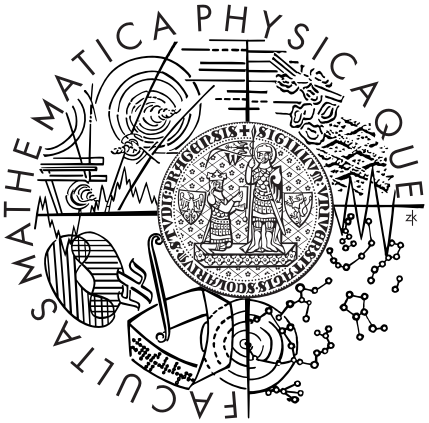
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Jan Kadlec

Přístupy počítačového vnímání hudebních nahrávek

Katedra teoretické informatiky a matematické logiky

Vedoucí bakalářské práce: Mgr. Vladan Majerech, Dr.
Studijní program: Informatika
Studijní obor: Obecná informatika

— 2006 —

Poděkování patří...

Mgr. Vladanu Majerechovi, Dr. za vedení práce, připomínky a za konzultace ohledně implementace projektu,

Ing. Petru Horákovi, Ph.D. za konzultace v oblasti zpracování signálu a analýzy řeči,

Prof. Ing. Janu Flusserovi, DrSc. za uvedení do oblasti počítačového vnímání a za objasnění metod používaných v klasifikaci

...a v neposlední řadě mým blízkým za motivaci, morální podporu a trpělivost.

Prohlašuji, že jsem svou bakalářskou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce a jejím zveřejňováním.

V Praze dne 10. srpna 2006

.....
Jan Kadlec

Obsah

Název práce: Přístupy počítačového vnímání hudebních nahrávek
Autor: Jan Kadlec
Katedra (ústav): Katedra teoretické informatiky a matematické logiky
Vedoucí bakalářské práce: Mgr. Vladan Majerech, Dr.
e-mail vedoucího: vladan.majerech@mff.cuni.cz

Abstrakt: V práci procházíme jednotlivá stádia vnímání skutečných hudebních signálů člověkem, vytváříme jejich modely a diskutujeme jejich implementaci. V úvodních kapitolách porovnáváme časově-frekvenční reprezentace a jejich rozklad na sinusové tóny a zbytkový šum. Kombinujeme různé přístupy k pospojování sinusových tónů do trajektorií. Dále se zabýváme metodami hledání parametrů not z trajektorií a metodami pro klasifikaci hudebních nástrojů. V závěru se zaměřujeme na globální hudebně teoretické vlastnosti nahrávek, jako je tempo, melodické fráze, akordický doprovod nebo tónina.

Klíčová slova: počítačové vnímání, časově-frekvenční reprezentace, modelování sinusovými tóny, hudební teorie, polyfonní transkripce hudby

---

Title: Approaches to computer cognition of musical recordings
Author: Jan Kadlec
Department: Department of Theoretical Computer Science and Mathematical Logic
Supervisor: Mgr. Vladan Majerech, Dr.
Supervisor's e-mail address: vladan.majerech@mff.cuni.cz

Abstract: In our study we deal with various stages of human cognition of real-life musical signals, create their models and discuss their implementation. The first part compares time-frequency representations and their factorization into sinusoids and background noise. We combine various approaches to coupling sinusoids into trajectories. Next we examine methods for the estimation of note parameters from trajectories and methods for musical instrument classification. The final part focuses on global music-theoretical features of recordings such as tempo, melodic phrases, chord accompaniment or key.

Keywords: computer cognition, time-frequency representations, sinusoidal modeling, music theory, polyphonic music transcription

1 Úvod

Okolní svět je člověku zprostředkován naprosto bezděčně a samozřejmě. Na začátku celého procesu dostanou naše smysly podnět, na konci naše mysl prostě vidí dům nebo slyší tón houslí. Jak ale probíhá „překlad“ informace do řeči našeho vědomí? A jak můžeme vysvětlit něco tak procítěného, jako je hudba, nemyslícímu stroji, schopnému jenom třídit jedničky a nuly?

Chtěl jsem prostudovat celý proces od prvního zachvění strun přes notový part, určený k reprodukci naposlouchané skladby, až po určení globálních vlastností písně, jako je tempo, předznamenání a akordy. Kdo někdy zkusil hledat, narazil na nepřeberné množství literatury. Má práce si klade za úkol podat čtenáři komplexní přehled oblasti počítačového vnímání hudby – skutečnosti jsou pro jednodušší vstřebání často zjednodušené a fakta jsou prokládaná mými postřehy a názory.

