\hat{\sigma}_{F_0}^{\text{sem}}, \sigma_{F_0}^{\text{sem}}]$ a představuje jednu promluvu. Pokud by se tyto hodnoty přesně rovnaly, tj. nastala by dokonalá shoda, budou ležet na šedivé přímce vyznačené v grafu. Čím jí jsou jednotlivé body blíže, tím je odhad přesnější. Zároveň je v legendě vyznačena hodnota příslušného Spearmanova korelačního koeficientu r, vypočtená podle 2.5. Spolu s tím je hodnota p, značící pravděpodobnost nulové korelace za předpokladu nenulové. Tedy např. hodnota $p = 0.05$ značí, že náš předpoklad existence korelace by měl být v 5% případů mylný.

Je vidět, že Praat, 2.1a, má některé chyby odhadu opravdu velké oproti zbylým dvěma estimátorům. Hodnota $r \cong 0.27$ také naznačuje, že tento estimátor v našem případě moc přesný nebude. BaNa, 2.1b, je o poznání lepší, což demonstruje i hodnota $r \cong 0.62$. Na rozdíl od Praatu zde nenajdeme tak výrazné chyby odhadu. Podobnou úspěšnost jako BaNa, ještě trochu vylepšenou zde prokazuje SWIPE, 2.1c. Jeho chyby odhadu jsou konzistentní a prakticky nenajdeme velmi vzdálené „outliery“. Hodnota $r \cong 0.67$ značí, že tento estimátor si zatím vede nejlépe.

V další sekci budeme zjišťovat úspěšnost z pohledu RMSE.

RMSE. Na obrázku 2.2 nalezneme výsledky při spočtení příslušného RMSE podle 2.3 pro každý z estimátorů pro 60 čistých promluv. Víceméně je zde vidět stejný trend jako v předchozí části, kdy Praat je o třídu horší než BaNa a SWIPE (rozdíl v RMSE je větší jak 1 semitón), kde SWIPE je ve výsledku o trochu přesnější (RMSE je menší o zhruba 0.3 semitónu než u BaNa).

V následující části porovnáme úspěšnost odhadu mezi jednotlivými estimátory.

Úspěšnost odhadu mezi estimátory. V této sekci zhodnotíme úspěšnost mezi jednotlivými estimátory, konkrétně určíme, v kolika případech byl daný algoritmus nejlepší. Na ilustrativním obrázku 2.3 jsou vidět odhady $\hat{\sigma}_{F0}^{\text{sem}}$ estimátorů pro 10 promluv zdravých lidí a 10 pacientů s Parkinsonovou chorobou spolu s Gold standardem, reprezentovaný černým křížkem. Čím je odhad blíže, tím je přesnější. Je zřejmé, že se výsledky dramaticky liší nejen pro typ estimátoru, ale také pro typ promluvy, kdy např. v nahrávce č. 4 byly všechny algoritmy velmi přesné a v nahrávce č. 19 mají poměrně velkou chybu odhadu.

V tabulce 2.2 je uvedeno, kolikrát měl daný estimátor mezi ostatními nejlepší odhad pro konkrétní promluvu, reprezentováno veličinou n_{best}. Nejlepší

	Praat	**BaNa**	**SWIPE**
n_{best}	14	18	28

Tabulka 2.2: Tabulka ukazující kolikrát má daný estimátor nejlepší odhad pro 60 čistých promluv.

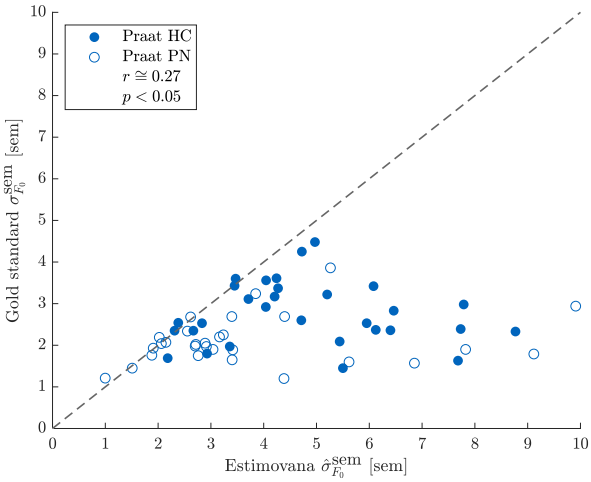
odhad je takový, jehož chyba odhadu $|\hat{\sigma}_{F_0}^{\text{sem}} - \sigma_{F_0}^{\text{sem}}|$ je nejmenší mezi ostatními estimátory pro danou promluvu.

Toto zhodnocení skýtá zajímavou informaci. I když je Praat horší než zbylé dva estimátory, v některých případech je jeho odhad velmi přesný, dokonce lepší než u BaNa a SWIPE. Počet těchto případů přitom není zanedbatelný (14 oproti 18ti, resp. 28mi).

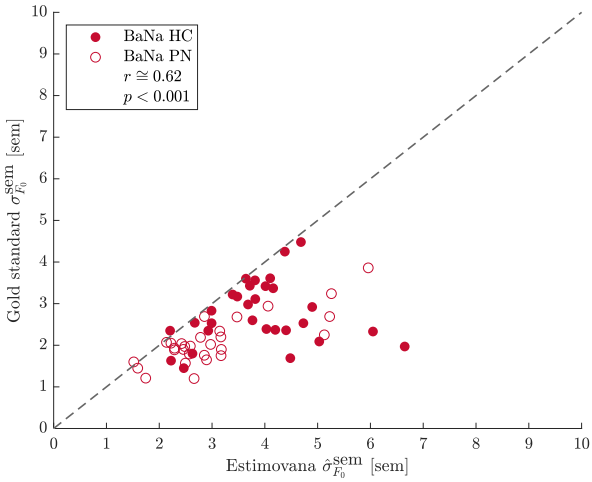
Zhodnocení. Z výsledků uvedených v této sekci je zřejmé, že pro čisté promluvy jsou estimátory SWIPE a BaNa poměrně přesné, kdy SWIPE má ale ve většině případů menší chybu odhadu než BaNa (viz tabulka 2.2). Estimátor Praat má o třídu horší úspěšnost z hlediska RMSE (viz obrázek 2.2), nicméně úplně k zahození také není, jelikož pro nezanedbatelný počet promluv je jeho odhad nejlepší (viz. tabulka 2.2). Pro zbylé promluvy je však nepřesný, v některých případech opravdu hodně (viz obrázek 2.1).

Je ovšem důležité zmínit, že v našem případě 100%ní přesnost není přímo požadována. Variabilita lidské mluvy je dosti velká, pro dvě stejné promluvy může mít ten samý člověk výsledky s odchylkou až do 20% [4]. Je tedy nutnost sledovat dlouhodobý trend, vývoj. Pokud bude estimátor poskytovat dostatek přesných odhadů pro sledované období, že „outliery“ bude možné identifikovat a odstranit, bude vývoj použitelný pro případnou diagnostiku i při nedosažení úplné přesnosti všech odhadů.

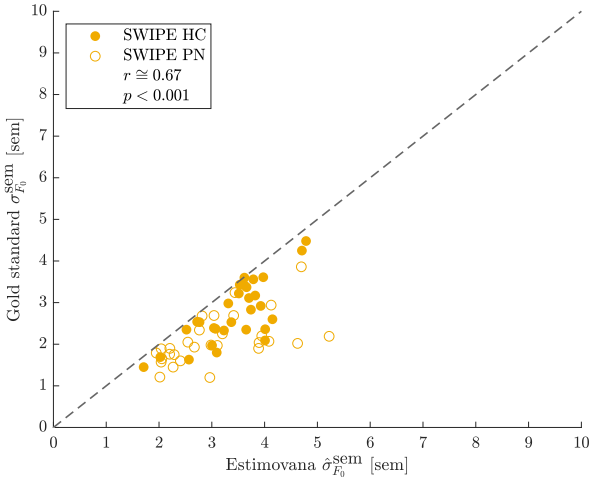
Tuto sekci tedy opouštíme s tím, že větší důvěru vkládáme do estimátorů SWIPE a BaNa, kdy hlavní váhu má SWIPE. Praat má horší výsledky, ale zatím s ním dále počítáme, jelikož pro nezanedbatelný počet případů měl i on velmi dobré výsledky. Je však otázkou, co s notoricky křehkými estimátory F_0 vůči kvalitě nahrávky udělá přidávání šumů, čímž se budeme zabývat v další kapitole.



(a) : Praat

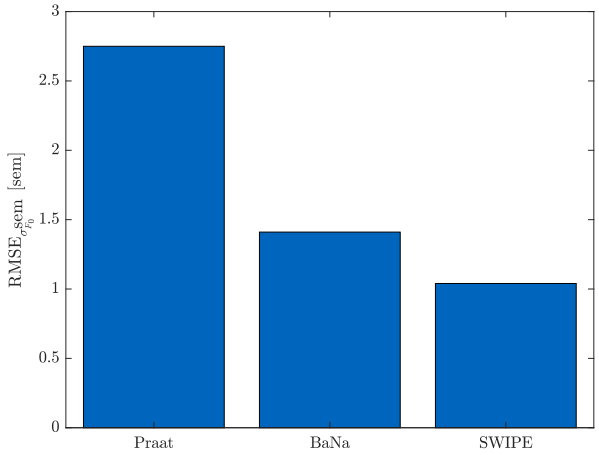


(b) : BaNa

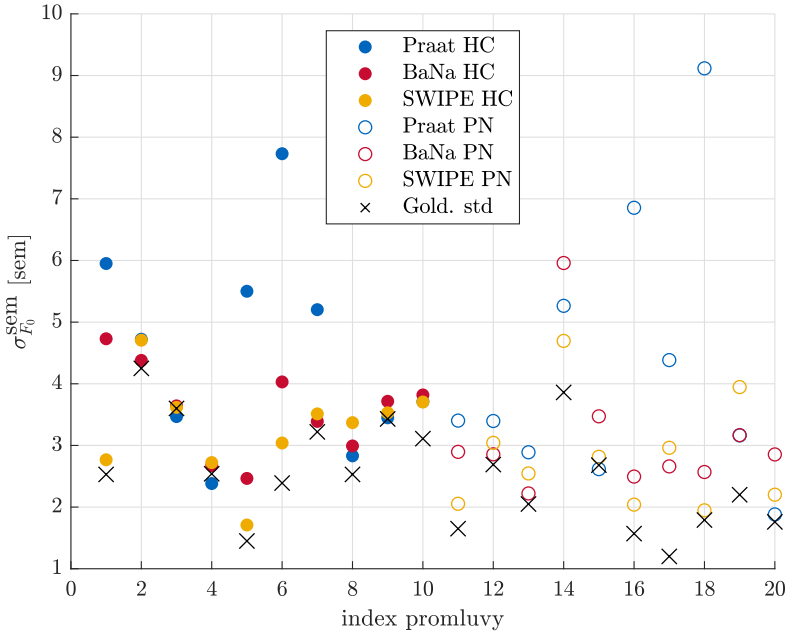


(c) : SWIPE

Obrázek 2.1: Srovnání jednotlivých algoritmů na 60ti čistých promluvách a jejich porovnání s referenční hodnotou. Šedivá úsečka značí pozice, kdy se odhad z estimátoru přesně rovná referenci. HC označuje zdravé lidi a PN pacienty s Parkinsonovou nemocí.



Obrázek 2.2: Porovnání úspěšnosti estimátorů podle příslušného RMSE pro 60 čistých promluv.



Obrázek 2.3: Odhady jednotlivých estimátorů spolu s Gold standardem pro 10 promluv zdravých lidí a 10 pacientů s PN. Na vodorovné ose je index promluv a na svislé příslušné hodnoty $\hat{\sigma}_{F_0}^{\text{sem}}$ a $\sigma_{F_0}^{\text{sem}}$.

2.4 Vyhodnocení šumové odolnosti

Klíčová vlastnost estimátorů pro naší aplikaci, kterou je nutno otestovat, jsou jejich výsledky na promluvách, které nebyly nahrány v tichém prostředí, jelikož pacienti mohou pořizovat záznamy své promluvy v podstatě kdekoliv. Aby byla možná objektivní analýza, vyhodnocení šumové odolnosti provedeme tak, že ke stejným promluvám z předchozí části, tedy nahraných v tichu, uměle přidáme 4 typy šumů reálného prostředí na různých hladinách odstupu signálu od šumu, SNR (signal-to-noise ratio).

