REAL-TIME MOŽNOSTI ZKRÁCENÉ AUTOKORELAČNÍ FUNKCE PRO DETEKCI ZÁKLADNÍ FREKVENCE

Jan Bartošek

Katedra teorie obvodů, ČVUT , Technická 2, 166 27 Praha

Abstract

Článek se zabývá možnostmi energeticky normalizované autokorelační funkce pro detekci základní frekvence (F0) na zkráceném okně signálu. Korelační funkce běžně operují na okně dlouhém alespoň dvě základní periody nejnižší detekovatelné frekvence a tím omezují dosažitelnou latenci. Článek přináší výsledky charakterizující změny v přesnosti algoritmu při zkrácení okna na 3/4 a 2/3 původní délky. Jako testovací data posloužila řečová databáze KEELE s referenčními hodnotami F0. Ukazuje se, že použitá metoda energeticky normalizované autokorelační funkce skutečně umožňuje detekovat základní frekvenci i na takto zkráceném okně, byť za cenu zvýšení chybovosti pro frekvence vyšší.

1 Úvod

Detekce základní frekvence (F0) pro použití v reálném čase s co nejmenším zpožděním je stále výzvou zejména v oblasti real-time hudebních technologií (převod skutečného signálu na MIDI události a následná syntéza, doplnění harmonie k hlasu a podobně). Fyzikálně jsme ale omezeni zejména nejnižší detekovatelnou frekvencí - čím nižší frekvence chceme hledat, tím déle musíme v digitálním světě čekat na dostatečný počet vzorků signálu, abychom odhad F0 mohli provést. Tyto dva principy jsou bohužel protichůdné.

S problémem snížení latence detekce F0 se odjakživa potýkají zejména výrobci hardwarových MIDI převodníků/syntezátorů obvykle úzce spjatých se speciálními nástrojovými multi-kanálovými "MIDI" snímači a po desetiletí vyvíjejí proprietární algoritmy, které často využívají specifických vlastností použitého nástroje a snímače. Například kytarový převodník firmy AXON pomocí magnetického snímače sleduje rozkmit strun a dokáže detekovat odraženou špičku signálu úderu trsátka od zmáčknutého pražce a podle doby trvání této prodlevy určit F0 mnohem dříve, než klasickou cestou. My budeme v našem experimentu operovat jen na holých řečových akustických signálech (tj. sejmutých mikrofonem).

Je nutné si uvědomit, že při nepříliš komplexních real-time DSP audio výpočtech se s dnešní výpočetní silou většinu času čeká na dostatek dat ke zpracování. Doba nutná k nasbírání těchto dat (nejčastěji naplnění bufferu zvolené velikosti zvukové karty) je určena fyzikálními principy a nelze ji urychlit. Tabulka 1 ukazuje vybrané kombinace vzorkovacích frekvencí, délek oken a příslušnou nejnižší detekovatelnou frekvenci klasickým přístupem (celé dvě periody signálu uvnitř okna). V důsledku se tedy můžeme často setkat se situací, kdy desítky milisekund čekáme na data, na kterých poté během jednotek milisekund provedeme potřebný výpočet (samozřejmě v závislosti na složitosti a implementaci DSP algoritmu). Přestože je běžné operovat v číslicovém zpracování signálů s okny délky mocnin dvou (zejména kvůli optimálnímu rozkladu pro "rychlé" algoritmy typu FFT), v tomto experimentu se na ně omezovat nebudeme. Časové rozlišení PDA je dáno jeho krokem, který se odvíjí od posunu oken ve zpracovávaném signálu.

Detekcí základní frekvence signálu se zabývá řada citací jak z historie, tak současnosti. Jen pár z nich se však dotýká aspektu co nejnižší latence a použití v reálném čase ([3], [2], [6]). Není nám ale známa studie, která by se snažila detekovat základní frekvenci z okna kratšího než celé dvě periody nejnižší detekovatelné frekvence.

