

Odhad základního tónu řeči s lokalizací hlasivkových pulsů a pitch-synchronní segmentace

P. Mizera, P. Pollák

České vysoké učení technické v Praze, Fakulta elektrotechnická, Katedra teorie obvodů
mizerpet@fel.cvut.cz, pollak@fel.cvut.cz

Abstrakt

Článek prezentuje výsledky analýzy přesnosti standardních algoritmů odhadu základní frekvence a vliv jednotlivých algoritmů PDA na přesnost lokalizace hlasivkových pulsů v realizovaném algoritmu PMA. Analýza je zaměřena na algoritmy na bázi autokorelační funkce (ACF), rozdílové funkce (AMDF) a bázi kepstra. V článku je prezentován postup pitch-synchronní segmentace a následná resyntéza z rozložených segmentů umožňující změny prozodických charakteristik.

1 Úvod

Odhad základního tónu, který obsahuje znělé úseky lidského hlasu, patří mezi základní parametry řečového signálu. Algoritmy realizující odhad základního tónu nalézají uplatnění v mnoha řečových aplikacích např. kodérech, rozpoznávacích řeči, které jsou založeny na artikulačních příznacích při kódování řeči a ve významné míře při syntéze řeči, či modifikaci prozodie. V článku je popsáno využití odhadu základního tónu při lokalizaci poloh hlasivkových pulsů, jež vznikají v okamžiku rázu uzavření hlasivek. Tyto pulsy jsou často označovány v anglické literatuře jako "Pitch Marks" a jsou základem pro pitch-synchronní segmentaci řečového signálu. Článek dále prezentuje resyntézu segmentovaného signálu, založenou na metodě PSOLA (Pitch Synchronous Overlap and Add) s možnostmi modifikací prozodických charakteristik.

2 Metody odhadu základního tónu lidského hlasu

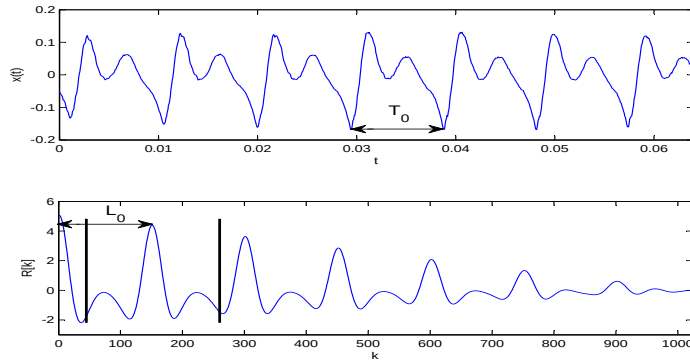
Základní hlasivkový tón nám podává informaci o výšce lidského hlasu. Vzniká periodickým kmitáním hlasivkových žláz. Obr. 1 zobrazuje znělý úsek řečového signálu, dále je zobrazena základní perioda, často označována T_0 , jejíž převrácenou hodnotou je základní frekvence hlasivkového tónu. Okamžiky s výraznou amplitudou vznikají v okamžiku rázu uzavření hlasivek a jsou nazývány hlasivkovými pulsy (ang. Pitch marks), jejichž lokalizace bude popsána dále.

2.1 Odhad na bázi ACF

Znělé úseky řeči obsahují informaci o velikosti základního tónu a mají periodický charakter. Základní přístupy pro detekci periodicity jsou založeny na autokorelační funkci. Vlastnosti autokorelační funkce umožňují detekovat periodu jako vzdálenost mezi opakujícími se maximy v jejím průběhu. Pro zachycení základní periody je doporučeno volit velikost segmentu v rozsahu 20-40ms [7].

Autokorelační funkce je sudou funkcí, proto se pro odhad základní periody využívá polovina průběhu funkce R_x , ve kterém je vyhledáno první postranní maximum. Vzdálenost nalezeného maxima od počátku udává velikost základní periody v počtu vzorků, označované často jako LAG nebo L_0 [4]. Pro potlačení chybných odhadů je výhodné první postranní maximum vyhledávat v intervalu, který odpovídá rozsahu lidského hlasu v mezích 50-500Hz [7]. Obr. 1 zobrazuje intervaly, ve kterých je prováděna detekce prvního postranního maxima.

Vlivem formantové struktury se mohou v průběhu autokorelace nacházet v okolí prvního výrazného maxima další vrcholy, které mohou způsobit chybu správné detekce. Pro potlačení vlivu formantové struktury bylo navrženo několik metod.