Typy šumů. Za pomocí chytrého telefonu Sony Xperia Z1 byly pořízeny 4 záznamy šumů na vhodně vybraných místech. Jednalo se o typická prostředí, ve kterých se lidé během dne pohybují, ale přitom obsahují nezanedbatelný šum na pozadí. Byla snaha o co nejrůznější prostředí, abychom pokryli co největší oblast jakým způsobem může být hlasový záznam rušen.

Jednalo se o tato prostředí

1. Rušná křižovatka s projíždějícími automobily, motorkami, tramvajemi a občasnými procházejícími chodci.
2. Ulice s velkým počtem chodců, často hovořících mezi sebou. Hluk dopravy (automobily, tramvaje) v dálce, cca 200 metrů od místa pořízení nahrávky.
3. Uvnitř jedoucí tramvaje, brždění, rozjíždění a hlášení stanic.
4. Nákupní centrum, větší počet chodců, hovořících mezi sebou, reprodukovaná hudba z obchodů a pípání pokladen supermarketu.

Fotky míst, kde byla nahrávka pořízena, jsou zobrazeny v příloze B.

Umělé přidávání šumů do nahrávek. Uměle zašuměnou promluvu $u[k]$ získáme jako součet zaznamenané čisté promluvy $x[k]$ s nahraným šumem $\xi[k]$

$$u[k] = x[k] + \xi[k], \tag{2.8}$$

kde $x[k]$ a $\xi[k]$ mají stejný počet vzorků. Na zhodnocení, do jaké míry je v nahrávce přítomný šum, budeme $\xi[k]$ váhovat zatlumovacím (příp. zesilovacím) koeficientem c, tak aby $u[k]$ mělo SNR na definovaných úrovních. Jako reprezentativní byly zvoleny úrovně 20, 10, 6 a 0 dB SNR.

Výpočet koeficientu c, odpovídajícího dané úrovni SNR provedeme následovně. Vyjdeme z definice SNR

$$\mathrm{SNR} = 10 \log_{10} \frac{P_x}{P_\xi} \ [\mathrm{dB}], \tag{2.9}$$

kde P_x a P_ξ jsou průměrné výkony čisté nahrávky a šumu, spočítané jako

$$P_x = \frac{1}{N}\sum_{k=1}^{N} |x[k]|^2, \tag{2.10}$$

kde N je počet vzorků nahrávky. Pro P_ξ platí analogicky stejný postup.

Pokud si označím $\xi'[k] = c \cdot \xi[k]$ a odpovídající průměrný výkon P'_ξ, bude platit vztah pro SNR

$$\begin{aligned} \mathrm{SNR} &= 10 \log_{10} \frac{P_x}{P'_\xi} \\ &= 10 \log_{10} \frac{\frac{1}{N}\sum_{k=1}^{N} |x[k]|^2}{\frac{1}{N}\sum_{k=1}^{N} |c \cdot \xi[k]|^2} \\ &= 10 \log_{10} \frac{\sum_{k=1}^{N} |x[k]|^2}{c^2 \sum_{k=1}^{N} |\xi[k]|^2}, \end{aligned} \tag{2.11}$$

z čehož spočítáme odpovídající hodnotu koeficientu c jako

$$c = \sqrt{\frac{\sum_{k=1}^{N} |x[k]|^2}{10^{\mathrm{SNR}/10} \sum_{k=1}^{N} |\xi[k]|^2}}. \tag{2.12}$$

Zašuměná nahrávka $u[k]$, kterou budeme analyzovat tedy je

$$u[k] = x[k] + c \cdot \xi[k], \tag{2.13}$$

kde c určíme ze vztahu 2.12 podle hladiny požadovaného SNR, tedy buď 20, 10, 6 nebo 0 dB.

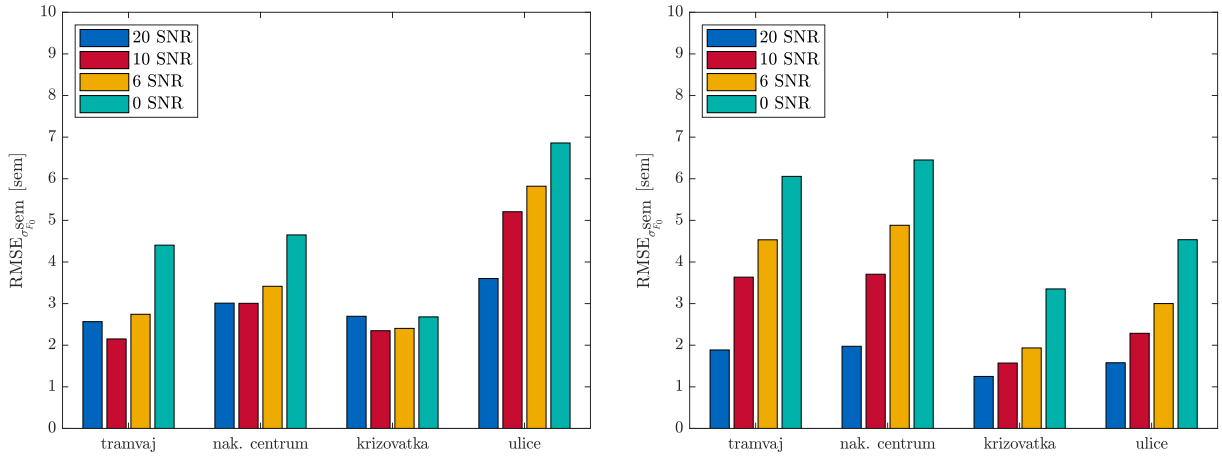
2.4.1 Výsledky

RMSE. Nejdříve prozkoumáme výsledky z hlediska RMSE. Ty získáme analogickým postupem jako v předchozí sekci u obrázku 2.2, jenom zde bude toto kritérium spočítáno pro každou promluvu 16x, jelikož máme 4 typy šumu a 4 úrovně SNR.

Výsledky jsou zobrazeny na obrázku 2.4. Potvrdilo se, z 2.4a, že Praat není přesný, jeho RMSE neklesne pod úroveň 2 semitóny, navíc se ukazuje jeho totální šumová neodolnost, kdy i při úrovni SNR 20 dB jsou hodnoty chyb velmi vysoké (okolo 2.5 semitónů).

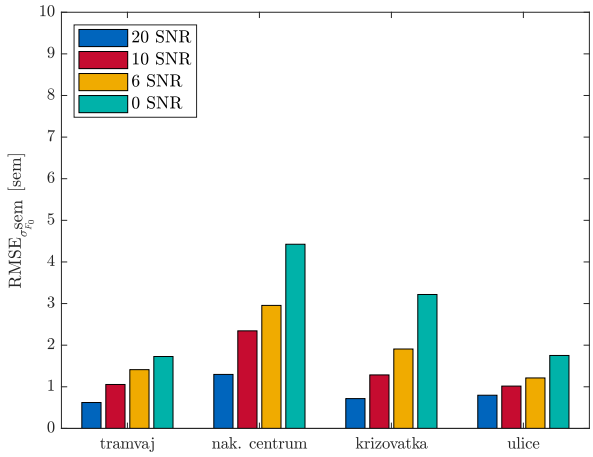
Překvapivé výsledky skýtá BaNa, na 2.4b, kdy i pro vysoké SNR se chyba poměrně dost zvětšila oproti čistým nahrávkám (rozdíl zhruba 0.5 semitónu pro SNR 20 dB a až 3 semitóny pro 10 dB). Pro nižší SNR jsou pak chyby opravdu velké (RMSE 4 semitóny a výše) a ukazuje se, že i když má být tento estimátor robustní i pro silný šum a v čistých nahrávkách byl poměrně přesný, v tomto případě to tak není a v přítomnosti šumu selhává.

Uspokojivé výsledky poskytuje SWIPE na 2.4c, který si pro vyšší a i některé nízké hladiny SNR udržuje svojí robustnost a chyby nejsou veliké (RMSE okolo jednoho semitónu). U určitých typů šumu pro nízká SNR však selhává i on (u nák. centra a křižovatky, RMSE jde až nad 3 semitóny), naštěstí u SNR 20 dB jsou výsledky stále rozumné.



(a) : Praat

(b) : BaNa



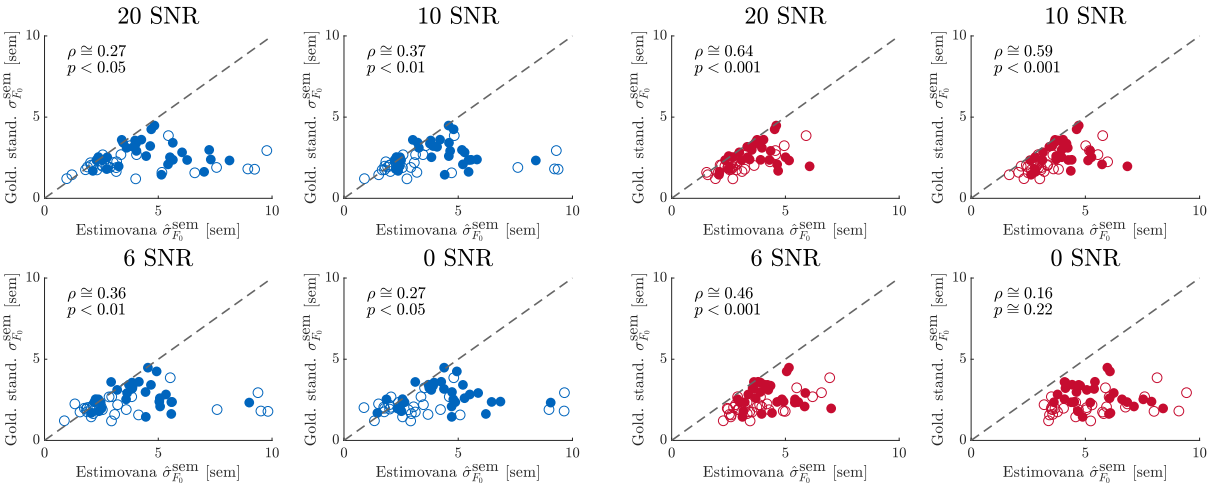
(c) : SWIPE

Obrázek 2.4: Hodnota RMSE odhadu estimátoru pro 60 promluv, zašuměné na čtyřech úrovních SNR pro 4 typy šumů.

Pro větší informaci prozkoumáme odhady pro zašuměné nahrávky v detailu, stejně jako v předchozí sekci u obrázku 2.1. Pro ilustraci uvedeme pouze výsledky pro typ šumu 4, tedy rušnou křižovatku, jako nejtypičtější reálné prostředí, obsahující velké množství typů hluku na pozadí. Na základě obrázku 2.4 budeme očekávat, že Praat bude mít velké chyby odhadů, zastoupené ve větší míře, BaNa nejspíše taky, ale mělo by jich být menší počet a méně signifikantní. SWIPE by měl být stále robustní až do SNR 6 dB, na nulové

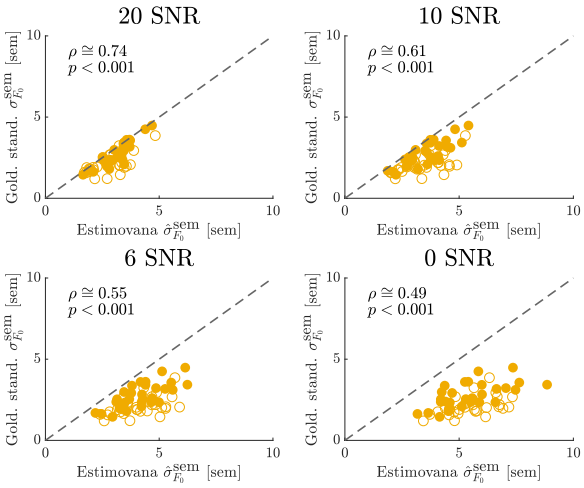
úrovni by měl selhat i ten.

Na ilustrativním obrázku 2.5 jsou zobrazeny výsledky. Pro každý estimátor jsou ukázány 4 grafy, odpovídající úrovni SNR. Šedivá úsečka reprezentuje pozice, kdy je odhad stejný jako referenční hodnota, tedy přesná shoda. U každého grafu je uvedena příslušná hodnota Spearmanova korelačního koeficientu r, spolu s p, značící pravděpodobnost nulové korelace za předpokladu nenulové. Snažíme se tedy o r co největší a p co nejmenší.



(a) : Praat

(b) : BaNa



(c) : SWIPE

Obrázek 2.5: Srovnání algoritmů na 60ti promluvách, zašuměných na úrovních 20, 10, 6 a 0 dB SNR pro šum typu 4, rušná křižovatka.