Běžné korelační metody, které se ze všech PDA (Pitch Detection Algorithm) užívají nejčastěji, operují nad oknem, které má délku N rovnu dvojnásobku periody nejnižší detekovatelné základní frekvence $F0_{min}$ (a tedy nejvyšší možné periody: $N = 2 * T0_{max}$). Tím je zaručena skutečnost, že pro všechny detekovatelné frekvence $F0 >= F0_{min}$ budou v korelačním okně obsaženy alespoň dvě celé základní periody signálu, jejichž podobnost lze poté snadno porovnávat. Hlavní myšlenka, kterou se zabývá tento článek, je založena na zkrácení délky okna N, v rámci které operuje autokorelační funkce, ideálně však při zachování možnosti detekce povodní nejhlubší frekvence. Chceme tedy ověřit, zda dokážeme ze signálu s rozumnou chybovostí detekovat F0 i v případě, že nemáme k dispozici celé dvě periody signálu. Za tímto účelem je jako základ testované metody využit odhad autokorelační funkce spolu s energetickou normalizací.

2 Teoretický úvod

2.1 Přehled metod

Jsou známy dva základní typy odhadů autokorelační funkce [5]: odhad vychýlený (1) a nestranný (2). V praxi je díky své větší numerické stabilitě častěji užívaný odhad vychýlený. Ze vzorců je zřejmé, že u vychýleného odhadu pro vyšší „lagy“ k (testované periody signálu) klesá počet realizovaných součinů a výsledná hodnota funkce má tedy klesající charakter. Nestranný odhad na rozdíl od vychýleného odhadu zohledňuje skutečný realizovaný počet součinů a pro větší k se snaží trend funkce narovnat.

$$ACF_{time}(k) = \frac{1}{N} \sum_{n=0}^{N-n-1} x(n)x(n+k), k = 0, 1, ...K \tag{1}$$

$$ACF_{time}(k) = \frac{1}{N-k} \sum_{n=0}^{N-n-1} x(n)x(n+k), k = 0, 1, ...K \tag{2}$$

$$CCF(k) = \sum_{n=0}^{N-1} x(n)x(n+k), k = 0, 1, ...K \tag{3}$$

Cross-korelace (CCF) [7] formálně odstraňuje závislost délky okna na maximální detekovatelné periodě. V literatuře se setkáváme zejména s myšlenkou zkrácení jinak dlouhého okna s více něž dvěma periodami signálu pro vyšší základní frekvence, kdy se základní frekvence ze začátku okna může na konci okna lišit (dojde ke rychlé změně F0 v průběhu zkoumaného úseku řeči) a podoba jednotlivých period signálu se tak bude v původně nepřiměřeně dlouhém okně také lišit (ve skutečnosti se mnohdy liší i přímo sousední periody signálu, což detekci F0 v řečovém signálu často znesnadňuje). Daná skutečnost vede k vyhlazení maxima autokorelační funkce, které poté mnohdy není nalezeno.

Table 1: Minimální detekovatelné F0 v závislosti na délce okna a vzorkovací frekvenci

vzorkovací frekvence FS [kHz]	délka vzorku [ms]	délka okna [vzorků]	délka okna [ms]	nejnižší detekovatelné F0, dvě periody [Hz]
11,025	0,0907	512	46,440	43,07
16	0,0625	512	32,000	62,50
22,05	0,0454	1024	46,440	43,07
44,1	0,0227	2048	46,440	43,07

$$NCCF(k) = \frac{\sum_{n=0}^{N-1} x(n)x(n+k)}{\sqrt{\sum_{n=0}^{N-1} x(n)^2 \sum_{n=0}^{N-1} x(n+k)^2}}, k = 0, 1, ...K \tag{4}$$

NCCF v rovnici (4) přidává energetickou normalizaci hodnot sumy individuálně pro každé testované zpoždění. Ve skutečnosti se jedná o geometrický průměr energií porovnávaných suboken, což by mělo pomoci zejména při rozdílných amplitudách sousedních period signálu - po energetické normalizaci by měla stačit tvarová podobnost period. Konečně nestranný odhad NCCF (5) kombinuje energetickou normalizaci s narovnáním trendu.

$$nNCCF(k) = \frac{1}{N-k} \frac{\sum_{n=0}^{N-1} x(n)x(n+k)}{\sqrt{\sum_{n=0}^{N-1} x(n)^2 \sum_{n=0}^{N-1} x(n+k)^2}}, k = 0, 1, ...K \tag{5}$$