Obrázek 1: Časový průběh a autokorelační funkce znělého segmentu

Jednou z nich je metoda modifikované autokorelační funkce MACF. Před aplikací autokorelační funkce je nejprve signál omezen dolní propustí a spektrálně zploštěn. Dolní propust je realizována FIR filtrem s mezní frekvencí 900Hz [8]. Spektrální zploštění je provedeno centrálním a amplitudovým omezením, definované rovnicí (1) a rovnicí (2).

$$c_1[n] = \begin{cases} x[n] - h_L & \text{pro } x[n] > h_L \\ 0 & \text{pro } |x[n]| \leq h_L \\ x[n] + h_L & \text{pro } x[n] < -h_L \end{cases} \quad (1)$$

$$c_2[n] = \begin{cases} 1 & \text{pro } c_1[n] > h_L \\ 0 & \text{pro } |c_1[n]| \leq h_L \\ -1 & \text{pro } c_1[n] < -h_L \end{cases} \quad (2)$$

Kde $c_1[n]$ a $c_2[n]$ je centrálně a amplitudově omezený signál, h_L je omezovací úroveň. Omezovací úroveň je určena rovnicí (3).

$$h_L = p \cdot \min\{\max(|s_1[n]|); \max(|s_2[n]|); \max(|s_3[n]|)\} \quad (3)$$

Kde $p=0,6$ až $0,8$, $s_1[n]$, $s_2[n]$, $s_3[n]$ jsou tři mikro-segmenty segmentu $x[n]$.

Další metoda pro potlačení vlivu formantové struktury využívá chybu lineární predikce. Odhad základní frekvence je prováděn aplikací metody ACF na chybový signál z lineární prediktivní analýzy (LPC), jež neobsahuje informaci o spektrálních modifikacích, které vznikají v artikulačních ústrojích. Predikovaný signál je získán filtrací analyzovaného signálu.

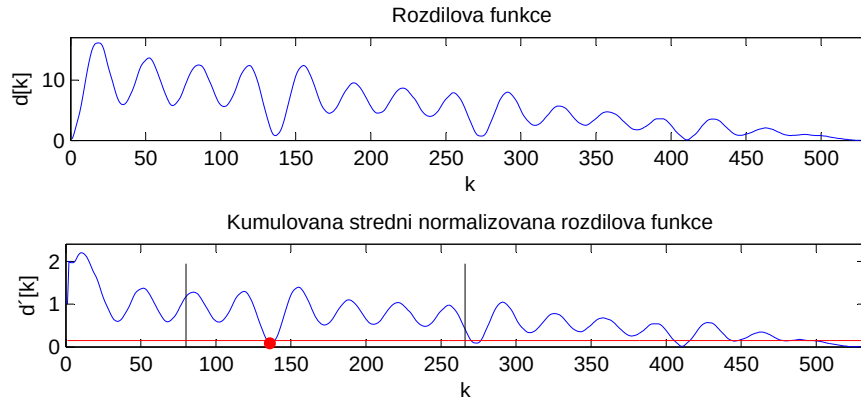
2.2 Odhad na bázi rozdílové funkce

Další metodou navrženou pro snížení chyby odhadu základní periody je metoda YIN. Základem této metody je rozdílová funkce AMDF definována rovnicí (4).

$$d[k] = \sum_{n=0}^{N-1} |x[n] - x[n - k]| \quad (4)$$

Kde k je velikost zpoždění, N je velikost segmentu. Metoda AMDF (Average Magnitude Difference Function) provádí rozdíl vzájemně posunutých vzorků. Na obr. 2 je zobrazen průběh rozdílové funkce. Velikost odhadnuté f_0 je rovna vzdálenosti prvního postranního minima.

Přesnost této metody je opět ovlivněna formantovou strukturou, která způsobuje vznik falešných údolí v okolí prvního postranního minima. Pro potlačení chybných odhadů byla funkce $d[k]$ modifikovaná do funkce označované jako kumulovaná střední normalizovaná rozdílová funkce $d'[k]$ (ang. Cumulative mean normalized difference function) definována rovnicí (5) a její průběh je zobrazen na obr. 2.