Pokud se podíváme na výsledky Praatu v 2.5a, vidíme, že náš předpoklad víceméně platí. Ukazuje se však fenomén, přítomný i u zbylých estimátorů, kdy více zašuměné nahrávky mají *větší úspěšnost* než ty méně zašuměné z pohledu korelačního koeficientu r. Konkrétně zde má nahrávka s úrovní SNR 10 a 6 dB r výrazně větší, než na SNR 20 dB ($r \cong 0.36$ a 0.37 oproti

0.27).

Estimátor BaNa na 2.5b má výsledky poměrně dobré, oproti pesimističtějšímu předpokladu. Způsobené to je hlavně tím, že BaNa má pro tento typ šumu nejlepší výsledky oproti jiným typům prostředí, patrné z RMSE na obrázku 2.4b. Pro SNR 20,10,i 6 dB se jedná stále o slušné výsledky, které se tolik neliší od úspěšnosti na čistých nahrávkách (obr. 2.2). Na nulové úrovni již selhává.

SWIPE na 2.5c vypadá dobře, kdy pro první úrovně SNR má skvělé výsledky $r \cong 0.74$ a 0.61 a pro nižší hladiny SNR se významněji nezhoršují. Jediná věc, která lehce zaráží je stejný fenomén zmíněný výše u Praatu. Úspěšnost odhadu v nahrávce na SNR 20 dB je vyšší, než pro čisté nahrávky, kdy bylo $r \cong 0.66$ (obr. 2.1c) oproti $r \cong 0.74$ jak je zde u SNR 20 dB.

Tento problém zjevně souvisí s šumovými vlastnostmi a bude podrobněji zkoumán v následujících sekcích, kdy se více zaměříme na průběh $\hat{F}_0$ během zašuměné promluvy. Je možné, že estimátor má velmi dobrou šumovou odolnost a při zašumění dojde ke zbavení se chybových odhadů, přítomných v čisté nahrávce. Stejně tak ale může pokládat šum za řečový segment a tím zkreslovat celkové výsledky.

Ukázalo se, že Praat i BaNa nejsou v zašuměném prostředí moc úspěšné, kdy BaNa měl uspokojivé výsledky alespoň v jednom případě pro šum rušné křižovatky (obr. 2.4b, 2.5b). SWIPE má výsledky pro zašuměné promluvy velmi slušné, kdy velkou chybovost obsahuje pouze na nízkých hladinách SNR u některých typů šumu. Bude ale nutné pečlivě prozkoumat vliv šumu na performanci estimátorů, kdy by mohly zásadně zkreslovat výsledky, pozorované na problému s lepší přesností pro více zašuměné nahrávky, popsaný výše.

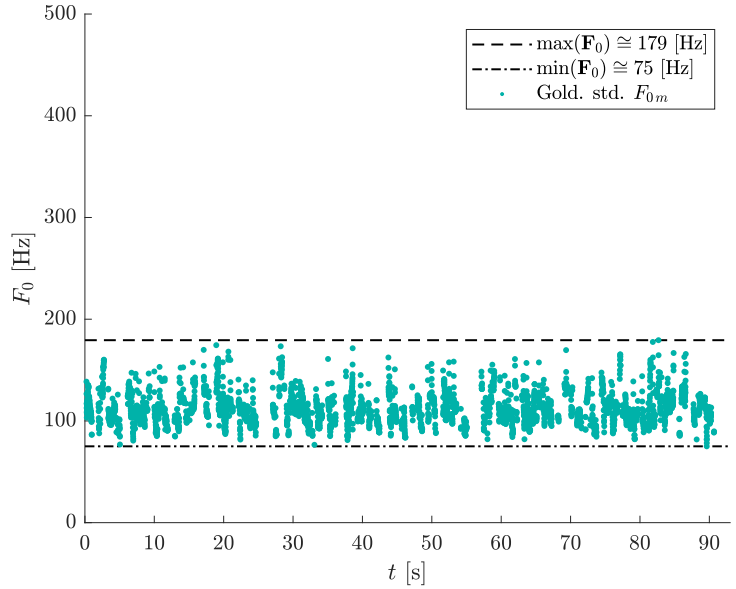
2.4.2 F_0 contour

V této sekci budeme zkoumat časový průběh odhadu $\hat{F}_0$ během promluvy, tzv. *contour*, pro čisté nahrávky i za přítomnosti šumu. Pokusíme se zdůvodnit, na čem estimátory chybují, jaké jsou zdroje jejich chyb co způsobující nepřesnosti, vyhodnocené v předchozích dvou částech.

Gold standard. Ilustrativně zde provedeme tento proces na jedné vhodně vybrané nahrávce, jelikož na zkoumání většího počtu promluv zde není prostor. Byl vybrán jeden z nejhorších případů, kdy pro tuto konkrétní promluvu nebyl žádný z estimátorů blízko Gold standardu, i pro čisté promluvy. Konkrétně hodnota reference $\sigma_{F_0}^{\text{sem}} \cong 2.3$ a odhady $\hat{\sigma}_{F_0}^{\text{sem}} \cong 2.9, 3.9, 5.1$ pro SWIPE, Praat a BaNa popořadě, pro *čisté* promluvy. Příslušný RMSE je 0.6, 1.6 a 2.8 respektive.

Jednalo se o mužský hlas zdravého člověka. Referenční průběh F_0 (Gold standard), je na obrázku 2.6. Každý bod v grafu reprezentuje hodnotu hlasivkové frekvence v Hertzích pro daný časový segment. V místech, kde řeč není přítomna, tedy např. pauzy, nádechy a některé neznělé hlásky není hodnota F_0 pro tento segment určena.

Můžeme pozorovat, že řečový průběh má velmi jasně vymezené pásmo frekvencí, dané rozdělením F_0 u (zde mužské) lidské řeči. Pokud by se jednalo o ženskou či dětskou mluvu, frekvence by s velkou pravděpodobností byly vyšší a celkové pásmo širší, ale trend bude podobný. Této vlastnosti můžeme využít při vyhodnocování průběhu odhadu $\hat{F}_0$ z estimátorů, kdy stanovíme počet hodnot *mimo* tento pás, tzv. „outliers“, daný Gold standardem.



Obrázek 2.6: Referenční časový průběh F_0 z vyznačenými úrovněmi maximální a minimální hodnoty. m značí daný časový segment, $\mathbf{F}_0$ celý průběh pro všechny úseky.

Odhad časového průběhu z estimátorů. Na obrázku 2.7 jsou vidět časové průběhy odhadu $\hat{F}_0$ pro příslušné estimátory. Jedná se o stejnou promluvu jako na obr. 2.6, se kterým budeme výsledky porovnávat. Na 2.7 jsou také zobrazeny hladiny maximální (čárková čára) a minimální (čára s čárkami a tečkami) hodnoty F_0 pro tuto promluvu, určené z Gold standardu.

Časové průběhy byly z ilustrativních důvodů vykresleny pro čistou promluvu a zašuměnou na 20, 10 a 6 dB SNR. Nulová úroveň byla vynechána jelikož průběh ztrácel svojí čitelnost a bylo těžké chyby pečlivě analyzovat. Také zde není takový dostatek prostoru.

Typ šumu byl zvolen stejně jako v předchozí části při konstrukci obrázku 2.5, tedy rušná křižovatka. Analýza časových průběhů na obrázku 2.7 probíhala

tak, že jsme se vždy zaměřili na časový úsek odhadu $\hat{F}_0$ obsahující větší množství chyb a jemu odpovídající úsek nahrávky a hledali možné příčiny chybného odhadu.

Estimátor Praat na obrázku 2.7a má evidentně velké chyby pro čistou i zašuměnou promluvu. Většinou se jedná o delší časové úseky, na kterých se estimátor evidentně chytá neřečových prvků v promluvě (nádechy, některé neznělé hlásky, mlasknutí apod.), jak je tomu např. v časovém segmentu okolo 20ti a 50ti sekund pro čistou promluvu (na 2.7a). Za přítomnosti šumu u nižších hodnot SNR je patrná decimace průběhu, kdy algoritmus vyhodnocuje více časových úseků jako neřečové a odhad $\hat{F}_0$ není použit. Chyby jsou stále přítomné ve velké míře, kdy estimátor mylně odhaduje hlasivkovou frekvenci pro směs vyšších frekvencí v šumu a neřečových prvků (např. časový segment 65 - 90 s pro 10 a 6 dB SNR u obrázku 2.7a).

V jaké míře jsou v časovém průběhu přítomny chyby způsobené odhady $\hat{F}_0$ pro nízké frekvence šumu nejsme schopni z obrázku 2.7 dobře rozeznat a problém bude analyzován později.

Estimátor BaNa, jehož časový průběh odhadu hlasivkové frekvence je na obrázku 2.7b, má evidentně chyby zastoupené v menší míře než Praat. Jsou ojedinělé, v rozsahu pár desítek milisekund, stále je jich je však velký počet. Příčina chyb je velmi podobná jako u eastimátoru Praat, BaNa je však robustnější vůči šumu.

SWIPE měl výsledky opět nejlepší. Jeho časový průběh odhadu $\hat{F}_0$ je na obrázku 2.7c. Obsahuje charakterově podobné chyby jako BaNa, je jich ale mnohem méně. Také je evidentní velmi dobrá šumová odolnost, výsledky se po zašumění zhorší jen nepatrně.

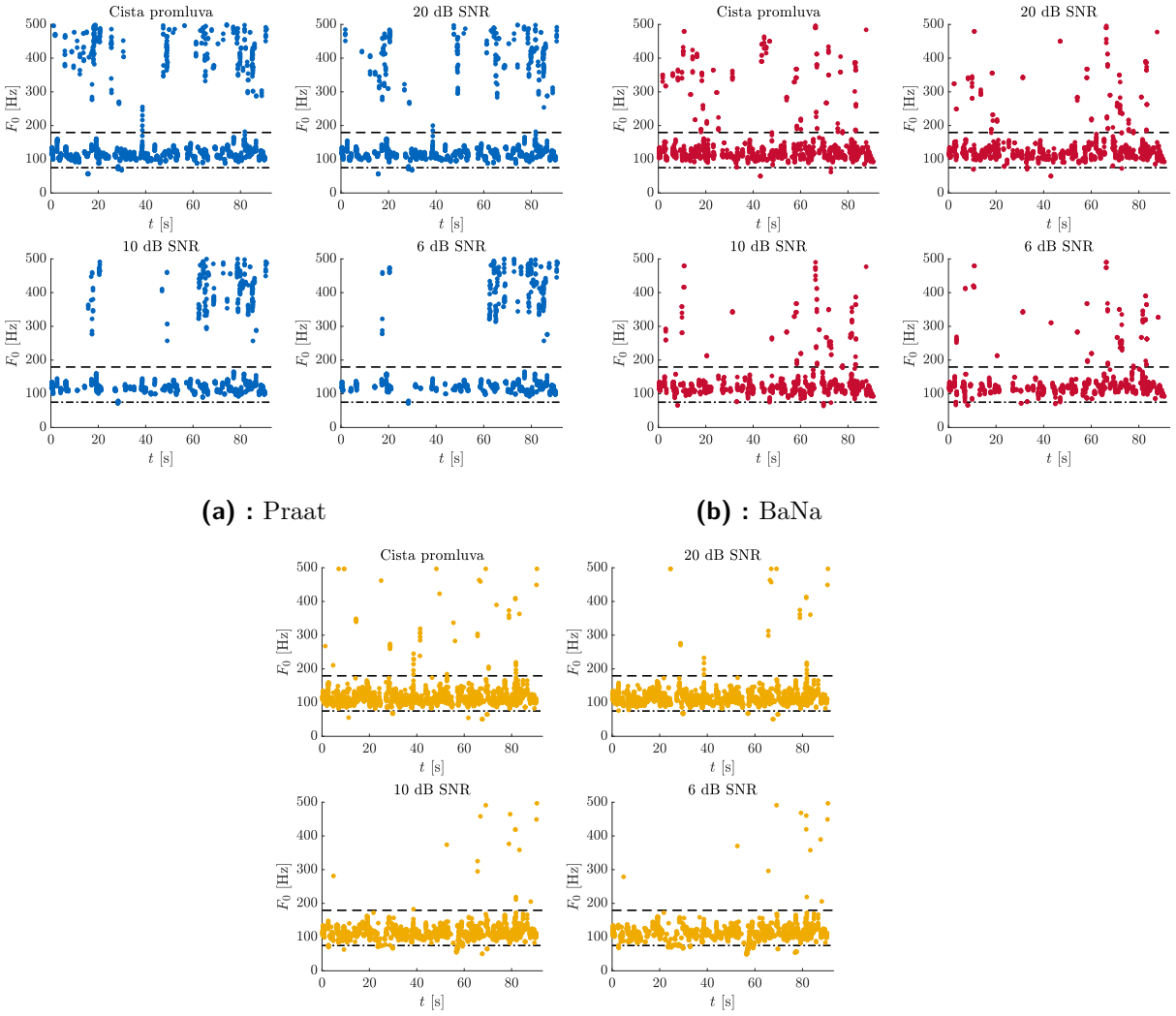
V tabulce 2.3 jsou uvedeny počty „outlierů“, tedy hodnot odhadu $\hat{F}_0$ ležící mimo pás vymezený Gold standardem na obrázku 2.6. Potvrdil se trend úspěšnosti estimátorů mezi sebou (SWIPE, BaNa, Praat), s výjimkou čisté promluvy, kdy byl Praat úspěšnější než BaNa. Je také evidentní, že počet „outlierů“ $\hat{F}_{0\,\mathrm{out}}$ v mnoha případech při nižším SNR klesá. Například u BaNa 210 pro čistou mluvu, 150 pro 20 dB SNR a 130 pro 10 dB SNR.