2.2 Vlastnosti autokorelačních funkcí s ohledem na nalezení nejkratší periody (tedy základní frekvence) signálu

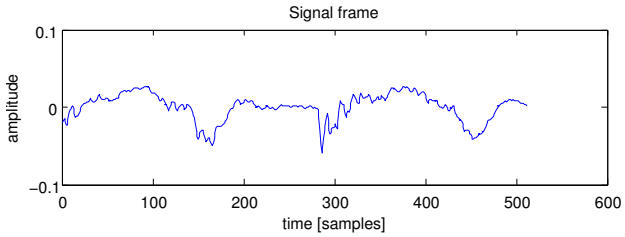
Předpokládáme, že vybrané okno signálu je periodické s periodou T_0. Pak platí, že je periodické i se všemi přirozenými násobky periody T_0:

$$x[k + nT_0] = x[k]$$

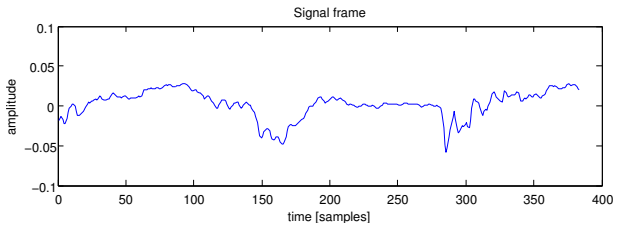
Očekávané vrcholy (peaky) autokorelační funkce budou proto nejen na hodnotě zpoždění T_0, ale i na zpoždění $2T_0$, $3T_0$, atd. Pokud bude signál stacionární v rámci celého okna (základní frekvence bude konstantní), pak lze očekávat i zcela totožné hodnoty nestranného odhadu autokorelační funkce v bodech $T_0, 2T_0, 3T_0$ atd. Hodnoty odhadu vychýleného (1) budou lineárně klesat spolu s úbytkem členů sumy pro rostoucí zpoždění k. Dodejme jen, že frekvenční analýzou signálu žádné sub-harmonické složky $F_0/2, F_0/3$, ... (odpovídající $2T_0, 3T_0$, ...) nenalezneme, jedná se o tzv. "virtuální" základní frekvence. Pokud je jako základní frekvence úseku označena frekvence $F_0/2$ (odpovídající $2T_0$), jedná se o "halving" (poloviční) oktávovou chybu odhadu. U nestranného odhadu se tedy často setkáváme s nadržováním nižších frekvencí, které odpovídající násobkům skutečné F_0.

2.3 Ukázka výstupu metod při zkrácení okna

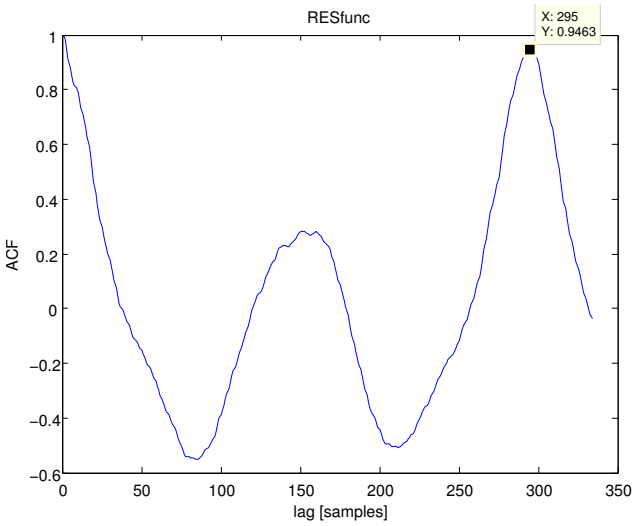
Máme znělý úsek řeči s F0=68Hz vzorkovaný frekvencí 20kHz. Plná délka takového úseku je 51,2ms (vybrali jsme okno dlouhé 1024 vzorků, tedy běžně detekovatelná nejnižší základní frekvence $F0_{min}$ při existenci dvou celých period signálu v úseku je až 39Hz). Základní perioda signálu tedy odpovídá zhruba 294 vzorkům. Nyní vezměme pouze levou polovinu úseku - prvních 25,6ms (512 vzorků, tedy běžně detekovatelná $F0_{min}$ vzroste na 78,13Hz), tento úsek je nakreslen na obrázku 1a. Všimněme si, že se do něj nevejdou celé dvě periody signálu, ty jsou totiž dlouhé zhruba 588 vzorků. Avšak funkce NCCF byla schopna detekovat F0 poměrně správně (obr.1c). Dále máme ještě více zkrácený úsek (obr.1b), který obsahuje pouze první 3/4 již zkráceného úseku, tedy 19,2ms (384 vzorků, čemuž odpovídá běžně detekovatelná $F0_{min}$=104,17Hz). I v tomto případě je funkce NCCF schopná F0 poměrně správně detekovat (obr.1d).