Obrázek 2: Odhad základní frekvence pomocí metody YIN

$$d'[k] = \begin{cases} 1, & \text{pro } k = 0, \\ \frac{d[k]}{\frac{1}{k} \sum_{j=1}^k d[j]}, & \text{pro } k = 1, 2, \dots, N-1, \end{cases} \quad (5)$$

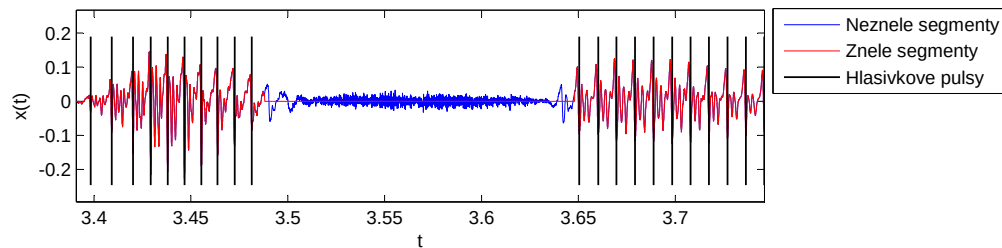
V obr. 2 je červenou čarou definována velikost prahu, pod který musí průběh funkce poklesnout. Následně je v intervalu, který poklesl pod stanovený práh, vyhledáno minimum. Nepoklesne-li průběh pod stanovený práh, je vyhledáno globální minimum funkce. Účelem stanoveného prahu je snížit chybu odhadu základní periody, ale v případě, kdy je vyhledáváno globální minimum, může docházet k hrubým chybám. Proto pro snížení této chyby je doporučeno provést kvadratickou interpolaci a vyhledávat minimum v určitém intervalu, ve kterém se předpokládá f_0 . Detailní popis této metody lze nalézt v práci [2].

2.3 Odhad na bázi kepra

Posledním uvedeným algoritmem odhadu f_0 je metoda využívající reálného kepra počítaného na bázi DFT. Tento odhad je založen na vlastnostech reálného kepra, které umožňuje oddělit část buzení a část, která obsahuje informaci o spektrální modifikaci řečového signálu. Několik prvních koeficientů reálného kepra obsahuje informaci o spektrální modifikaci, proto detekce základní frekvence je prováděna v intervalu 32 - 320 [-], jež odpovídá intervalu 50-500Hz při vzorkovací frekvenci 16kHz. V tomto intervalu je vyhledáno maximum, jehož vzdálenost od počátku je přepočtena na velikost odhadnuté základní frekvence.

3 Realizace algoritmu PMA a pitch-synchronní segmentace

V této části bude popsán postup realizovaného algoritmu PMA, jež byl vytvořen na základě provedené analýzy několika PMA algoritmů [3], [5], [6]. Cílem algoritmů PMA je lokalizace poloh hlasivkových pulsů ve znělých úsecích řeči, viz. obr. 3. Přesnost lokalizace algoritmu je založena na přesnosti odhadu základní frekvence. Lokalizované polohy hlasivkových pulsů jsou dále využity pro pitch-synchronní segmentaci řečového signálu.



Obrázek 3: Lokalizované hlasivkové pulsy

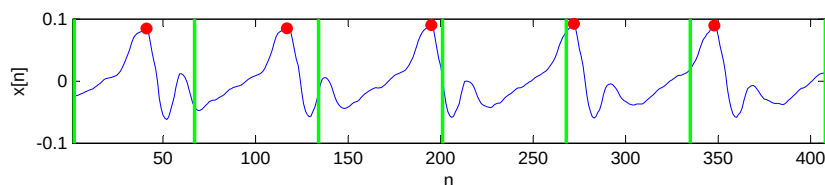
Vstupem do algoritmu PMA je analyzovaný signál a již odhadnutý průběh základní frekvence pomocí algoritmu PDA. Lokalizace hlasivkových pulsů je prováděna pouze ve znělých segmentech. Algoritmus každý znělý segment rozdělí do N mikro-segmentů na základě odhadnuté základní frekvence dle rovnice (6) a rovnice (7)

$$MS_l(i) = \text{floor}\left(\frac{f_s}{f_{0seg(i)}} \cdot p_A\right), \quad (6)$$

kde $MS_l(i)$ je velikost mikro-segmentu v i -tém segmentu, f_s je vzorkovací frekvence, $f_{0seg(i)}$ je velikost odhadnuté základní frekvence v i -tém segmentu, p je parametr ovlivňující velikost mikro-segmentu ($p_A = 0, 8$).