Práce je určena pro programátora s muzikálními sklony, pro hudebníka, který uvítá pomoc při interpretaci skladeb z poslechu, nebo prostě pro každého, kdo se chce dozvědět více o těch podivných zvucích a o pochodech, které nám umožňují říkat jim „hudba“.

2 Vnímání hudby člověkem

2.1. Zjednodušený model lidského sluchu

Zájemce o fyziologii sluchového ústrojí a další podrobnosti odkazuji na [2].

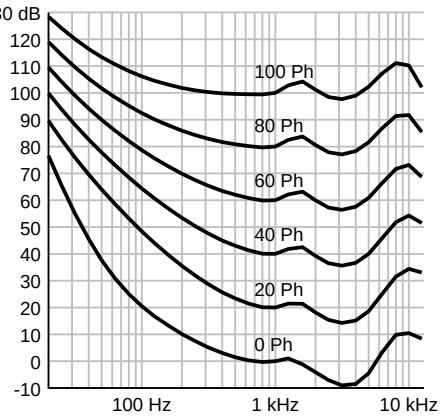
Zvuk tvoří změny tlaku prostředí v čase. V počítačích se pro tuto reprezentaci používá *pulzní kódová modulace* (PCM), která vznikne vzorkováním akustického tlaku v rovnoměrných intervalech a následnou kvantizací.

První zpracování zvuku se odehrává v hlemýždi, kam se přenáší ve formě vibrací tekutiny. Různé části hlemýždě rezonují na různých frekvencích a dráždí nervová zakončení umístěná po obvodu.

Nervová zakončení vysílají impulzy (jejich četnost a úvodní zpoždění odpovídá logaritmu intenzity vibrací – hlasitost se měří v logaritmické stupnici dB) do centrální nervové soustavy k dalšímu zpracování: korelace vjemů z obou uší, spojování různých frekvencí nebo porovnávání s krátkodobou a dlouhodobou pamětí.

Fyzikálními vlastnostmi hlemýždě a nervů je ovlivněno velké množství parametrů:

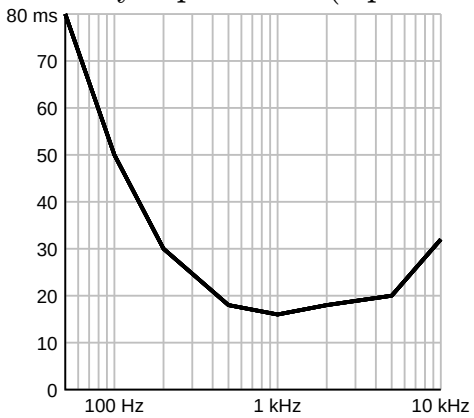
- **Citlivost na různé frekvence**
  Maximální slyšitelná frekvence se pohybuje kolem 20 kHz, se stářím se snižuje. Sluch je nejcitlivější na frekvence kolem 3 kHz. Obrázek znázorňuje křivky stejné hlasitosti podle ISO 226:2003 [3] – nejnižší odpovídá *prahu slyšení*, 120 fónů odpovídá *prahu bolesti* (ten už se z pochopitelných důvodů musí extrapolovat).



- **Maskování a spojování tónů**
  Silný vjem na jedné frekvenci (tj. na jednom místě hlemýždě) dráždí i sousední nervy, takže slabší vjemy v okolí zaniknou – jsou *maskovány*. Toho využívají například složitější metody pro kompresi zvuku. S tím souvisí i rozlišovací schopnost sluchu: když hrají dva tóny rychle za sebou nebo mají blízké frekvence,

má mozek problém jejich vjemy oddělit (na podobném principu nám splývají i jednotlivé odrazy ozvěny). Empiricky byla nalezena frekvenční stupnice (s jednotkou *bark*), ve které se maskovací křivka chová uniformně. Křivku lze aproximovat po částech lineární funkcí: útlum je 25 dB/bark k nižším frekvencím a 10 dB/bark k vyšším.

- **Rozpoznávání výšky a hlasitosti tónu**
  Tóny se skládají ze společně znějících sinusovek, ideálně v celočíselných poměrech (v přírodě tomu tak není, působí faktory jako pnutí strun klavíru nebo nenulová tloušťka flétny). Výšku tónu lze určit jako nejmenší společný násobek frekvencí. Je třeba několika nervových impulzů, abychom výšku a hlasitost tónu rozpoznali (na slabší zvuky potřebujeme víc času). V grafu je vynesena minimální doba trvání sinusovky, aby ji člověk nevnímal jako klepnutí.



Lidé mají dvě uši, což je pro sluch nepochybně přínosem a byla by škoda toho nevyužít i v našem modelu. Binaurální práh slyšení se sníží o 3 dB, maskovací křivka dokonce o 15 dB (při nenulovém rozdílu fází).