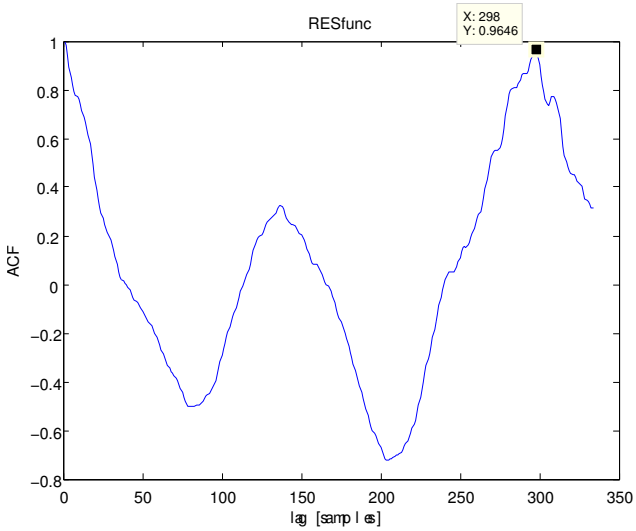
(a) Zkrácené okno signálu 512 vzorků, T0=294 vzorků



(b) Zkrácené okno signálu 384 vzorků, T0=294 vzorků



(c) NCCF funkce na okně 512 vzorků



(d) NCCF funkce na okně 384 vzorků

Figure 1: Zkrácená okna signálu a detekce jejich základní frekvence pomocí NCCF

3 Použitá databáze

V experimentu byla použita referenční F0 databáze KEELE Pitch DB [4]. Jedná se o jednokanálovou databázi, ve které jsou referenční hodnoty získány z přidruženého laryngographového signálu. Obsahem databáze je anglický text o délce zhruba 30 vteřin čtený 10 neprofesionálními mluvčími, pěti muži a pěti ženami. Formát databáze je následující: vzorkovací frekvence 20kHz, Mono, 16-bit Little Endian.

4 Použitá kritéria pro hodnocení PDA

4.1 Znělost/neznělost (V/UV), chyby znělosti

Znělost úseku je často rozhodována na základě prahování velikosti maxima spočtené korelační funkce - práh je obvykle nastaven na fixní hodnotu podle povahy testovaného materiálu (přítomnost šumu v nahrávkách) a může pro jeden daný PDA fungovat uspokojivě. Tento postup s jedním pevným prahem ale nebude fungovat pro různé varianty PDA, kdy dochází k nelineárním modifikacím korelačních křivek. Detekce znělosti však není primárním cílem práce, a proto jsme pro porovnatelnost všech metod ve finální fázi našeho experimentu vyřadili V/UV blok. Všechny úseky byly tedy algoritmy považované za znělé a došlo u nich k výpočtu odhadu F0. Ve fázi hodnocení výsledků jsou však brány v potaz pouze ty úseky, které jsou referenčně označeny jako znělé. Všechny algoritmy tedy v experimentu dosahují stejných hodnot chyb znělosti, což dále umožňuje objektivní měření.

4.2 Chyby přesnosti

Gross Error High GEH (Gross Error Low GEL) je podíl odhadů F0 (správně určených jako znělých), které se nevejdou do 20% horní (dolní) frekvenční tolerance v Hz. Chyby GEH10 a GEL10 byly zavedeny analogicky pro přísnější toleranci jen 10%. Mezi oktávové chyby patří halving chyby (HE - odhad frekvence je polovina referenční hodnoty) a doubling chyby (DE - dvojnásobek), zde používáme toleranci 1 půltónu na obě strany od referenční hodnoty F0.

4.3 Chyby v pásmech

Výše zmíněné chyby přesnosti představují jen jednu část pohledu na danou skutečnost. Pro podrobnější analýzu a ucelení tohoto pohledu na zkoumanou metodu uvádíme také výčet procentuální chybovosti metody v určitých referenčních frekvenčních pásmech. Osvědčilo se nám rozdělit hlasový rozsah na pět 2/3-oktávových pásem (57Hz-88Hz, 88Hz-141Hz, 141Hz-225Hz, 225Hz-353Hz, 353Hz-565Hz). Z vlastní zkušenosti můžeme tvrdit, že valná většina referenčních F0 v mužských promluvách se nachází ve druhém frekvenčním pásmu a v ženských promluvách ve třetím frekvenčním pásmu. Pro tento experiment jsme použili 20% toleranci pro měření chyb v pásmech, udávána hodnota je tedy součtem GEH+GEL pro dané pásmo.