$$N = \text{floor}\left(\frac{S_l}{MS_l(i)}\right) \quad (7)$$

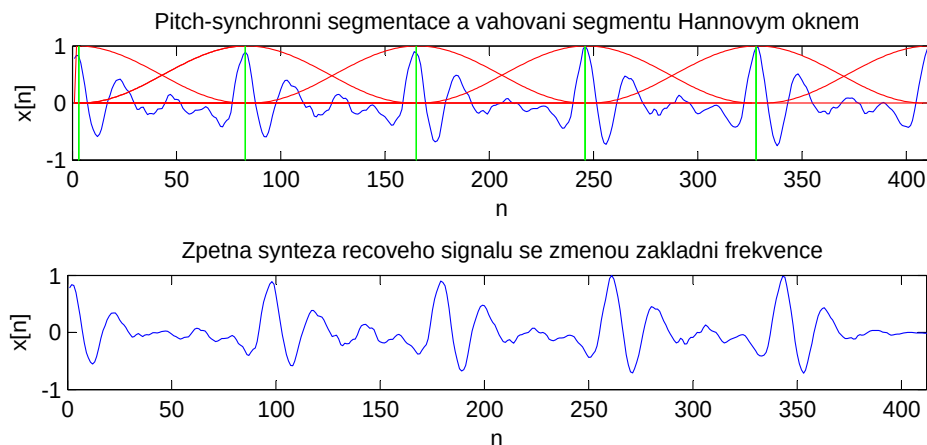
Kde N je počet mikro-segmentů v segmentu i , S_l je velikost segmentu. V dalším kroku algoritmus v každém mikro-segmentu vyhledává maximální hodnotu, která odpovídá pozici hlasivkového pulsu. Obr. 4 zobrazuje postup lokalizace hlasivkových pulsů ve znělých segmentech řečového



Obrázek 4: Odhad umístění hlasivkových pulsů

signálu. Hlasivkové pulsy jsou zobrazeny červenou barvou. V posledním kroku algoritmus provádí optimalizaci umístění hlasivkových pulsů, jehož cílem je odstranění falešných hlasivkových pulsů, které vznikají při chybném odhadu základní frekvence.

Pitch-synchronní segmentace a resyntéza z rozložených segmentů je zahrnuta v algoritmu TD-PSOLA, který je často používán při konkatenační syntéze řeči a své uplatnění nachází i v aplikacích, jejichž cílem je modifikovat prozodické charakteristiky lidského hlasu. V prvním



Obrázek 5: Pitch-synchronní segmentace a zpětná syntéza

kroku algoritmus provádí pitch-synchronní segmentaci, která je založena na lokalizovaných hlasivkových pulsech. Obr. 5 zobrazuje rozklad signálu synchronně s hlasivkovými pulsy, jejichž poloha je zobrazena zelenou barvou. Při rozkladu signálu je použito Hannovo okno. Ve druhém kroku dochází k časovým či frekvenčním modifikacím, které se projevují jako změny prozodických charakteristik (změna rychlosti řeči, či základní frekvence). Více informací lze nalézt v [7]. Posledním krokem je resyntéza z rozložených a modifikovaných segmentů, jež je provedena pomocí algoritmu OLA (Overlap-add). Obr. 5 zobrazuje syntetizovaný signál, ve kterém byla modifikována velikost f_0 .

4 Experimenty

Přesnost uvedených algoritmů PDA a PMA byla analyzována na databázi s referenčními hodnotami základní frekvence a poloh hlasivkových pulsů. Tato databáze byla vytvořena z malé části španělské databáze řady SPEECON v rámci projektu ECESS (European Center of Excellence in Speech Synthesis). Databáze se skládá ze souboru 1020 řečových signálů, které vznikly na základě 60 mluvčích (50% muži a 50 % ženy) ve věku 19-79 let.

Pro hodnocení algoritmů PDA bylo využito standardně používaných kritérií [4].

- GEH (Gross error high), procentuální vyjádření počtu znělých segmentů, v nichž hodnoty odhadnuté f_0 převýšily o 20% referenční hodnoty f_0 .
- GEL (Gross error low), vyjadřuje opačnou informaci, kolik procent znělých segmentů má odhadnutou základní frekvenci nižší než 20% referenční hodnoty.
- VE (Voiced error)/ UVE (Unvoiced error), procentuální vyjádření představující, jaká část znělých segmentů byla chybně určena jako neznělá a naopak.