- **Rozpoznávání směru**
  Hlavní kritéria, pomocí nichž mozek zjišťuje směr vlevo/vpravo, jsou *rozdíl fází* a rozdílná doba příchodu prvních nervových impulzů. Ta může být způsobena *přirozeným zpožděním* (v řádu stovek mikrosekund) nebo *akustickým stínem* (hlava tlumí vysoké frekvence, např. na 5 kHz je útlum 25 dB). Směry nahoru/dolů a dopředu/dozadu jdou nahrubo rozpoznat opět pomocí stínu nebo *druhotnými odrazy* od ramen a ušních boltců, ale lidé spíš podvědomě otáčejí hlavu, aby směr zpřesnili na své hlavní ose. Rozlišovací schopnost je až 1° pod 1 kHz, ve vyšších frekvencích už jsou potíže s fází a potřebný úhel se rychle zvyšuje.

Zprostředkování informací z našich smyslů lze chápat jako určitou formu komprese [8] – lidský mozek získává svůj výpočetní výkon z paralelismu a redukce zpracovávaných dat je nutností. Jak vidíme, vývoj sluchu provázelo velké množství kompromisů. Přestože tedy nemůžeme očekávat, že se přiblížíme schopnosti mozku porovnávat

vzory, máme aspoň lepší vstupní data, se kterými pracujeme. Navíc se na skladbu můžeme dívat jako na celek – nemusíme se omezit na online zpracování.

2.2. Co je to hudba

Hudba je v podstatě jenom posloupnost zvuků. Je dobré si uvědomit, čím se liší od ostatních zvuků, a znalosti použít na specializaci algoritmů. Online výkladový slovník CoJeCo [1] definuje hudbu jako *druh umění, odlišující se od ostatních*

a) materiálem, tj. seskupením uměle vytvořených zvuků, převážně tzv. tónů (o určité výšce, síle, délce a barvě neboli témbru),

b) průběhem v reálném čase,

c) obsahem, jehož jádrem jsou psychické stavy a reakce člověka na vnější realitu.

Vlastnosti tónů jdou dobře matematicky popsat. Tón se skládá z periodických nebo kvaziperiodických kmitů, jejichž frekvence bývají celočíselné násobky *základního tónu* (*vyšší harmonické*). *Výška* tónu se definuje jako frekvence základního tónu v Hz. Pro vnímání hudby je důležitý relativní poměr výšek tónů, proto se používá dvanáctitónová logaritmická stupnice (*temperované ladění* s poměrem sousedních výšek $\sqrt[12]{2}$), která dobře aproximuje malé celočíselné poměry. Poměry mají historické hudební názvy, např. *oktáva* 2:1, *kvinta* 3:2, *velká tercie* 5:4, *malá tercie* 6:5 a další.

Kromě tónů používá hudba i *hluky* bez dobře definovatelné výšky – skládají se z několika málo kmitů (náš sluch je nestihne klasifikovat jako tón) nebo ze šumu. Hluky získají uměleckou hodnotu až časovým průběhem.

Většina hudebních zvuků jde v čase rozdělit na *nájezd* a *dozvuk*. Po nájezdu má zvuk největší energii – zde se určuje *síla* tónu (v dB), doba trvání dozvuku zase určuje *délku*. *Barva* zvuku se určuje nejobtížněji, protože nám nestačí jen jeden rozměr. Tvar frekvenční obálky a poměr energie jednotlivých harmonických vystihuje pojem *barva* nejlépe. Pro poslech je důležitá i *poloha* zdroje zvuku – lidský sluch je obzvlášť účinný pro oddělení zvuků přicházejících z různých směrů. Musíme také počítat s tím, že výška, barva, síla i směr se mohou v čase měnit, navíc je nutné očekávat *harmonii* – více zvuků znějících současně.

Časový průběh zvuku nám naopak může práci usnadnit. Jednotlivé tóny jdoucí za sebou označujeme jako *melodii*. Většina hudebních skladeb obsahuje jednu nebo několik *frází* – melodií, které se v různých obměnách opakují. Pojem můžeme rozšířit i na *rytmus* – zvuky vyskytující se v pravidelných intervalech. S tím souvisejí další termíny jako *takt* nebo *tempo*.

Psychické reakce člověka na hudbu se modelují jen velmi obtížně a není to ani předmětem této práce. Když porovnáme parametry sluchu a řeči, vidíme, že se sluch vyvíjel právě pro poslech řeči – schopnost předávat informace hrála a hraje v lidském životě ústřední roli. Při troše nadsázky má ale hudba vlastnosti, které jsou s řečí komplementární [4] (viz tabulku); cvičí, matou a „škádlí“ naše zařízení na klasifikaci řečových formantů, emocionálního podbarvení hlasu a přízvuku. Hudbu tedy můžeme chápat jako hru pro naše sluchové ústrojí, stejně jako optické klamy pro zrakové nebo lechtání pro hmatové.