Jak již bylo zmíněno v sekci 2.3.1, může to být způsobené dvěma důvody. Estimátory jsou buď velmi robustní na šum a „outliera“ v čisté promluvě, k jehož příslušnému časovému segmentu je následně přidám šum, prohlásí za neřečový segment a odhad se neurčuje, či nevyužije, nebo jsou výsledky ovlivněny nízkými frekvencemi šumu, které mohou chybové hodnoty „stahovat“ dolů, kde jsou pak k nerozeznání od správných odhadů. Případná přítomnost druhého problému by byla pro nás velice nepříjemná. V následující sekci se tedy pokusíme tuto záležitost prozkoumat.

Po zhodnocení výsledků této části také klasifikujeme estimátor Praat jako nedostatečně přesný a málo šumově robustní. V porovnání se zbylými dvěma algoritmy obsahoval řádově větší chybovost a v dalších částech již proto není jeho průběh dále analyzován.

$\hat{F}_{0\,\mathrm{out}}$	**Praat**	**BaNa**	**SWIPE**
čistá prom.	134	210	75
20 dB SNR	325	150	50
10 dB SNR	264	130	72
6 dB SNR	262	130	133

Tabulka 2.3: Tabulka ukazující zastoupení „outlierů“, tedy hodnot ležící mimo pás, vymezený na obrázku 2.6 pro čistou i zašuměnou nahrávku.



(a) : Praat

(b) : BaNa

(c) : SWIPE

Obrázek 2.7: Časový průběh odhadu $\hat{F}_0$ z estimátorů pro čistou náhrávku a zašuměné, na 20, 10 a 6 dB SNR. Vyznačené meze určují maximální a minimální hodnotu referenčního průběhu F_0, určené na obr. 2.6.

2.4.3 Analýza dopadu přítomnosti šumu na časový průběh odhadu $\hat{F}_0$

V této sekci prozkoumáme fenomén, zjištěný v předchozích sekcích, kdy má estimátor v některých případech „lepší“ výsledky pro nižší hladiny SNR. Tato neshoda je nejvíce patrná z výsledků, uvedených v tabulce 2.3. Chceme otestovat, zdali tomu tak je díky velmi dobré šumové robustnosti, nebo zavádějících mylných odhadů pro nízké frekvence, obsažených v šumu.

První přístup byl takový, že jsme estimátory otestovali *pouze* na šumové nahrávky. Pro ilustraci byl vybrán šum z prostředí rušné křižovatky. Průběh odhadu $\hat{F}_0$ pro tuto nahrávku od estimátoru SWIPE a BaNa je na obrázku 2.8 spolu s výběrovou střední hodnotou $\hat{\mu}$ a směrodatnou odchylkou $\hat{\sigma}_{F_0}$ v Hertzích i semitónech.

Z levého grafu pro SWIPE, na 2.8a, je evidentní, že tento estimátor má velký problém pro nízké frekvence v šumu, které považuje za řečové segmenty. Z výběrové střední hodnoty $\hat{\mu}$ je vidět, že většina určených hodnot odhadů je na hodnotách, které má samotná promluva (mužská) a tím výrazně zkresluje výsledky. Šumy nízké frekvence v těchto nahrávkách představuje např. jedoucí tramvaj nebo rozjíždějící se auta.

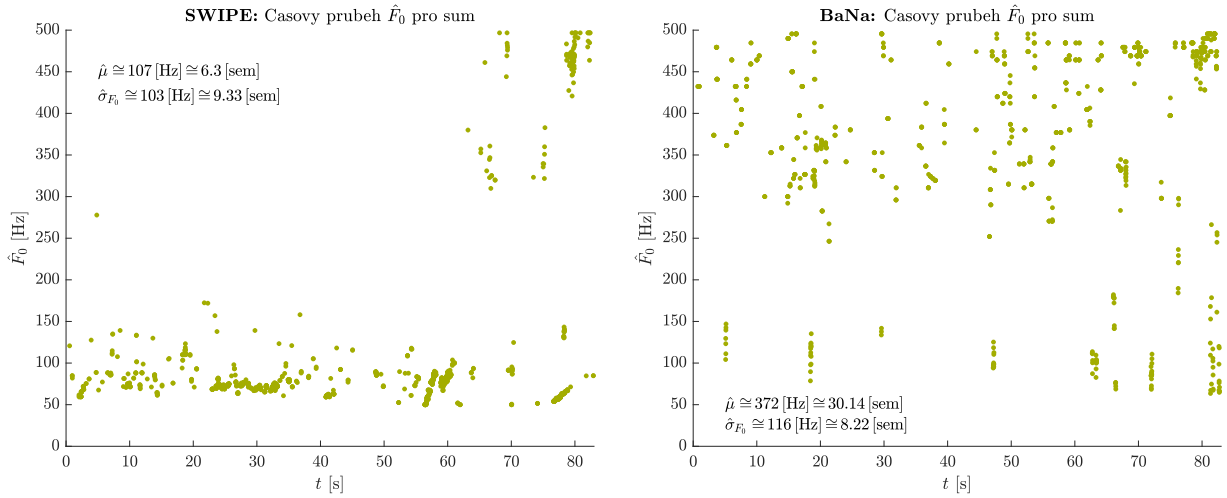
Oproti tomu BaNa, z obrázku 2.8b, má chyby odhadu na vyšších frekvencích ($\hat{\mu} \cong 372$ Hz) a počet špatně určených odhadů $\hat{F}_0$ na nižších je zanedbatelný. To sice zavádí nepřesnosti do výsledků, které ale jsou pozorovatelné v časových průbězích F_0. Ve výsledné zašuměné nahrávce jdou z průběhu rozlišit chybně určené odhady od těch správných, což u SWIPE nemůžeme.

Pitch strength. Naštěstí je tento problém řešitelný změnou jednoho ze vstupních parametrů SWIPE, nastavující práh pro tzv. „pitch strength“, viz sekce 1.2.3. Pokud má odhad $\hat{F}_0$ pro daný časový segment tuto hodnotu nižší než stanovený práh, je segment hodnocený jako neřečový.

Nový časový průběh pro stejnou šumovou nahrávku s pozměněným parametrem hodnoty prahu pro „pitch strength“ je na obrázku 2.9. Je vidět, že změnou nám chybné odhady $\hat{F}_0$ šumových segmentů téměř vymizely a řeší se tím problém zkreslování výsledků nízkými frekvencemi šumu u SWIPE. V v průběhu odhadu na 2.9 je oproti 2.8a s nezměněným parametrem pouze 4.25% hodnot.

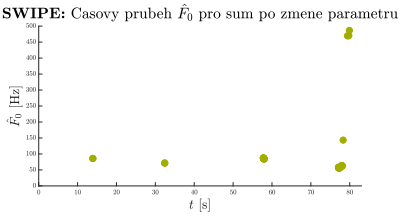
Nabízela by se možnost provést analogický proces pro BaNa a zbavit se chyb způsobených šumem na vyšších frekvencích a estimátor tak zpřesnit. Bohužel, tato změna časový průběh odhadu $\hat{F}_0$ výrazně decimuje, určených hodnot je velmi málo. Takový průběh již není příliš objektivní na určování směrodatné odchylky, našeho pozorovacího parametru.

Na ilustrativním obrázku 2.10 je znázorněn průběh odhadů $\hat{F}_0$ SWIPE pro původní a změněnou hodnotu prahu pro „pitch strength“ (na (2.10b resp. 2.10c)), spolu s Gold standardem (2.10a). Vybraná nahrávka je zašuměná



(a) : SWIPE **(b) :** BaNa

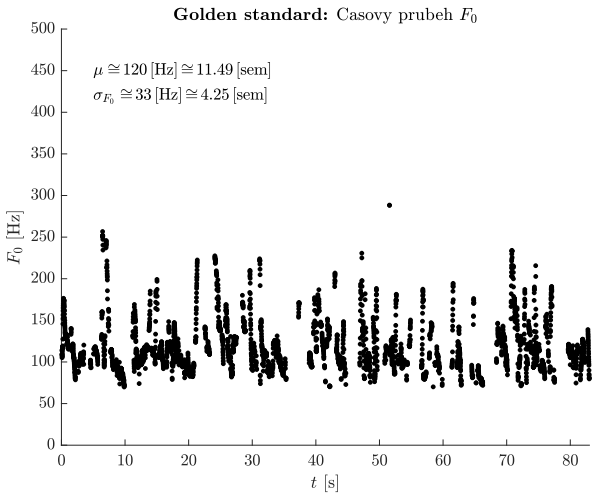
Obrázek 2.8: Časový průběh odhadu $\hat{F}_0$ šumové nahrávky, tedy neobsahujících žádnou přímou mluvu. Šum je z rušné křižovatky. V grafem jsou také uvedeny hodnoty výběrové střední hodnoty $\hat{\mu}$ a směrodatné odchylky $\hat{\sigma}_{F_0}$ v Hertzích i semitónech.



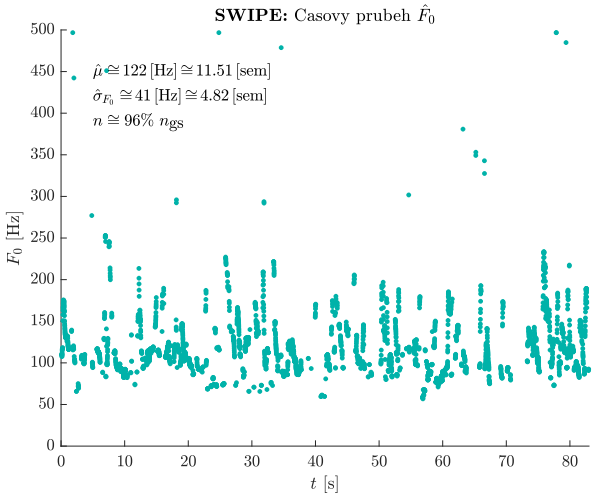
Obrázek 2.9: Časový průběh odhadu $\hat{F}_0$ pouze šumové nahrávky pro estimátor SWIPE po změně hodnoty prahu pro „pitch strength“.

pro hladinu 10 dB SNR, kdy šum je z rušné křižovatky. Je vidět, že změnou se značně zdecimuje počet hodnot (56% oproti 96% počtu hodnot v Gold. standardu), stále je však průběh objektivní, jelikož jeho uvedené parametry (zejména výběrová směrodatná odchylka $\hat{\sigma}_{F_0}$) jsou velmi blízko referenčním. Zároveň jsme se tak zbavili zavádějících chybných odhadů pro nízké frekvence šumu.