Přesnost lokalizace hlasivkových pulsů algoritmem PMA byla hodnocena za základě dynamického programování, které porovnává polohy odhadnutých hlasivkových pulsů a referenčních pozic hlasivkových pulsů [5]. Algoritmus vyhodnocující přesnost využívá třech matematických operací, pomocí nichž se snaží soubor odhadnutých hlasivkových pulsů přeuspořádat tak, aby se rovnal souboru s referenčními hlasivkovými pulsy.

Substituce značená S má za cíl odhadnutý hlasivkový puls, který se nachází ve vzdálenosti větší jak 10% od referenční pozice hlasivkového pulsu, umístit na tuto referenční pozici. Odebrání (delete) D, nastane-li situace, že odhadnutý hlasivkový puls se nachází na správné referenční pozici a v okolí této pozice je další odhadnutý hlasivkový puls, je tento odhadnutý hlasivkový puls falešný a ze souboru odhadnutých hlasivkových pulsů je odebrán. Doplnění (include) I, do souboru odhadnutých hlasivkových pulsů bude doplněn další hlasivkový puls v případě, že v okolí referenční pozice hlasivkového pulsu se nenacházel žádný odhadnutý hlasivkový puls.

Počet provedených operací, kterými bylo docíleno rovnosti obou souborů, je využito pro výpočet přesnosti algoritmu PMA dle rovnice (8).

$$P = \frac{N_R - N_S - N_D - N_I}{N_R} \times 100\% \quad (8)$$

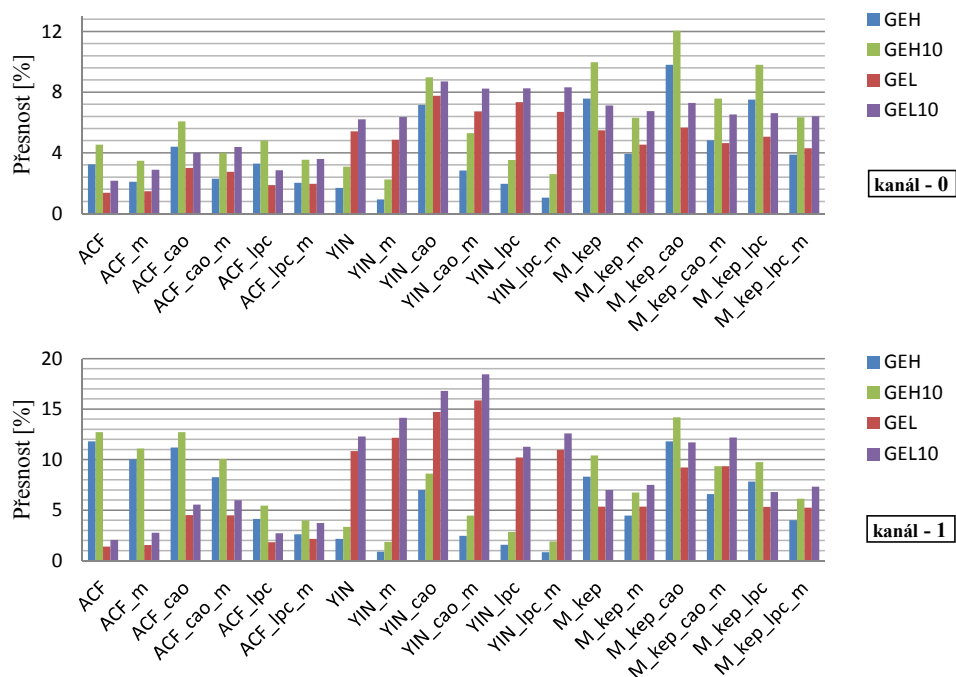
Kde N_R je počet hlasivkových pulsů v souboru referenčních hlasivkových pulsů, N_S je počet operací substituce, N_D je počet operací odebrání, N_I je počet operací doplnění.

Odhadnuté hlasivkové pulsy, které jsou ve vzdálenosti do 10% od referenční pozice, jsou považovány za správně určené. Vzdálenost 10% odpovídá přepočítané hodnotě 10% aktuálně odhadnuté základní frekvence v daném segmentu. Hodnota 10% je uvedena v [4] a vychází z výsledku v práci [8], kde uvádějí, že kvalita syntetizované řeči vytvořená pomocí metody PSOLA není snížena, jsou-li pozice hlasivkových pulsů určeny s touto nepřesností.