Vlastnost	**Význam v hudbě**	**Význam v řeči**
Základní frekvence	výška tónu určující přesná	větná melodie ve většině jazyků jen pro artikulaci
Kvantizace v čase	rytmus, tempo určující zapisuje se přesná	rytmus řeči neurčující nezapisuje se přesnost není důležitá
Krátké pauzy	artikulace, výraz neurčující nezapisují se	součást souhlásek určující zapisují se
Formanty	barva zvuku neurčující	rozlišení samohlásek určující
Detaily nájezdů	barva zvuku podle nástroje	rozlišení souhlásek určující

3 Časově-frekvenční reprezentace

Historický přehled různých reprezentací v transkripci hudby lze nalézt například v [5].

V dnešní době se používá několik reprezentací frekvence v čase. Pro naše účely jsou některé vhodnější, jiné už méně – probereme jejich výhody i nevýhody. Výchozí bod všech reprezentací je PCM signál $x[m]$.

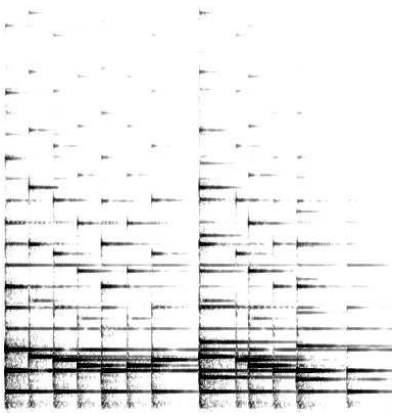
3.1. Fourierova transformace

Fourierovu transformaci znal už Euler, ale masově se začala používat až po vydání Cooleyho a Tukeyho článku [6]. Tam popisují efektivní algoritmus na výpočet diskrétní Fourierovy transformace v čase $\mathcal{O}(N \log N)$.

Aby bylo možné lokalizovat frekvence v čase, používá se *krátkodobá Fourierova transformace* (STFT) s délkou okna N

$$\mathrm{STFT}_m[k] = \sum_{n=-N/2}^{N/2-1} x[n+m]\, w[n]\, e^{-\frac{2\pi k n i}{N}}.$$

STFT se nepočítá pro každý vzorek: m se zvyšuje o konstantní krok h (*hop-size*). Okénko w odfiltruje postranní laloky těch sinusovek, jejichž perioda nedělí N.



Je mnoho druhů okének s různými vlastnostmi, podrobnosti viz např. [7]. Při zpracování hudebního signálu se používá von Hannovo okno $w[n] = \frac{1}{2} + \frac{1}{2}\cos(\frac{2\pi n}{N})$, protože vnáší nejmenší chybu do lokalizace frekvencí.

Výhodou STFT je to, že jde spočítat velmi rychle – nevýhodou je, že neurčitost ve frekvenci a čase je na všech frekvencích konstantní a nám by se hodila spíš logaritmická závislost.

Obrázek vpravo znázorňuje typický výstup STFT několika klavírních tónů – na vodorovné ose je čas, na svislé lineární frekvenční stupnice.

3.2. Waveletová transformace

Mějme *škálu* $s > 0$, *posun* t a *matečnou waveletu* – funkci ψ. *Spojitou waveletovou transformaci* (CWT) funkce f definujeme jako

$$f_\psi(s,t) = \int |s|^{-1/2} f(x)\, \psi\left(\frac{x-t}{s}\right)^* \mathrm{d}x.$$

Při zpracování zvuku lze použít například Mortlettovu nebo Gaborovu waveletu [9]. CWT dobře využívá neurčitost v čase a ve frekvenci, ale musí se počítat pomocí konvolucí, což je pomalé. Někdy je variabilní přesnost i na škodu – chceme-li např. hledat společný začátek sinusovek, po CWT se rozmaže každá jinak.

Existuje i velmi rychlá *diskrétní* dyadická verze (DWT), ale ta trpí přesně opačným problémem než STFT – na vysokých frekvencích je moc malé rozlišení ve frekvenci a na nízkých zase v čase. Pro filtraci šumu a zpracování obrazu se ale osvědčila. Článek [10] porovnává praktické použití CWT a STFT na zpracování zvuku.

3.3. Vícepásmová Fourierova transformace

Myšlenka spočívá v rozdělení frekvencí na pásma (buď logicky, nebo pomocí kaskády převzorkování) a počítání různých STFT pro každé pásmo. Výhodou je nastavitelnost přesnosti v každém pásmu (i adaptivně podle charakteru hudby) a zpomalení oproti STFT jen o konstantní faktor. Problém je s navázáním pásem a s redundancí výstupních dat.