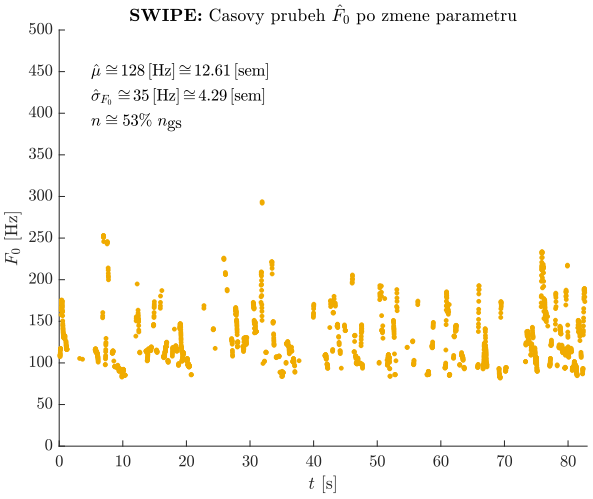
Důvod, proč zavádíme změnu vstupního parametru až nyní je, že v části 2.3.1 kdy jsme ladili vstupní parametry estimátorů nebyl přítomen šum a tedy jsme tento aspekt nemohli obsáhnout. Problém byl objeven až v části 2.4.1, kdy se již se šumy nepracovalo.



(a) : Gold standard



(b) : SWIPE



(c) : SWIPE s upraveným parametrem

Obrázek 2.10: Časové průběhy Gold standardu a odhadů $\hat{F}_0$ est. SWIPE pro původní a změněnou hodnotu prahu „pitch strength“. Nahrávka je zašuměná na hladině 10 dB SNR. Jsou uvedeny hodnoty výběrové střední hodnoty a směrodatné odchylky v Hertzích i semitónech, spolu s počtem odhadnutých $\hat{F}_0$.

2.4.4 Filtrace časového průběhu odhadu $\hat{F}_0$

I když to není v zadání této práce, pokusíme v této sekci jen velmi zběžně prozkoumat a navrhnout možnou jednoduchou filtraci časového průběhu $\hat{F}_0$, například zbavení se „outlierů“.

První přístup byl uplatnit na průběh mediánový filtr. Při pozorování průběhu byly výsledky poměrně uspokojivé, naprostá většina „outlierů“ zmizela a průběh se vyhladil. Bohužel, tento přístup se ukázal jako nevhodný při následném počítání výběrové směrodatné odchylky, našeho hlavního ukazatele. Odhadnuté hodnoty $\hat{\sigma}_{F_0}^{\text{sem}}$ z filtrovaného průběhu byly značně nižší, než zjištěné referenční $\sigma_{F_0}^{\text{sem}}$, pro všechny případy. Tento problém přetrvával i při použití nižších řádů filtru. Tento přístup tedy nevyužijeme.

Při pozorování časového průběhu estimátorů na obrázku 2.7 se nabízí možnost filtrovat „outliery“ stanovením frekvenčních mezí průběhu a hodnoty mimo tento pás neuvažovat. Jelikož se v praxi často rozdělení F_0 v průběhu mluvy aproximuje normálním [27], určíme šířku tohoto pásu $2\hat{\sigma}_{F_0}$ okolo výběrové střední hodnoty $\hat{\mu}_{F_0}$. Dále budeme tento pás označovat jako 1σ. Tento přístup měl o něco lepší výsledky než pás daný středem v mediánu časového průběhu, ohraničený prvním a třetím kvartilem.

Při testech na RMSE se ukázalo, že tento přístup má dobré výsledky u estimátoru BaNa. U algoritmu Praat, se kterým již nepočítáme, byl tento pás natolik široký, že obsahoval i velké množství „outlierů“. Šířka pásu byla daná velkým počtem chyb na vyšších frekvencích. U estimátoru SWIPE je již časový průběh $\hat{F}_0$ po změně vstupního parametru hodnoty prahu „pitch strength“ (viz předchozí část 2.4.3) natolik přesný, že tento způsob filtrace nepřináší významně lepší výsledek. Oříznutí průběhu pásem 1σ tedy budeme provádět pouze u BaNa.

2.4.5 Výsledky pro SWIPE a BaNa po filtraci a úpravě vstupních parametrů

V této sekci určíme finální podobu výsledků, které jsme již určovali v kapitole 2.4.1. V předchozích sekcích, kdy jsme se detailně zabývali působením aditivního šumu na nahrávky, jsme zjistili, že SWIPE má problémy z nízkými frekvencemi šumu, zkreslující výsledky. V části 2.4.3 jsme tento problém vyřešili změnou hodnoty prahu pro „pitch strength“ daného odhadu $\hat{F}_0$. Dále jsme v sekci 2.4.4 otestovali, že jednoduchou filtrací časového průběhu $\hat{F}_0$ od BaNa oříznutím v pásu 1σ dosáhneme lepších výsledků z hlediska RMSE.

Estimátor Praat byl v závěru části 2.4.2 klasifikován jako nedostatečně přesný a pro naší aplikaci nevhodný. V těchto výsledcích již tedy uveden není.

Na obrázku 2.11 jsou zobrazeny výsledky, z hlediska Spearmanovy korelace, pro šum rušné křižovatky. U SWIPE, na 2.11a je vidět až extrémně přesná shoda s referenční hodnotou, minimálně pro 20 a 10 dB SNR ($r > 0.9$). Pro 6 dB je přesnost také velmi dobrá, $r > 0.8$. U nulové hladiny je již horší, $r \cong 0.65$, nicméně na těžké podmínky při této úrovni je to stále dosti dobrý výsledek. Pozorujeme, že se oproti obrázku 2.5c přesnost značně zvýšila.

Estimátor BaNa, na obr. 2.11b, si již tak dobře nevedl. Ukazuje se, že jednoduchá filtrace časového průběhu odhadu $\hat{F}_0$, popsaná v sekci 2.4.4, příliš nepomohla. Naopak, sice uvidíme, že RMSE je nižší než předtím, nicméně korelace s Gold standardem je nižší. Vidíme to např. u 20 dB SNR, kdy bez úpravy byla hodnota $r \cong 0.64$ (z obr. 2.5b) a zde je $r \cong 0.51$.

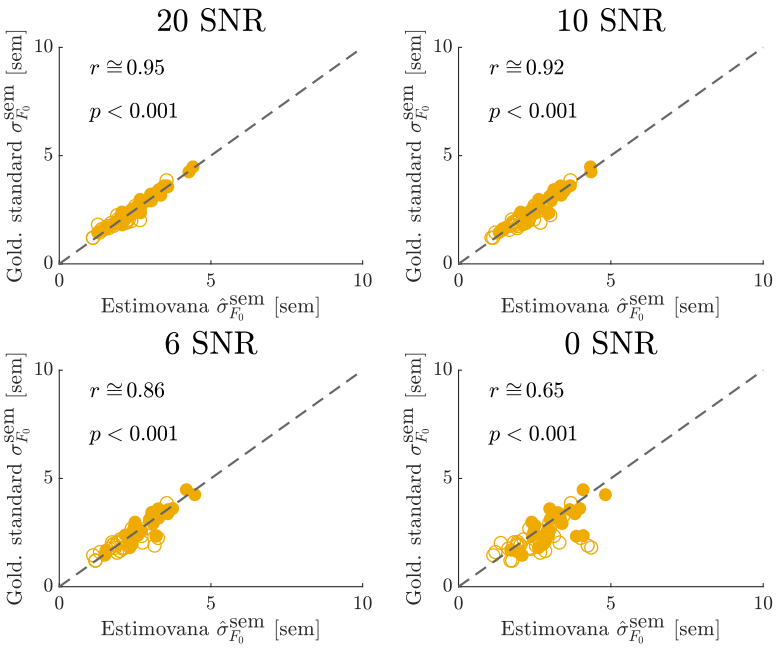
Pokud bychom s BaNa pracovali dále, použijeme tedy nejspíše jiný typ post-processingu, kdy bychom se na tento problém podívali detailněji. Již teď je ale zřejmé, že SWIPE je natolik přesný, že na zjištění $\sigma_{F_0}^{\text{sem}}$ v praxi, použijeme pouze tento algoritmus.

RMSE. Na obrázku 2.12 jsou zobrazeny výsledky z hlediska RMSE. Potvrzují se závěry zmíněné výše o velmi dobré přesnosti SWIPE. Jediný typ šumu, kde je estimátor o třídu horší je nákupní centrum, viz. obr.2.12, s RMSE nad 2 semitóny pro 6a 0 dB, všechny ostatní šumy mají výsledky podobné, kdy si estimátor zachovává velmi dobrou šumovou robustnost (RMSE okolo 1 sem. pro 20 a 10 dB a okolo 1.5-2 sem. pro 6 a 0 dB SNR). U případu nákupního centra je SWIPE přesný alespoň pro vyšší hodnoty SNR (pro 20 dB SNR zhruba 1 sem). Porovnáním s výsledky na obrázku 2.4c vidíme, že se RMSE značně zmenšilo.

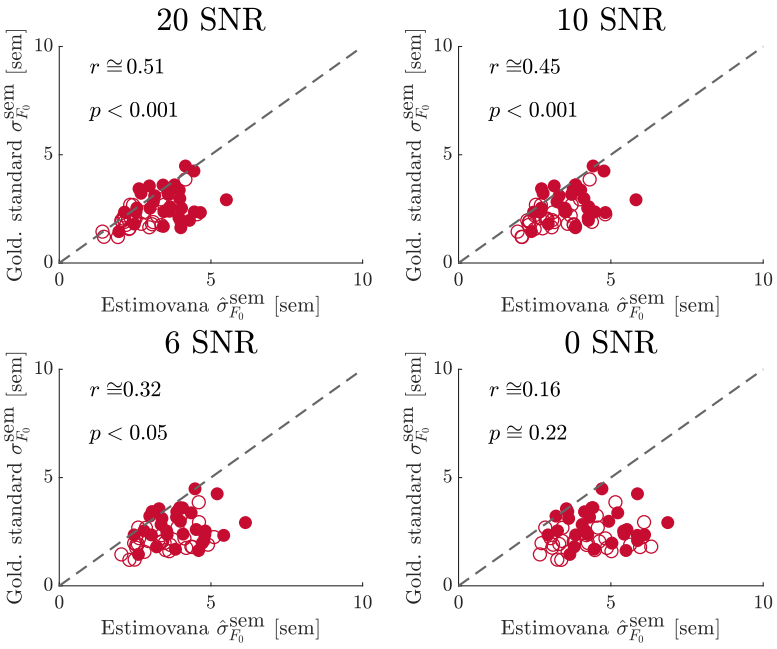
BaNa, na obr. 2.12b, má také lepší výsledky než na 2.4b, nicméně hodnota r byla v tomto případě nižší, viz výše. Hodnoty pro vysoká SNR jsou poměrně dobré (RMSE < 1 pro 20 dB), nicméně oproti SWIPE (RMSE < 0.3) stále o třídu horší.

Závěr. Obrázek 2.11a, spolu s 2.12a, jsou patrně nejvýznamnější z celé práce. Ukazují, že SWIPE s upraveným vstupním parametrem prahu „pitch strength“ oproti sekci 2.4.1, má velmi vysokou přesnost z hlediska Spearmanovy korelace a RMSE a je robustní i pro šum na nižších úrovních SNR. Takovýto estimátor by se již dal použít na případnou klinickou praxi, kdy zkoumáme $\sigma_{F_0}^{\text{sem}}$ promluv pacientů.

S estimátorem BaNa již nepočítáme, jelikož se nám ani jednoduchým post-processingem nepodařilo výsledky významně zlepšit. Z celé této sekce tedy vychází jasný vítěz SWIPE, jako kandidát do možného výzkumu detekce Parkinsonovy choroby pomocí chytrých telefonů.

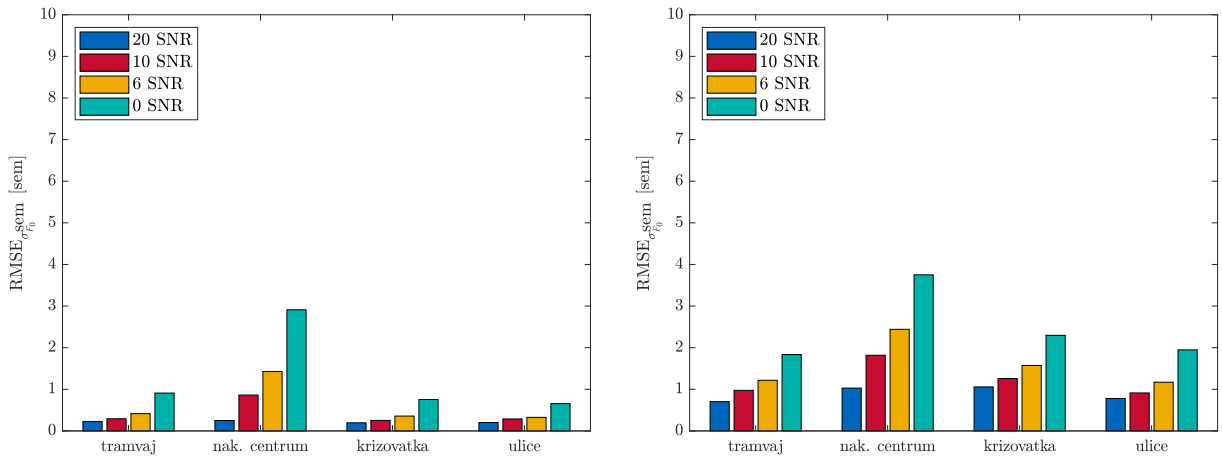


(a) : SWIPE



(b) : BaNa

Obrázek 2.11: Srovnání estimátoru SWIPE a BaNa na 60ti promluvách, zašuměných na úrovních 20, 10, 6 a 0 dB SNR pro šum typu 4, rušná křižovatka. Algoritmy byly oproti výsledkům na obrázku 2.5 modifikovány podle sekce 2.4.5.