Při provádění experimentů bylo nastavení parametrů pro všechny algoritmy stejné. Velikost segmentu byla rovna 30 ms, posun segmentu byl roven 30 ms. Pro segmentaci signálu bylo použito obdélníkové okno. Parametry předpokládané odhadované základní frekvence byly nastaveny na hodnoty Max $f_0 = 305$ [Hz] a Min $f_0 = 80$ [Hz]. Při vyhlazování odhadu f_0 byl použit mediánový filtr 3. řádu.

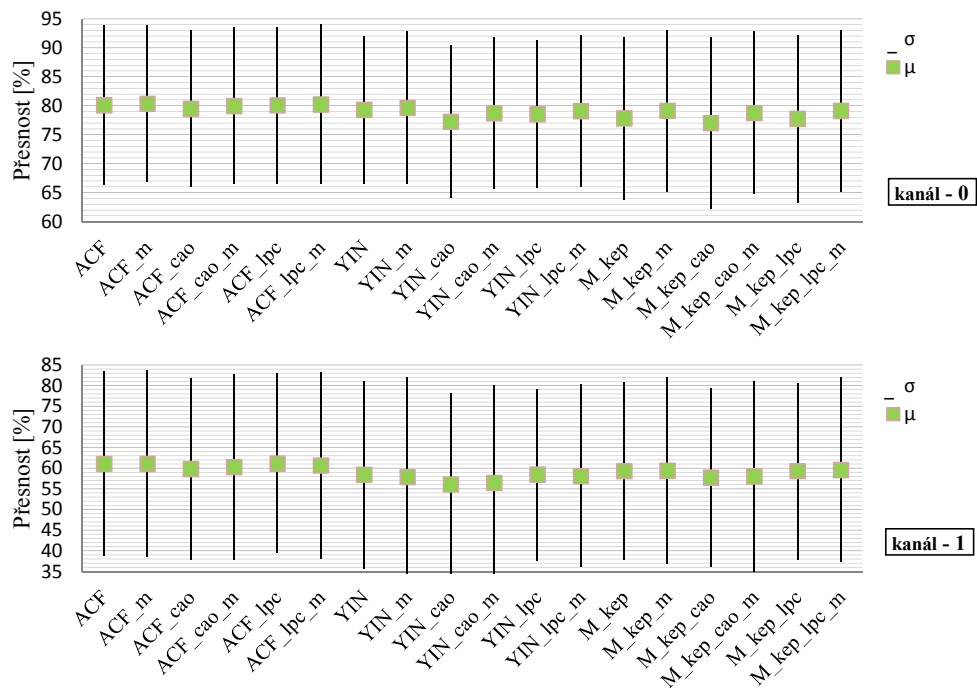
4.1 Analýza algoritmů PDA a jejich vliv na přesnost lokalice algoritmu PMA

V prvním experimentu byly algoritmy testovány na signálech ze španělské databáze, které byly zaznamenány mikrofonom v těsné blízkosti - kanál 0. Graf na obr.č. 6 zobrazuje hodnoty kritérií, která byla vypočítána pro každý testovaný algoritmus. Kriteria VE/UVE nejsou zobraze, protože



Obrázek 6: Porovnání výsledků algoritmů PDA

pro detekci znělých a neznělých segmentů byl použit detektor na bázi energie a ZCR (Zero Crossing Rate). Proto se vliv jednotlivých algoritmů neprojeví na těchto kritériích. Pro kanál 0 bylo dosaženo hodnoty $VE = 4.41\%$, $UVE = 6.91\%$. Dále jsou v grafu uvedena kritéria GEH10 a GEL10, která se liší od kritérií GEH, GEL jen ve velikosti prahu. U kritérií GEH10 a GEL10 byl práh snížen na hodnotu 10% [1].