3.4. Další reprezentace

Různé *constant-Q reprezentace* se snaží sadou vhodně zvolených filtrů dosáhnout logaritmické škály, ale CWT problém řeší většinou lépe. V lékařství a při zpracování radarových snímků se používá *Wigner-Villeova distribuce*, ale kvůli velké interferenci sinusovek je pro zpracování hudby nevhodná. Další transformace, které jsem zavrhnul, jsou *Radonova* (součty energie na přímkách – vhodné pro nalezení „okamžité frekvence“), *chirp-z transformace* (jako STFT, ale bázi tvoří „sinusovky“ s lineární změnou frekvence), *chiplety* (spojení myšlenek waveletů a chirp-z) nebo různé metody založené na adaptivní filtraci Wigner-Villeovy distribuce.

4 Dominantní sinusové tóny

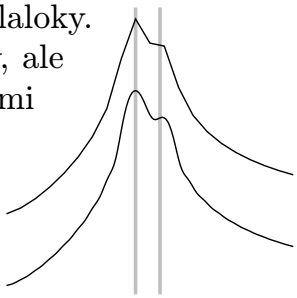
Tato kapitola je výtah z [11] a [12] doplněný o poznámky.

Dalším krokem je vyjádření každého časového okamžiku jako součet několika málo sinusových tónů. Metody jsou většinou vyvinuty pro STFT, ale je možné je přizpůsobit i jiným reprezentacím.

Sinusové tóny si můžeme představit jako lokální maxima frekvenční reprezentace. Kvůli rychlosti pracují metody hladově – po nalezení extrémů se zbytkové spektrum prohlásí za šum a další, „schované“ tóny se už nehledají. Z výstupu se nakonec vyhází sinusovky, jejichž amplituda se moc neliší od okolí (lidský sluch vnímá jen ty, které jsou ve svém pásmu dominantní).

4.1. Doplnění nulami (zero padding)

Rozšíříme-li okénko STFT před transformací o nuly, získáme vyšší rozlišení a tedy i přesnost. Ve frekvenční oblasti se metoda chová jako interpolace normovanou $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$. Nevýhodou je, že při vyšší výpočetní složitosti nezískáme žádnou novou informaci a že se jako falešné frekvence objeví i postranní laloky. Po doplnění nulami lze vracet už přímo extrémy, ale lepší výsledky dostaneme v kombinaci se složitějšími metodami.



Na obrázku vidíme dvě blízké sinusovky po STFT – po doplnění nulami (dolní křivka) se objeví obě lokální maxima.

4.2. Interpolační metody

Hledáme-li lokální maxima spojité navzorkované funkce (jako je frekvenční reprezentace), vyplatí se jít s přesností pod vzorkovací mřížku. V praxi se používají různé metody. Popíšu nejjednodušší z nich – *parabolickou interpolaci*.

Najdeme lokální maximum amplitudy X_n a vedeme parabolu vzorky X_{n-1}, X_n a X_{n+1} (případně použijeme metodu nejmenších čtverců na širší okolí). Zpřesněné maximum frekvence a amplitudy leží ve vrcholu paraboly.

4.3. Metody využívající fázi

Fázové metody jsou přesnější než interpolační, mají větší odolnost vůči šumu a nalezené sinusovky se tolik neovlivňují.

Obyčejná STFT přiřadí amplitudu a fázi středům obdélníků časově-frekvenční mřížky. Myšlenka *nového přiřazení* (reassignment) je posunout tato přiřazení do nových bodů, kde budou lépe vystihovat vlastnosti signálu (do těžišť původních přihrádek). Pro lokální maximum s indexem n je nová frekvence definována jako

$$\hat{f} = n\frac{F_s}{N} - \mathrm{Im}\left(\frac{\mathrm{STFT}'[n]}{\mathrm{STFT}[n]}\right)\frac{F_s}{2\pi},$$

kde F_s je vzorkovací frekvence a STFT' je STFT signálu s okénkem vzniklém derivací původního okénka w. Podobnou operaci můžeme provést i v čase. Do podrobností se tomu věnuje [5].