4.4 Statistická kritéria

K vyhodnocení používáme také vylepšená statistická kritéria [1] - střední hodnotu rozdílu $\overline{\Delta}_{\%}$ (6) a směrodatnou odchylku rozdílů $\delta_{\%}$ (7), obě počítané v centech půltónů (100centů=1půltón).

$$\overline{\Delta}_{\%} = \frac{1200}{N} \sum_{n=1}^{N} \log_2 \frac{F_{est}(n)}{F_{ref}(n)} \tag{6}$$

$$\delta_{\%} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} [1200 \log_2 \frac{F_{est}(n)}{F_{ref}(n)} - \overline{\Delta}_{\%}]^2} \tag{7}$$

5 Popis a nastavení experimentu

5.1 Zkoumané PDA metody

Všechny testované algoritmy ve své podstatě vycházejí z autokore počítané v časové oblasti:

- M0 - ACF - vychýlený odhad na plném okně
- M1 - ACF - nestranný odhad na plném okně
- M2 - NCCF na plném okně
- M3 - Nestranný odhad NCCF na plném okně
- M4 - NCCF na 3/4 zkráceném okně
- M5 - Nestranný odhad NCCF na 3/4 zkráceném okně

- M6 - NCCF na 2/3 zkráceném okně
- M7 - Nestranný odhad NCCF na 2/3 zkráceném okně

Důsledky zkrácení okna na 3/4 a 2/3 délky pro úsek s původně nejnižší detekovatelnou frekvencí (plné okno tedy obsahuje přesně dvě periody signálu) jsou následující: Pokud zkrátíme plné okno na 3/4 jeho délky, pak pro takový úsek se budou v korelační sumě pro zpoždění k odpovídající maximální detekované periodě MAX_PER porovnávat pouze první poloviny původních period (zprava zkrácené okno nám neumožní porovnat další vzorky). Pro případ zkrácení okna na 2/3 se v takovém případě porovnávají jen levé třetiny původních period.

5.2 DSP blok

Celý experiment probíhá v offline režimu - máme tedy v každém okamžiku dostupný celý signál, který nejprve normalizujeme tak, aby rozsah amplitud byl v intervalu $< 0; 1 >$. Dále následuje běžné DSP - ze signálu jsou s 50% překryvem brány úseky, z každého úseku dostáváme odhad jedné základní frekvence. Žádné předzpracování úseku signálu není provedeno. V případě potřeby dosažení vyšší robustnosti algoritmu lze například zařadit high-pass filtr se zlomovou frekvencí 50Hz, který odstraní síťovou složku. Nejnižší hledaná frekvence $F0_{min}$ byla nastavena na 62.5Hz, nejvyšší $F0_{max}$ na 450Hz. Délka plného okna byla nastavena přesně na dvojnásobek periody $F0_{min}$, tedy na 32ms.

6 Výsledky a jejich zhodnocení

Díky úmyslné absenci rozhodovací logiky VUV jsou všemi testovanými metodami považovány všechny úseky za znělé, což vede k hodnotám Voiced Error (VE) = 0% a Unvoiced Error (UE)=1. Tyto hodnoty tedy v tabulkách s výsledky uvedeny záměrně nejsou, protože pro tento experiment nepřinášejí žádnou informaci. Výsledky pro KEELE řečovou databázi jsou v tabulce 2 a 3.

Table 2: Výsledky naměřené na KEELE databázi, první část

PDA method	GEH [%]	GEL [%]	GEH10 [%]	GEL10 [%]	DE [%]	HE [%]	$\overline{\Delta}_{\%}$ (cents)	$\delta_{\%}$ (cents)
M0	3,17	1,83	4,56	3,94	0,88	0,57	5,3	313
M1	1,98	9,37	3,24	11,56	0,50	5,30	-126	473
M2	1,57	6,78	2,80	8,97	0,35	3,49	-93	399
M3	0,75	66,18	1,42	67,57	0,14	18,40	-1246	994
M4	1,06	11,19	1,96	13,01	0,24	6,27	-165	494
M5	0,43	68,80	0,98	70,13	0,06	18,60	1374	978
M6	0,74	14,71	1,52	16,31	0,15	8,15	-224	561
M7	0,30	70,65	0,71	71,68	0,02	18,90	-1344	964