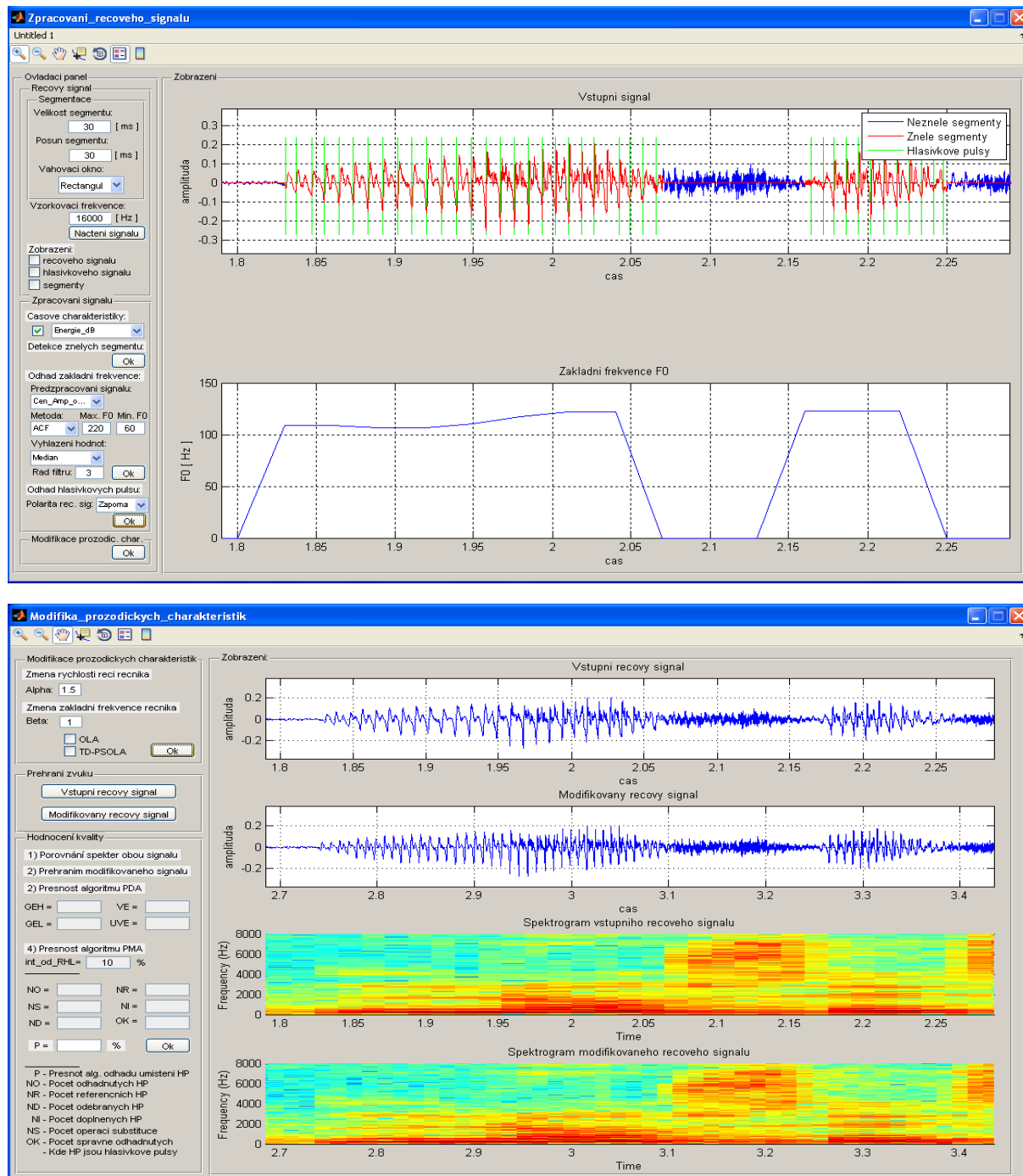
Obrázek 7: Hodnocení algoritmu PMA

Další experiment byl zaměřen na sledování chování algoritmů PDA v případě, kdy signál obsahoval vyšší poměr šumu. Nižší poměr SND se projevil zvýšením hodnot kritérií znělosti $VE = 17.3\%$, $UVE = 19.2\%$. Algoritmy PDA byly aplikovány na stejné signály, které zaznamenal vzdálenější mikrofon - kanál 1. Ze změn v grafu na obr. 6 je vidět, že snížení hodnoty SNR se projevilo u metody ACF nárůstem chybovosti kritéria GEH, oproti metodě YIN u které

narostla hodnota chybovosti GEL. U metody na bázi kepstra hodnoty kritérií GEH, GEL se zvýšily minimálně. V třetím experimentu byl sledován vliv jednotlivých algoritmů PDA na přesnost lokalizace hlasivkových pulsů algoritmem PMA. NA obr. 7 jsou zobrazeny výsledky dosažené přesnosti na kanálu 0.