Princip *vokodéru* spočívá v tom, že frekvenci můžeme vyjádřit jako rozdíl rozbalených fází v sousedních vzorcích. Spočítáme úhel svíraný dvěma STFT kolem vzorků x a $x+1$ v komplexní rovině:

$$\hat{f} = \frac{F_s}{\pi}\arccos\left(\frac{\mathrm{STFT}_x[n]\cdot\mathrm{STFT}_{x+1}[n]}{\left|\mathrm{STFT}_x[n]\right|\left|\mathrm{STFT}_{x+1}[n]\right|}\right).$$

Nakonec zpřesníme i amplitudu – dopočítáme ji z původní frekvence $f = nF_s/N$ a (spojité) Fourierovy transformace použitého okénka $\mathcal{F}_w(\omega)$:

$$\hat{A} = 2\,\frac{|\mathrm{STFT}_x[n]|}{\left|\mathcal{F}_w\left(2\pi\frac{\hat{f}-f}{f}\right)\right|}.$$

Existují i další metody, ale většinou jsou s některou z těchto dvou ekvivalentní. Pro zmírnění vlivu šumu je možné zprůměrovat frekvence z několika transformací poblíž vzorku x.

4.4. Azimut

Někdy se kromě frekvence a amplitudy zavádí ještě *azimut* – úhel ve vodorovné rovině, pod kterým sinusovka přichází. Nejsnazší je pro každou sinusovku porovnat fáze a amplitudy levého a pravého kanálu a odhadnout úvodní zpoždění nervových impulzů z hlemýždě.

Zpoždění lze najít také přímo pomocí *korelace* levého a pravého kanálu, ale signál se musí filtrovat, aby se azimuty sinusovek rozlišily. Zájemce o problematiku binaurálního slyšení odkazuji na článek [17].

5 Trajektorie, noty a nástroje

Zde jsem se netradičně pokusil propojit dva pohledy na věc do jediné metody. Přiřazení sinusovek tónům se obvykle provádí buď nezávisle v různých časových okamžicích (např. Klapuri [15]), nebo se sinusovky spojí v čase do *trajektorií* a dál se nezpracovávají.

Začněme spojováním do trajektorií. Můžeme předpokládat, že se amplituda, frekvence ani azimut nebudou měnit moc rychle.

5.1. Lineární predikce

Nejúspěšnější model, který jsem objevil, je *lineární predikce* – v jistém smyslu je to vlastně jen další frekvenční reprezentace. Používá se například při kompresi hlasového signálu (tam je vstupem přímo PCM) nebo v ekonomii. Pro predikci vývoje sinusovek byla prvně použita ve [13], zájemce tam najde i srovnání se staršími metodami. Model byl dále rozšířen například ve [14].

Předpokládejme, že se aktuální hodnota $x[n]$ (teď nerozlišujeme, zda jde o frekvenci, amplitudu nebo azimut) dá aproximovat lineární kombinací p hodnot z minulých okének

$$\tilde{x}[n] = \sum_{i=1}^{p} a_i \, x[n-i].$$

Pro výpočet koeficientů a_i se osvědčila *Burgova metoda*, která rekurzivně minimalizuje průměr dopředné a zpětné chyby predikce. Odvození najdete ve [13], zde uvedu rovnou algoritmus.

$f \leftarrow x, \quad b \leftarrow x, \quad a \leftarrow \{1\}$
for $(m = 0 \ldots$ délka historie $h)$
  $f' \leftarrow f$ bez prvního prvku
  $b' \leftarrow b$ bez prvního prvku
  $k \leftarrow -2\, b' \cdot f' \,/\, \left(b'^2 + f'^2\right)$
  $f \leftarrow f' + kb', \quad b \leftarrow b' + kf'$
  $a \leftarrow \{a[0], a[1], \ldots, a[m], 0\} +$
    $k\{0, a[m], a[m-1], \ldots, a[0]\}$

5.2. Váhová funkce

Predikcí stanovíme nejpravděpodobnější pokračování trajektorie. V dalším časovém snímku však může být možných pokračování několik, nebo mohou dokonce chybět (trajektorii nám například „rozhází“ bicí nástroje). V praxi se postupuje tak, že každé možné pokračování je ohodnoceno *váhovou funkcí*. Její parametry je třeba vyvážit tak,

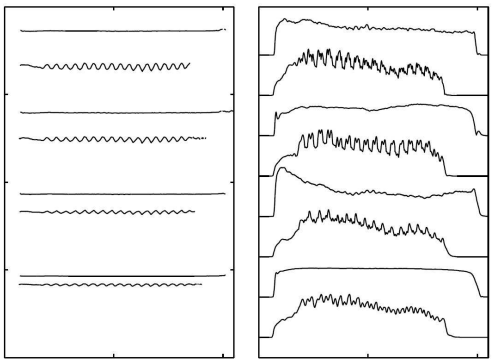
aby co nejlépe vystihovaly hledané nástroje, případně pro každý nástroj použít jinou.