(a) : SWIPE **(b) :** BaNa

Obrázek 2.12: Hodnota RMSE odhadu estimátoru pro 60 promluv, zašuměné na čtyřech úrovních SNR pro 4 typy šumů. Algoritmy jsou oproti výsledkům na obr. 2.4 modifikovány podle sekce 2.4.5.

2.5 Odhad SNR na nahrávkách s reálným šumem

V této sekci se zaměříme na prozkoumání nahrávek, pořízených v reálném prostředí na chytrý telefon. Konkrétně nás bude zajímat, jak jsou algoritmy obsažené v telefonu schopné potlačovat šum na pozadí. Hlavní důvod tohoto zkoumání je zjistit, zda je vůbec nutné aby měly estimátory dobré výsledky i pro nízká SNR, když reálně nebude vůbec možné těchto hladin dosáhnout.

Nahrávky byly pořízené na těch samých místech jako se nahrávaly aditivní šumy, přidávané uměle do čistých nahrávek. Odhad SNR probíhal velmi jednoduchou metodou, kdy byly vybrány krátké úseky (okolo 1.5 sekundy) s mluvou a podobně dlouhé bez mluvy (např. v pauzách), kde je přítomen pouze šum na pozadí. Jelikož jsou v řečovém segmentu přítomny obě složky, určíme odhad SNR, $\hat{\mathrm{SNR}}$, jako

$$\hat{\mathrm{SNR}} = 10 \log_{10} \left(\frac{P_{\mathrm{mluva}} - P_{\mathrm{sum}}}{P_{\mathrm{sum}}} \right) \quad [\mathrm{dB}], \tag{2.14}$$

kde P_{mluva} a P_{sum} je průměrný výkon pro mluvený a šumový segment, spočítaný podle vztahu 2.10.

Nahrávky byly pořízeny dvěma telefony, Sony Xperia Z1 a Nokia 6, aby mohl být patrný jejich případný rozdíl. Použita byla aplikace [29]. Prostředí byly rušná křižovatka, ulice s větším počtem lidí a vnitřek jedoucí tramvaje (oba telefony). Tedy šumu 1, 2 a 3 popsané na začátku sekce 2.4. Pro každý záznam bylo SNR odhadnuto dvakrát, z časových segmentů na rozdílných místech.

V tabulce 2.4 jsou uvedeny výsledky. Je vidět, že díky velké nestacionaritě šumů na pozadí se hladiny SNR můžou během promluvy měnit, někdy i velmi výrazně (např. na 4. řádku - Tramvaj (Sony Xperia)). Na nulovou úroveň se nedostaneme nikdy, v některých případech jsme ale poměrně blízko (1.70 dB). Nicméně většina hodnot se pohybuje mnohem výše, kolem 5ti, 18ti i za 20ti dB SNR, velmi závisí na aktuální situaci na pozadí v jednotlivých prostředích.

Konkrétní rozdíl v obou telefonech přesně stanovit nemůžeme, je však evidentní, že nahrávání funguje odlišně. U telefonu Nokia 6 je úroveň SNR nižší, zároveň však nevykazuje moc velké odchylky. Sony Xperia Z1 tlumí šum více, občas ale tento proces evidentně nezvládne tak dobře (hodnota 1.70 dB SNR u tramvaje).

Důležitý závěr z této sekce je, že většina hodnot odhadnutého SNR leží v mezích, kdy estimátor SWIPE je schopen fungovat stále velmi přesně (z obrázků 2.12a a 2.11a). Nemusíme se tedy zabývat jeho účinností na extrémních hodnotách SNR, jako je nulová hladina a níže.

	$\hat{\mathrm{SNR}}_1$ [dB]	$\hat{\mathrm{SNR}}_2$ [dB]
Rušná křižovatka (Nokia 6)	7.16	3.56
Tramvaj (Nokia 6)	4.12	16.90
Tramvaj (Sony Xperia)	1.70	25.49
Ulice (Sony Xperia)	18.81	18.87

Tabulka 2.4: Hodnoty hladin odhadu $\hat{\mathrm{SNR}}$ pro nahrávky pořízené v rušném prostředí.

Kapitola 3

Závěrečné zhodnocení

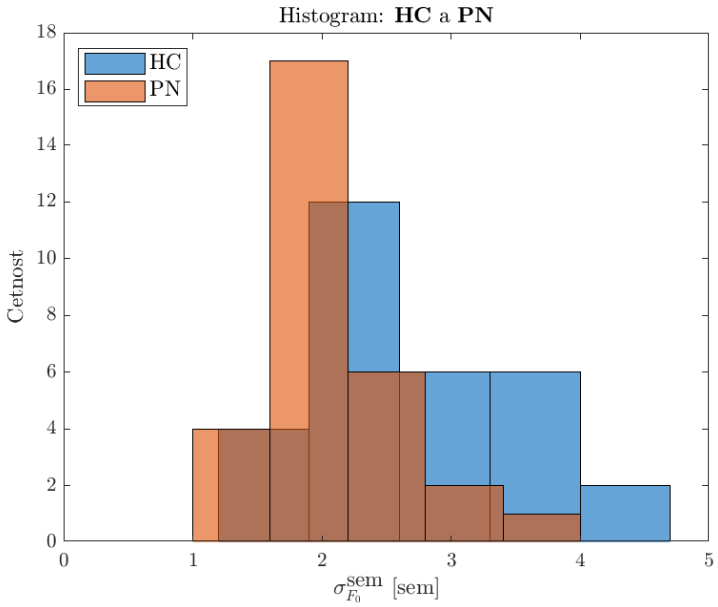
V předchozí kapitole jsme se zabývali testováním jednotlivých estimátorů. Z konečných výsledků vyšel nejlépe algoritmus SWIPE, především na základě hodnot RMSE na obr. 2.12a a korelace s Gold standardem na obr. 2.11a. Tento estimátor tedy navrhujeme jako kandidáta na zkoumání směrodatné odchylky F_0 z nahrávek řeči, pořízené na chytrý telefon v reálném prostředí.

V reálné aplikaci by se případná diagnóza určovala tak, že se sestaví rozdělení odhadnutých $\hat{\sigma}_{F_0}^{\text{sem}}$ za delší časový úsek (několik měsíců až roky) a toto se bude porovnávat s rozdělením, jaké mají zdraví lidé. Pokud budou výrazně odlišná, je možná přítomnost nemoci.

Tento klíčový aspekt zde zkusíme přibližně otestovat, kdy budeme zkoumat zda-li přítomnost šumu v nahrávkách výrazněji neovlivní rozdělení $\hat{\sigma}_{F_0}^{\text{sem}}$ tak, aby z toho pak již nešlo věrohodně určovat rozdíly mezi rozdělením u pacientů s PN a zdravých jedinců. Toto budeme provádět nepárovým t-testem dvou výběrů.

3.1 T-test pro výsledky zdravých lidí a pacientů s PN

Na obrázku 3.1 je zobrazen histogram odhadnutých hodnot $\sigma_{F_0}^{\text{sem}}$ Gold standardu pro zdravé (HC) a pacienty s Parkinsonovou nemocí (PN). Pro každou skupinu bylo analyzováno 30 promluv. Na vodorovné ose je hodnota směrodatné odchylky v semitónech a na svislé je četnost, počet odhadů, ležících ve vymezené oblasti na vod. ose.



Obrázek 3.1: Histogram odhadnutých hodnot $\sigma_{F_0}^{\text{sem}}$ Gold standardu pro zdravé (HC) a pacienty s Parkinsonovou nemocí (PN).

Je zde vidět, že rozdělení jsou velmi odlišná. Konkrétně pacienti s PN mají tendenci mít $\sigma_{F_0}^{\text{sem}}$ menší, což souvisí s malou flexibilitou hlasivkových svalů a velké řečové monotónnosti (viz sekce 1.1.2). Na rozdělení zobrazené na grafu 3.1 se toto projeví zejména *odlišnou střední hodnotou pro oba výběry.*

Budeme zkoumat, zdali i odhady $\hat{\sigma}_{F_0}^{\text{sem}}$ od estimátoru SWIPE pro čisté a zašuměné nahrávky budou vykazovat podobné výsledky a rozdělení HC a PN půjde jednoznačně rozlišit.

Na posouzení, jak jsou jednotlivá rozdělení odlišná, použijeme nepárový t-test dvou nezávislých výběrů. Nulová hypotéza je, že střední hodnoty jsou pro oba stejné, $\mu_1 = \mu_2$. První výběr, $\mathbf{x}$, bude představovat hodnoty $\hat{\sigma}_{F_0}^{\text{sem}}$ zdravých lidí a druhý výběr, $\mathbf{y}$, pacientů s Parkinsonovou chorobou. Oba jsou stejně velké. Je evidentní, že tyto výběry jsou navzájem nezávislé, jedná se o různé lidi a různé podoby nahrávek. Další předpoklad t-testu, který musíme ověřit, je zdali výběry pochází z normálního rozdělení. Toto budeme ověřovat pomocí *Pearsonova* χ^2 *testu.*

Pearsonův χ^2 test

Tento test použijeme k ověření nulové hypotézy, že oba výběry, $\mathbf{x}$ a $\mathbf{y}$ pochází z normálního rozdělení oproti opaku, tedy že z něj nepochází.

Data ve výběrech jsou nejprve roztříděna do k stejně velkých sekcí, podobně jako u obrázku histogramu na 3.1. Zde byla zvolena jako optimální hodnota $k = 8$. Statistika má pak tvar

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{3.1}$$

kde O_i je *pozorovaná* četnost a E_i je *předpokládaná* četnost pro i-tou sekci. E_i se určí jako

$$E_i = n\left(F(Y_\mathrm{h}) - F(Y_\mathrm{d})\right), \tag{3.2}$$

kde F je distribuční funkce normálního rozdělení, hodnoty Y_h, Y_d jsou horní a dolní mez pro i-tou sekci a n je počet vzorků ve výběru, tedy zde $n = 30$.

Výslednou hodnotu statistiky, vypočítané ze vzorce 3.1, následně porovnáme s příslušným kvantilem χ^2 rozdělení s $k - (p_\mathrm{param} + 1)$ stupni volnosti, kde se do k počítají jen neprázdné sekce, obsahující vzorky a p_param značí v našem případě počet parametrů normálního rozdělení, které musíme z výběrů odhadnout. Tedy $p_\mathrm{param} = 2$ (výběrová střední hodnota a směrodatná odchylka). Dále je také kvantil daný hladinou významnosti α, zde stanovenou jako $\alpha = 0.05$. V případě, kdy je hodnota statistiky větší, než tento kvantil, nulovou hypotézu zamítáme na hladině významnosti α. V opačném případě jí nezamítáme, ovšem nepotvrzujeme.

Společně se s hodnotou statistiky je určena také hodnota p, značící pravděpodobnost pozorovaných výsledků za předpokladu, že platí nulová hypotéza. Čím je p menší, věříme jí stále méně. Zamítnutí nulové hypotézy nastane v případě, kdy je $p < \alpha$.

Pearsonův χ^2 test byl proveden pro všech 36 výběrů (čtyři typy šumu na čtyřech úrovních SNR plus navíc Gold standard a čisté nahrávky, celé pro HC i PN skupinu). Pouze ve *dvou případech* (tedy zhruba v 5.5%) jsme byli nuceni nulovou hypotézu o normálním rozdělení zamítnout. Jednalo se o jeden případ na SNR 6 dB, kdy šum je typu ulice a jeden na SNR 0 dB pro šum nákupního centra, obojí pro zdravé lidi. Výsledné hodnoty p a χ^2 statistiky jsou uvedeny v příloze, v tabulce A.1. Na základě toho můžeme konstatovat, že s přijatelnou mírou nepřesnosti předpokládáme normální rozdělení zkoumaných výběrů. T-test tedy můžeme použít.