Čtvrtý experiment byl proveden na signálech z kanálu 1, cílem bylo opět sledování přesnosti lokalizace PMA. Z grafů na obr. 7 je vidět, že snížení hodnoty SNR se projevilo na přesnosti všech algoritmů rovnoměrným poklesem. Přesnost algoritmu PMA poklesla přibližně o 20 %.

Všechny prezentované algoritmy a analýzy byly realizovány v prostředí MATLAB, ve kterém bylo vytvořeno i interaktivní prostředí, které umožňuje provádět změny prozodických charakteristik řečového signálu, viz. obr. 8.



Obrázek 8: GUI prostředí pro analýzu PDA a PMA algoritmů

5 Závěr

V článku byla provedena analýza standardních algoritmů odhadu základní frekvence. Pro signály z kanálu 0 dosáhl nejnižší chybovosti algoritmus ACF v kombinaci s medianovým filtrem. Algoritmus ACF je jeden z nejvíce používaných algoritmů. Dále byl sledován vliv šumu na přesnost odhadu f_0 u jednotlivých algoritmů. Z provedené analýzy nejlepších hodnot dosáhla opět metoda ACF, která byla aplikována na předzpracovaný signál pomocí LPC.

Dále byla v článku provedena analýza vlivu jednotlivých algoritmů PDA na přesnost realizovaného algoritmu lokalizace hlasivkových pulsů při různé hodnotě SNR. Nejpresnější lokalizace dosáhl PMA při odhadu f_0 pomocí ACF, i když podle hodnocení GEH a GEL tato metoda nebyla nejpresnější při odhadu f_0 . Je tedy vidět, že vyšší hodnota GEH, která způsobuje detekci falešných hlasivkových pulsů, významně neovlivňuje přesnost algoritmu PMA jako hodnota GEL, protože falešné hlasivkové pulsy jsou odstraněny při optimalizaci poloh hlasivkových pulsů.

Pro realizaci a analýzu bylo využito prostředí MATLAB, ve kterém bylo vytvořeno GUI, které mohou uživatelé využít pro analýzu řečových signálů, či pro modifikace prozodických charakteristik hlasu.

6 Poděkování

Tato práce vznikla za podpory vnitřního grantu ČVUT SGS12/143/OHK3/2T/13 Algoritmy a hardwarové realizace číslicového zpracování signálu.

Reference

- [1] BARTOŠEK, J. - HANŽL, V.: *Comparing Pitch Detection Algorithms for Voice Applications* In Digital Technologies 2010 [CD-ROM]. Žilina, 2010, ISBN 978-80-554-0304-5
- [2] DE CHEVEIGNÉ, A. - KAWAHARA, H.: *YIN, a fundamental frequency estimator for speech and music*. J. Acoust. Soc. Amer., 2000.
- [3] CHENG-YUAN, L. - JYH-SHING, R. J.: *A two-phase pitch marking method for TD-PSOLA synthesis*. In: Proc. INTERSPEECH, Jeju, Korea, pp. 1189-1192, 2004.
- [4] KOTNIK, B. - HÖGE, H. - KAČIČ, Z.: *Evaluation of pitch detection algorithms in adverse conditions*. In: Proceedings of Speech Prosody 2006, Dresden, 2006.
- [5] LEGÁT, M. - MATOUŠEK, J. - TIHELKA, D.: *On the detection of pitch marks using a robust multi-phase algorithm*. http://www.kky.zcu.cz/en/publications/LegatM_2011_Onthedetectionof.
- [6] MATTHEYSES, W. - LATACZ, L. - KONG, Y. - VERHELST, W.: *Robust pitch marking for prosodic modification of speech using TD-Psola*. V: Proc. IEEE Benelux Signal Processing Symposium, SPS-DARTS, 2006.
- [7] PSUTKA, J. - MÜLLER, L. - MATOUŠEK, J. - RADOVÁ, V.: *Mluvíme s počítačem česky*. Praha: Academia, 2006.
- [8] RABINER, L. R. - CHENG, M. J. - ROSENBERG, A. E. - A MCGONEGAL, C.: *A Comparative Performance Study of Several Pitch Detection Algorithms*. IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 5, pp. 399-418, 1976.

Petr Mizera

České vysoké učení technické v Praze, Fakulta elektrotechnická, Katedra teorie obvodů
e-mail: mizerpet@fel.cvut.cz

Petr Pollák

České vysoké učení technické v Praze, Fakulta elektrotechnická, Katedra teorie obvodů
e-mail: pollak@fel.cvut.cz