Nestranný odhad zklamal již ve variantách na plné délce okna (M1 a M3). M1 oproti základní autokorelační metodě vychýleného odhadu (M0) sice přináší drobné zlepšení chybovosti typu Gross Error High (GEH,GEH10 a DE) a snížení chybovosti v nejnižším pásmu, to je ale vykoupeno výrazným zvýšením chyb typu GEL a chybovosti ve třetím a čtvrtém frekvenční pásmu. Potvrdila se tedy tendence nestranného odhadu nadhodnocovat vyšší zpoždění (lagy) a detekovat tak frekvence nižší než referenční. Samotná energetická normalizace na plném okně (M2) přináší zlepšení GEH, které je opět kompenzováno výraznějším zhoršením GEL.

Table 3: Výsledky naměřené na KEELE databázi, druhá část

	procento chyb ve 2/3 oktávových pásmech (20% tolerance)				
PDA method	57Hz-88Hz	88Hz-141Hz	141Hz-225Hz	225Hz-353Hz	353Hz-565Hz
M0	28,3	5,1	6,7	2,8	38,1
M1	18,9	6,7	15,8	12,6	39,8
M2	17,0	5,0	12,6	8,6	41,5
M3	12,0	28,1	92,3	89,6	85,6
M4	14,2	7,9	18,5	12,5	42,4
M5	11,5	29,3	95,5	92,8	92,4
M6	13,8	10,2	23,3	15,9	41,5
M7	13,0	31,7	96,3	94,4	92,4

Energetická normalizace dohromady s nestranným odhadem (M3) má zcela nepřípustnou hodnotu GEL>60% a oba principy použité u této metody se evidentně podporují v nadhodnocování nižších frekvencí. Zkrácení okna na 3/4 původní délky vede u NCCF (M4) ke zhoršení GEL z 6.78% na 11,19%. Zkrácení na 2/3 (M6) pak oproti celému oknu sice snižuje GEH na polovinu, ale současně zhruba dvakrát zvyšuje GEL. Je tedy jasně vidět, že NCFF je schopna detekovat původní nízké frekvence i se zkráceným oknem, ale za cenu zvýšení GEL chyb.

7 Závěr

Představili jsme myšlenku, která je založena na energeticky normalizované korelační funkci a umožňuje detekovat základní frekvence i z okna kratšího než dvě periody signálu. Nestranný odhad korelační funkce v kombinaci s normalizovanou energií bohužel nedává dobré výsledky a systematicky velmi zvýhodňuje nižší frekvence. Avšak i samotná energetická normalizace nadržuje nižším frekvencím. Jistě bude na dalším zkoumáním, zda-li se podaří dalšími úpravami algoritmu více se na zkráceném okně přiblížit výsledkům původní autokorelační funkce na nezkráceném okně.

8 Poděkování

Tento výzkum je podporován Českou grantovou agenturou v rámci grantu SGS12/143/OHK3/2T/13 "Algoritmy a hardwarové realizace číslicového zpracování signálů".

References

[1] Hynek Bořil and Petr Pollák. Direct time domain fundamental frequency estimation of speech in noisy conditions. *in Proceedings of EUSIPCO 2004 (European Signal Processing Conference, Vol. 1)*, pages 1003–1006, 2004.

[2] Patricio De La Cuadra and Aaron Master. Efficient pitch detection techniques for interactive music. In *In Proceedings of the 2001 International Computer Music Conference, La Habana*, 2001.

[3] J. J. Dubnowski and R. W. Schafer. Digital hardware for pitch detection. *The Journal of the Acoustical Society of America*, 56(S1):S16–S16, 1974.

[4] G. Meyer F. Plante and A. Ainsworth. A pitch extraction reference database. In *Eurospeech*, pages 837–840, 1995.

[5] P. Sovka J. Uhlíř. *Číslicové zpracování signálů*. ČVUT Praha, 1995.

[6] Fei Sha and Lawrence K. Saul. Real-time pitch determination of one or more voices by nonnegative matrix factorization. In *in Advances in Neural Information Processing Systems 17*, pages 1233–1240. MIT Press, 2005.

[7] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech Coding and Synthesis, Elsevier Science*, pages 495–518, 1995.

Jan Bartošek
bartoj11@fel.cvut.cz