Váhová funkce může obsahovat například

- kritéria pro začátek nové trajektorie a její konec
- jak moc nám vadí přeskočené snímky (bez přiřazení sinusovky)
- jakou váhu přisuzujeme (nejčastěji kvadratickým) odchylkám od predikované amplitudy, frekvence a azimutu
- jak moc nám vadí rychlé kmity v predikovaných hodnotách (*tremolo* v amplitudě, *vibrato* ve frekvenci)
- počet „agentů“ – kolik možností jedné trajektorie sledovat

Minimalizace váhové funkce probíhá buď hladově (snímek po snímku), nebo se použije dynamické programování (pro každou trajektorii se sleduje několik nejlepších možností a na konci se jedna zafixuje).

5.3. Seskupování trajektorií do tónů



Trajektorie, které odpovídají vyšším harmonickým jednoho tónu, se v čase chovají hodně podobně. Na obrázku vidíme časové průběhy frekvencí (vlevo) a amplitud dvou takových tónů.

Náš mozek si spojuje změny, které nastaly naráz, do jediného vjemu – pokusíme se o podobný přístup. Seskupování trajektorií do tónů může hrát roli i ve váhové funkci z přechozí podkapitoly.

- **Vzdálenost koeficientů predikce různých trajektorií**
  Zvolíme si metriku na koeficientech predikce (stačí třeba vážená euklidovská) a blízké trajektorie označíme jako možné kandidáty na seskupení. Pokud se trajektorie chovají podobně jen v určitých souvislých úsecích, můžeme je na tyto úseky rozsekat.
- **Vyšší harmonické**
  Trajektorie reprezentují časový průběh nějakých vyšších harmonických. Zkusme je tedy seskupit do tónů: každé trajektorii

přiřadíme takovou základní frekvenci, aby byla pokud možno co nejvíce sdílena i dalšími trajektoriemi. Při tom můžeme využít informaci o barvě nástrojů, případně ji zpětně z nalezených tónů odvodit a proces iterovat.

Potíže nastanou, když si uvědomíme, že tóny v harmonických vztazích (jak je v hudbě běžné) také obsahují celočíselné poměry frekvencí. Je tedy nevyhnutelné, že spolu budou interferovat a jednotlivé trajektorie bude nutné přiřadit více tónům.

V [15] je problém vyřešen postupným odfiltrováváním tónů (od nejhlasitějšího). Při našem přístupu ale můžeme zachovat jednoznačné přiřazení *trajektorie* → *tón*: najdeme-li kolizi, trajektorii logicky rozštěpíme a každou část sledujeme zvlášť. Parametry jednotlivých částí odvodíme z vyšších harmonických, které jsme už k základním tónům přiřadili.

Také je vhodné využít znalost vývoje nástroje v čase – amplitudová obálka nájezdu může pomoci v obtížných případech, jako je rychlé opakování stejného tónu nebo hraní stejného tónu dvěma různými nástroji (jejich vlastnosti spočítáme z jiných částí skladby).

- **Dominance v pásmech**
  Když si uvědomíme, že v jednom frekvenčně-azimutovém pásmu může být pro člověka dominantní jen jeden nástroj, můžeme se ho tam pokusit uměle najít. Tím jsme schopni rozpoznat i doprovodné nástroje, které jsou zvukově skryty za ostatními a člověk se na ně musí specificky soustředit, aby je objevil (např. akordický doporovod nebo basový part). V moderní hudbě se navíc používají digitální efekty, které ořezávají spektrum nástrojů – takové tóny jinou metodou objevíme jen stěží.

Zbylé trajektorie a sinusovky, které byly moc slabé, krátké nebo neodpovídaly nalezeným tónům, můžeme přidat do zbytkového šumu.

5.4. Hluky

Na chvíli si odpočiňme od trajektorií a podívejme se, co se zbytkovým šumem. Můžeme předpokládat, že jej tvoří už výhradně jen hluky a šumové pozadí skladby. U hluků máme tu výhodu, že se nemusíme zabývat základním tónem. Velmi dobře pro ně funguje například metoda *časově-frekvenčních předloh* [16].

Vezměme časově-frekvenční reprezentace zbytkového šumu N a předlohy P (zvlášť pro každý hledaný hlukový nástroj). Pro každou kombinaci časů t_h, t_p se spočítá podobnost spekter N_{t_h} a P_{t_p}. Řádky výsledné časově-časové matice se posunou tak, aby sloupce reflektovaly časový průběh předlohy. Jsou-li pak hodnoty v některém sloupci podobné (tzn. hlukový nástroj udržoval svůj charkteristický průběh po celou dobu trvání), prohlásí se za začátek noty.

Předlohy můžeme iterativně zlepšovat: za nové předlohy zvolíme medián$(N_{t_i} \dots N_{t_i+l})$, kde t_i jsou začátky nalezených not a l je délka původní předlohy.