Podoba statistiky t t-testu závisí na tom, zdali mají $\mathbf{x}$ a $\mathbf{y}$ shodné rozptyly. To určíme pomocí tzv. F-testu.

F-test

Testujeme nulovou hypotézu, že $\sigma_1^2 = \sigma_2^2$, tedy, že rozptyl je shodný pro oba soubory. Testovací statistika F má tvar

$$F = \frac{S_{\mathrm{x}}^2}{S_{\mathrm{y}}^2}, \tag{3.3}$$

kde S_{x}^2 a S_{y}^2 jsou výběrové rozptyly $\mathbf{x}$, $\mathbf{y}$ spočítané jako

$$S_{\mathrm{x}}^2 = \frac{1}{n}\sum_{n=0}^{n-1}(x_i - \overline{x})^2, \tag{3.4}$$

kde n je velikost výběru a $\overline{x}$ je výběrová střední hodnota (spočtená analogicky jako ve vztahu 2.2). Stejně bychom dostali i S_{y}^2. Pokud by se hodnota S_{y}^2 ukázala větší, než S_{x}^2, ve vzorci 3.3 bychom zaměnili čitatel a jmenovatel.

Výslednou hodnotu F pak porovnáváme s příslušným kvantilem F rozdělení s $((n-1),(n-1))$ stupni volnosti pro danou hladinu významnosti α (zde zvolena $\alpha = 0.05$). Nulovou hypotézu zamítáme, pokud hodnota F přesáhne mez, danou příslušným kvantilem rozdělení, odpovídající stupňům volnosti a zvolené α.

Ukázalo se, pomocí F-testu, že nulovou hypotézu o rovnosti rozptylů zamítáme na hladině významnosti $\alpha = 0.05$ pouze ve čtyřech případech (což odpovídá zhruba 20%). V těchto případech bude mít t statistika jiný počet stupňů volnosti. Výsledné hodnoty p a F statistiky jsou uvedené v příloze, v tabulce A.2.

Statistika t má tvar (pro stejnou velikost souborů n)

$$t = \frac{|\overline{x} - \overline{y}|}{\sqrt{\frac{S_{\mathrm{x}}^2 + S_{\mathrm{y}}^2}{n}}}, \tag{3.5}$$

s t rozdělením o $2n-2$ stupních volnosti pro rovnost rozptylů obou výběrů. V případě kdy toto neplatí (tedy celkem pro 4 případy) je počet stupňů volnosti roven

$$\frac{\left({S_x}^2 + {S_y}^2\right)^2 (n-1)}{{S_x}^4 + {S_y}^4}. \tag{3.6}$$

Velikost t se pak porovnává s příslušným kvantilem rozdělení, odpovídající hladině významnosti α. Zde je opět $\alpha = 0.05$. Pokud se hodnota t nevejde do daných mezí, nulovou hypotézu zamítáme na hladině významnosti α.

Při každém počítání statistiky určujeme hodnotu p, značící pravděpodobnost pozorovaných výsledků (tedy zde výběrových středních hodnot souborů) za předpokladu, že platí nulová hypotéza. Čím je tedy p menší, věříme nulové hypotéze méně. Pokud je $p < \alpha$, hypotézu zamítáme.

3.1.1 Výsledky

V tabulce 3.1 jsou uvedeny výsledné hodnoty p a t statistiky pro $\hat{\sigma}_{F_0}^{\text{sem}}$ ze SWIPE pro 30 promluv zdravých lidí a 30 pacientů s PN. U řádků s příslušnými hodnotami SNR jsou uvedeny údaje 4x, zvlášť pro každý typ šumu.

Platí, že čím je hodnota statistiky t větší, rozdíl v obou rozděleních je významnější, což následně demonstruje hodnota p. Jak je vidět z výsledků pro Gold standard na prvním řádku, snažíme se o co největší rozdíl. Klíčová je pak hranice $p = \alpha = 0.05$, kdy na takovéto hladině významnosti nezavrhujeme nulovou hypotézu o shodnosti středních hodnot. Vyšší hodnoty p pak znamenají, že prakticky nelze rozlišit mezi výsledky pro pacienty s PN a zdravé lidi.

	t	p	
Gold standard	3.92	< 0.001	
Čistá promluva	3.34	0.002	
SNR 20 dB	3.73	< 0.001	křižovatka
	3.75	< 0.001	tramvaj
	3.88	< 0.001	ulice
	3.39	0.001	nák. centrum
SNR 10 dB	3.65	< 0.001	křižovatka
	3.34	0.002	tramvaj
	3.54	< 0.001	ulice
	2.57	0.013	nák. centrum
SNR 6 dB	3.38	0.001	křižovatka
	2.44	0.018	tramvaj
	3.58	< 0.001	ulice
	1.96	0.055	nák. centrum
SNR 0 dB	2.46	0.017	křižovatka
	1.14	0.260	tramvaj
	2.61	0.011	ulice
	1.60	0.110	nák. centrum

Tabulka 3.1: Hodnoty t statistiky a příslušného procenta p t-testu $\hat{\sigma}_{F_0}^{\text{sem}}$ ze SWIPE pro 30 PN a 30 HC promluv. U řádku pro příslušné SNR jsou hodnoty 4x, pro každý typ šumu. Případy, kdy $p > \alpha$, značící, že příslušné výběry mají střední hodnoty podobné, jsou označeny modře.

Ukazuje se, že až na výjimky dává t-test dobré výsledky, kdy se hodnoty p drží na velmi nízké úrovni, buď hluboko pod 0.001 a nebo velmi blízko této hodnotě. V některých případech, které jsou v tabulce označeny modře, je však p velmi vysoké (a odpovídající hodnota t nízká). V těchto případech již nelze rozlišit mezi výsledky od nemocných a zdravých. Jedná se o promluvy s nižším SNR (jeden případ na 6 dB a 2 na 0 dB), ukazuje se tedy, že přítomnost silného šumu může výsledné zhodnocení zkreslovat.

Také je vidět, že velmi záleží na typu šumu, kdy pro jednu hodnotu SNR

se mohou p lišit. Například u SNR 10 dB jsou dvě hodnoty p řádu 10^{-4} (v tabulce uvedené jako < 0.001) což je oproti zbylým dvěma rozdíl jednoho řádu.

Výsledkem této části je zjištění, že SWIPE nekazí rozdělení směrodatných odchylek $\sigma_{F_0}^{\text{sem}}$ od nemocných a zdravých lidí. Jediný pozor si musíme dát pro nižší SNR (6 dB a převážně pak 0 dB), kdy pro některé typy šumu nebyly výsledky použitelné na případnou diagnózu.

3.2 Porovnání s podobnými studiemi

Pokusíme se práci porovnat se studiemi, zabývajícími se podobným tématem. Tedy hodnocení kvality řeči, nahrané na chytrý telefon. Jedná se o články [20, 21, 22, 23] ze kterých jsme vycházeli při zvolení časového průběhu F_0 a konkrétně jeho směrodatné odchylky jako náš hlavní cíl zkoumání.

Co mají všechny práce společné je podoba analyzovaného řečového segmentu. Jedná se o krátké úseky v řádu jednotek sekund, typicky znělé hlásky /a/ [20, 21, 22, 23], nebo jejich kombinací, /aiu/ v [21]. V [21] do studie zahrnuje i plynulou řeč s pauzami, v řádu jedné až dvou krátkých vět. V pracích, [21, 22], jsou použity syntetizované řečové segmenty, v ostatních se jedná o nahrávky pacientů. Jedná se tedy o značně odlišný přístup oproti našemu, kdy zkoumáme dlouhou řečovou promluvu. Tento přístup byl zvolen jako vhodnější a univerzálnější pro případnou aplikaci, kdy by mohly být například analyzované záznamy telefonních hovorů.

Velikosti databázi nejsou výrazně větší či menší než zde (60 promluv). Konkrétně 50, pro [20], 4 syntetizované segmenty pro 11 případů úrovně šumu na pozadí a jednoho bez zašumění v [21], 2 typy syntetizovaných segmentů pro 3 úrovně šumu a tichého prostředí u [22] a 118 krátkých řečových úseků ve [23].

Práce s šumy se velice liší. Studie [22, 23] s nimi nepracují vůbec, v [20, 21] se šum přidává pomocí reproduktorů umístěných v bezodrazové komoře v určité vzdálenosti od nahrávacího telefonu, kdy jeden vysílá čistý řečový segment a druhý šum, na hlasitosti dané odpovídajícímu hrubému SNR. Tedy stejně jako zde se jedná o „umělé“ vytvoření zašumělé nahrávky, kde my jsme pouze sčítali oba signály, převedené do digitální podoby. Druhý způsob bere více v potaz charakteristiku mikrofonů chytrých telefonů, na druhou stranu se musí dobře pohlídat a zanalyzovat zvukové vlastnosti reproduktorů a celé nahrávací soustavy.

Celkový rozdíl je tedy následující. Uvedené studie měly za hlavní cíl určit robustní ukazatele kvality řeči (mimo F_0 zkoumaly mnohé další, např. tzv. „jitter“ či „shimmer“) při nahrávání na mikrofon chytrého telefonu, u [20, 21] i za přítomnosti šumu na pozadí. Navíc [20] používá pouze typy šumu spojené

s plynulou řečí, tedy nádechy, zadrhávání apod. [21] využívá šum reálného prostředí, ale jen jeden typ. Analýza řečového segmentu byla určována *pouze algoritmem Praat*, u [20, 21, 22], a softwarem Dr. Speech u [23]. Výsledek byl, že nejvěrohodnější je časový průběh F_0, konkrétně jeho směrodatná odchylka.

V této práci jsme se již věnovali jen zkoumání F_0 jako ukazatele řečové kvality, na „obtížnější“ úloze (dlouhé plynulé promluvy) avšak hodnotnější pro případnou aplikaci. Hlavní cíl byl otestovat metody estimace F_0, čímž se výše uvedené studie vůbec nezabývaly, každá používala pouze jednu. Navíc většina využívala estimátor Praat, který se zde ukázal jako velmi nepřesný. Tyto studie tedy tvoří jakýsi předstupeň této práce, kdy určily F_0 jako věrohodný ukazatel a *zde potom uvádíme, jak tento průběh robustně určit*, za daných podmínek.

3.3 Závěr

Výsledky práce. V této práci jsme se primárně zabývali testováním estimátorů základní hlasivkové frekvence F_0 na nahrávkách pořízených chytrým telefonem a v přítomnosti uměle přidaného šumu, simulující reálné prostředí. Z vybraných estimátorů (sekce 2.1), Praat, SWIPE a BaNa, popsaných v části 1.2.2, 1.2.3 a 1.2.4, byl jako suverénně nejlepší zhodnocen SWIPE.

Oproti zbylým estimátorům měl robustnější výsledky v přítomnosti šumu z hlediska RMSE a celkově byly jeho výsledky přesnější, což jsme pozorovali na korelaci s Gold standardem (viz. sekce 2.3.1, 2.4.1 a 2.4.5). Ostatní algoritmy nebyly tak přesné, navíc v zašuměných nahrávkách selhávaly (důvody byly zkoumány v sekci 2.4.2), což se v případě BaNa ani nedalo zlepšit jednoduchou filtrací, v části 2.4.4. Jako možného kandidáta do mobilního výzkumu monitoringu F_0 a její směrodatné odchylky jsme tedy navrhli SWIPE.

Ten má velmi dobré výsledky pro většinu typů šumů (z výsledků v kapitole 2.4.5). V případě nákupního centra je však přesnost horší, stejně jako pro nulovou hladinu SNR u ostatních typů, např. u tramvaje. V sekci 3.1 jsme pomocí t-testu zkoumali, zda-li nám tyto nepřesnosti ve výsledcích neznehodnotí výsledné rozdělení nashromážděných dat, kdy by bylo nemožné rozlišit trend mezi zdravým a člověkem s Parkinsonovou chorobou. Ukázalo se, na základě výsledků v sekci 3.1.1, že tento případ se vyskytl pouze na nízkých hladinách SNR a jen pro určitý typ šumu.

V části 2.5 jsme pak zjistili, že mobilní telefony dokáží šum prostředí do jisté míry tlumit. Konkrétně se hodnoty odhadu $\hat{\text{SNR}}$ nikdy nedostaly na nulovou hladinu a naprostá většina ležela v oblasti, kdy SWIPE funguje prakticky bez problémů.

Lze tedy konstatovat, že všechny pokyny v zadání byly splněny.

Na základě těchto výsledků můžeme tvrdit, že SWIPE je opravdu vhodný

algoritmus pro zamýšlenou aplikaci. Estimátor je schopný dávat poměrně přesné výsledky pro většinu analyzovaných případů a chovat se dostatečně robustně i ve velmi hlučných prostředích. Pomocí odhadnutých $\hat{\sigma}_{F_0}^{\text{sem}}$ lze pak úspěšně rozlišit mezi hodnotami od zdravého člověka a od nemocného s Parkinsonovou nemocí až do SNR 6 dB.

Přínos této práce je tak pro klinickou praxi poměrně rozsáhlý. Kromě případné brzké diagnostiky onemocnění lze sledovat efekty farmakoterapeutické léčby a její účinnost nebo detailní monitorování progrese nemoci. Také je možno určovat vývoj a stav řečové kvality, která nemusí být nutně spojená s motorickou poruchou, což poskytuje cenné informace pro případnou řečovou terapii. Veškerá diagnostika přitom může probíhat pomocí chytrých telefonů pacientů, prakticky bez omezení na typ prostředí.

Navazující výzkum. Další studie v této oblasti by se mohly zabývat rozsáhlejším a detailnějším výzkumem na podobné téma, například otestováním více estimátorů (oproti třem v této práci) a jejich aplikace na *reálně* pořízené nahrávky, kde by šum byl přítomen přirozeně a ne přidán uměle, jako zde. Tím by se lépe prozkoumaly aspekty reálné aplikace.

Dále bude výzkum zaměřen na kompletaci databáze promluv, snímaných na chytrý telefon v reálných podmínkách a na následnou statistickou analýzu trendů vývoje F_0.

Literatura

[1] Athanasios Tsanas, Matías Zanartu, Max A. Little, Cynthia Fox, Lorraine O. Ramig and Gari D. Clifford. *Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering.* Journal of the Acoustical Society of America 2014; 135: 2885-2901.

[2] Yang N, Ba H, Cai W, Demirkol I, Heinzelman W. BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music. IEEE/ACM Transactions on Audio, Speech and Language Processing 2014; 22: 1833-1848.

[3] Rusz J, Hlavnicka J, Tykalova T, Buskova J, Ulmanova O, Ruzicka E, Sonka K. Quantitative assessment of motor speech abnormalities in idiopathic REM sleep behaviour disorder. Sleep Medicine 2016;19:141-147.

[4] Rusz J, Cmejla R, Ruzickova H, Ruzicka E. *Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease.* J Acoust Soc Am. 2011;129:350-367.

[5] Ray D. Kent, Ph.D., Gary Weismer, Ph.D., Jane F. Kent, Houri K. Vorperian, Joseph R. Duffy. *Acoustic studies of dysarthric speech: methods, progress, and potential.* J. COMMUN. DISORD. 32 (1999), 141–186

[6] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang. *Handbook of Speech Processing.* Springer, 2008. ISBN: 978-3-540-49125-5

[7] David Talkin, *A Robust Algorithm for Pitch Tracking (RAPT).* Speech Coding and Synthesis, W. B. Kleijn and K. K. Palatal (eds), Elsevier Science B.V., 1995, 497-518

[8] Paul Boersma, *Acurrate short-term analysis of the fundamental frequency and the harminics-to-noise ratio of a sampled sound*, Institute of Phonetic Sciences, University of Amsterdam, Proceedings 17 (1993), 97-110.

[9] Arturo Camacho. *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*. J Acoust Soc Am. 2008; 124:1638-1652.

[10] A. V. Oppenheimer, R. V. Schafer, *Discrete Time Signal Processing*, Prentice Hall, New Jersey, 1989. ISBN: 978-0-132-16292-0

[11] Cheveigné, A., and Kawahara, H. *YIN, a fundamental frequency estimator for speech and music*. J. Acoust. Soc. Am. 2002; 111:1917-1930.

[12] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, E*Motion recognition by speech signals*. Proc. 8th Eur. Conf. Speech Commun. Technol., 2003.

[13] Jan Roth, Marcela Sekyrová a Evžen Růžička. *Parkinsonova nemoc*. Maxdorf, 1999. ISBN: 978-8-073-45178-3

[14] Poewe W, Seppi K, Tanner CM, Halliday GM, Brundin P, Volkmann J, Schrag AE, Lang AE. *Parkinson disease*. Nat Rev Dis Primers 2017;23:17013.

[15] Postuma RB, Lang AE, Gagnon JF, Pelletier A, Montplaisir JY (2012). *How does parkinsonism start? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder*. Brain 135:1860–1870

[16] Ho AK, Iansek R, Marigliani C, Bradshaw JL, Gates S (1998). *Speech impairment in large sample of patients with Parkinson's disease*. Behav Neurol 11:131–137

[17] Logemann JA, Fisher HB, Boshes B, Blonsky ER (1978). *Frequency and coocurence of vocal tract dysfunction in the speech of a large sample of Parkinson patients*. J Speech Hear Disord 43:47–57

[18] R. B. Postuma, D. Berg, M. Stern, W. Poewe, C. W. Olanow, W. Oertel, J. Obeso, K. Marek, et al. *MDS clinical diagnostic criteria for Parkinson's disease*, Mov. Disord. vol. 30, no. 12, pp. 1591–1601, 2015.

[19] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, et al. *Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results*. Mov. Disord., vol. 23, pp. 2129–2170, 2008.

[20] Maryn Y., Ysenbaert F., Zarowski A., Vanspauwen R. *Mobile Communication Devices, Ambient Noise, and Acoustic Voice Measures*. Journal of Voice 2016;31(2):248-248.

[21] Lebacq, Jean and Schoentgen, Jean and Cantarella, Giovanna and Thomas Bruss, Franz and Manfredi, Claudia and DeJonckere, Philippe.

Maximal Ambient Noise Levels and Type of Voice Material Required for Valid Use of Smartphones in Clinical Voice Research. Journal of Voice 2017;31(5):550-556.

[22] Manfredi, Claudia and Lebacq, J and Cantarella, Giovanna and Schoentgen, Jean and Orlandi, Silvia and Bandini, Andrea and DeJonckere, P.H. *Smartphones Offer New Opportunities in Clinical Voice Research.* Journal of Voice 2017;31(1):111-111.

[23] Uloza, Virgilijus and Padervinskis, Evaldas and Vegiene, Aurelija and Pribuisiene, Ruta and Saferis, Viktoras and Vaiciukynas, Evaldas and Gelzinis, Adas and Verikas, Antanas. *Exploring the feasibility of smart phone microphone for measurement of acoustic voice parameters and voice pathology screening.*Archives of Oto-Rhino-Laryngology, 2015 ;272(11):3391-3399.

[24] E. Zwicker, G. Flottorp, S.S. Stevens: *Critical band- width in loudness summation*, J. Acoust. Soc. Am. 1957; 29:548-557

[25] C. Spearman, *The Proof and Measurement of Association between Two Things.* The American Journal of Psychology 1904, 15(1):72-101.

[26] Paul van Alphen, Dick R.van Bergem: *Markov Models and their Applications in Speech Recogniction.* Proc. Inst. Phon. Sci. Univ. of Amsterdam, 1989, 1–26.

[27] Hartmut Traunmüller and Anders Eriksson, *The frequency range of the voice fundamental in the speech of male and female adults.* Institutionen för lingvistik, Stockholms universitet, 1995, Sweden.

[28] Paul Boersma and David Weenink. *The Praat software*, `http://www.fon.hum.uva.nl/praat/`

[29] SmartMob. *Smart Recorder - High-quality voice recorder (Beta) application for Android.* `https://play.google.com/store/apps/details?id=com.andrwq.recorder&hl=en`

Příloha A

Tabulky výsledných hodnot pro χ^2 a F testy

HC	p	χ^2
Golden standard	0.093	4.73
Čistá promluva	0.941	0.12
SNR 20 dB, šum 1	0.057	5.71
SNR 20 dB, šum 2	0.221	3.01
SNR 20 dB, šum 3	0.160	3.66
SNR 20 dB, šum 4	0.614	0.97
SNR 10 dB, šum 1	0.168	1.89
SNR 10 dB, šum 2	0.160	1.97
SNR 10 dB, šum 3	0.850	0.04
SNR 10 dB, šum 4	0.345	0.89
SNR 6 dB, šum 1	0.136	2.18
SNR 6 dB, šum 2	NaN	0.65
SNR 6 dB, šum 3	0.016	8.23
SNR 6 dB, šum 4	0.517	0.41
SNR 0 dB, šum 1	0.100	2.70
SNR 0 dB, šum 2	0.361	0.83
SNR 0 dB, šum 3	0.326	2.23
SNR 0 dB, šum 4	0.043	4.06

PN	p	χ^2
Golden standard	0.064	5.49
Čistá promluva	0.304	1.05
SNR 20 dB, šum 1	0.327	1.54
SNR 20 dB, šum 2	0.543	0.36
SNR 20 dB, šum 3	0.452	0.56
SNR 20 dB, šum 4	0.505	0.44
SNR 10 dB, šum 1	0.568	1.12
SNR 10 dB, šum 2	0.870	0.27
SNR 10 dB, šum 3	0.120	4.23
SNR 10 dB, šum 4	0.734	0.61
SNR 6 dB, šum 1	0.222	3.00
SNR 6 dB, šum 2	0.337	2.17
SNR 6 dB, šum 3	0.151	2.06
SNR 6 dB, šum 4	NaN	0.70
SNR 0 dB, šum 1	0.279	1.32
SNR 0 dB, šum 2	0.407	1.79
SNR 0 dB, šum 3	0.280	1.16
SNR 0 dB, šum 4	0.450	0.56

Tabulka A.1: Tabulky výsledných hodnot p a statistiky χ^2 z χ^2 testu pro výběry zdravých lidí (HC, tabulka vpravo) a a pacientů s PN (PN, levá tabulka). Hodnota p určuje pravděpodobnost pozorovaného výsledku za předpokladu, že platí nulová hypotéza. Tedy, že výběry pochází z normálního rozdělení.
Hodnota NaN znamená, že jsme neměli dostatečný počet stupňů volnosti na určení pravděpodobnosti. Podle velikosti příslušné statistky však usuzujeme, že nulová hypotéza platí.
Modře jsou vyznačeny případy, kdy musíme nulovou hypotézu o normálním rozdělení zamítnout, na hladině významnosti $\alpha = 0.05$.
Šum 1 značí rušnou křižovatku, šum 2 jedoucí tramvaj, šum 3 ulici a šum 4 nákupní centrum.

	p	F
Gold standard	0.157	1.70
Čisté promluvy	0.425	1.34
SNR 20 dB. šum 1	0.130	1.76
SNR 20 dB. šum 2	0.049	2.10
SNR 20 dB. šum 3	0.230	1.56
SNR 20 dB. šum 4	0.278	1.50
SNR 10 dB. šum 1	0.077	1.94
SNR 10 dB. šum 2	0.011	2.63
SNR 10 dB. šum 3	0.307	1.46
SNR 10 dB. šum 4	0.885	0.94
SNR 6 dB. šum 1	0.116	1.80
SNR 6 dB. šum 2	0.011	2.60
SNR 6 dB. šum 3	0.574	1.23
SNR 6 dB. šum 4	0.382	0.72
SNR 0 dB. šum 1	0.099	1.86
SNR 0 dB. šum 2	0.017	2.47
SNR 0 dB. šum 3	0.287	0.67
SNR 0 dB. šum 4	0.433	0.74

Tabulka A.2: Výsledné hodnoty p statistiky F pro provedený F-test, popsán v kapitole 3.1. Hodnota p značí pravděpodobnost pozorovaného výsledku za předpokladu, že platí nulová hypotéza (tedy zde rovnost rozptylů obou výběrů). Modře jsou vyznačeny případy, kdy jsme nulovou hypotézu byli nuceni zamítnout, na hladině významnosti $\alpha = 0.05$.
Šum 1 značí rušnou křižovatku, šum 2 jedoucí tramvaj, šum 3 ulici a šum 4 nákupní centrum.

Příloha B

Fotky prostředí, kde byly pořízeny šumové nahrávky

Obrázek B.1: Rušná křižovatka

Obrázek B.2: Nákupní centrum

Obrázek B.3: Rušná ulice

Obrázek B.4: Uvnitř jedoucí tramvaje