5.5. Parametry not

Články zabývající se klasifikací nástrojů se zaměřují buď na samostatné tóny nebo na fyzikální modelování nástrojů. Z první skupiny mě zaujal zajímavý přehled příznaků a jejich kombinací [18], vyčerpávající přehled fyzikálních vlastností nástrojů dává [2].

My potřebujeme ale spíš zjistit parametry not – výšku (tu už máme jako základní frekvenci), délku, hlasitost a chování v čase. Potřebujeme také robustně odlišit nástroje od sebe. Tato a následující podkapitola je jen stručný přehled možných přístupů.

- **Základní frekvence**
  Různé nástroje mají různé frekvenční rozsahy. Chování jednoho nástroje v nejnižších a nejvyšších tónech se také dost liší. Je tedy vhodné určit, jak vysoko nástroj hrává, a má-li široký rozsah, odvodit jeho parametry pro několik výšek zvlášť.

- **Časová obálka**
  Nástroje se v čase chovají různě. Stádia časového vývoje tónu lze zjednodušeně rozdělit v maximu amplitudy na *nájezd* a *dozvuk* [19]. Hlasitost tónu se nejlépe určuje ke konci nájezdu, délku tónu lze definovat jako dobu, kterou trvá pokles amplitudy z maxima například o 12 dB.

  Pro rozlišení nástrojů lze použít například přesnost frekvence při nájezdu, intenzitu tremola a vibrata v různých frekvenčních pásmech a pro různé uvažované periody kmitů (25 Hz – „drsnost“, 6 Hz – „kolísání“), dobu nájezdu nebo tvar amplitudy dozvuku (používá se aproximace exponenciální funkcí, polynomem nebo spliny).

- **Frekvenční obálka**
  Tvoří hlavní součást barvy zvuku. Je ji možné najít *lineární predikcí* (LP) nebo ještě lépe *lineární predikcí se změněnou škálou* (WLP) [18], ale nejvíce se mi osvědčily *mel-frekvenční kepstrální koeficienty* (MFCC), případně spojené se svými derivacemi v čase (DMFCC) [20]. MFCC se definují jako koeficienty kosínové transformace průměrů amplitud v pásmech definovaných speciální frekvenční stupnicí s jednotkou *mel*. Pro nás je důležité si uvědomit, že MFCC obálka je nezávislá na základní frekvenci.

  Kromě průměrných MFCC a DMFCC nájezdu a dozvuku lze nástroje rozlišit dalšími momenty MFCC (rozptylem a šikmostí), průměrnou „rozmazaností" trajektorií, pohybem mediánu spektra po vyšších harmonických nebo prostě povoleným rozsahem základních frekvencí.

- **Globální vlastnosti**
  Jak často a jak rychle nástroj hrává? Hraje polyfonně, nebo zní vždy jen jeden jeho tón? Je to sólový nástroj? Hraje pořád, nebo ho můžeme omezit jen na některé pasáže? Jak je korelován s rytmem?

  Všechny tyto globální příznaky a mnohé další (jejich určování se věnuje další kapitola) lze také použít při rozlišování nástrojů – záleží jen na charakteru skladby a fantazii programátora.

5.6. Klasifikace nástrojů

Při tak velkém množství příznaků by byl zázrak, kdyby měly dva tóny určený stejný nástroj. Z příznaků je třeba vybrat nebo nakombinovat několik málo takových, které nám nástroje rozliší nejlépe (tomuto procesu se říká *redukce dimenzionality příznakového prostoru*). Můžeme začít z ničeho a přidávat ty nejlepší příznaky (SFG), nebo začít se všemi a po jednom je odebírat (SBG). Můžeme příznaky chápat jako metrický prostor, *natočit* jeho osy tak, aby se veškerý rozptyl (PCA) nebo rozlišovací schopnost (*Fischer Discriminant Analysis*) přesunula jen do několika málo souřadnic, a zbylé zahodit. Možností je nespočetně, zájemce odkazuji na literaturu věnující se umělé inteligenci nebo rozpoznávání vzorů.

Nakonec musíme příznakový prostor rozdělit na oblasti, které přísluší jednotlivým nástrojům. Populární *Gaussian mixture model* (GMM) reprezentuje každou oblast součtem (mnoha) Gaussových

skvrn. Velmi jednoduchý, ale mocný, je *k-NN klasifikátor* – každý nástroj vybere „typické“ reprezentanty a výhercem je nástroj, jehož k-tý nejbližší reprezentant od klasifikovaného bodu k němu má nejblíž (dle použité metriky rozlišujeme *euklidovský k-NN*, *harmonický k-NN* a další).

Parametry modelů se musí nastavit v závislosti na tom, kolik je dostupných dat, jak jsou ovlivněny šumem, na kolik chceme klasifikovat tříd nebo kolik času chceme klasifikaci věnovat.

6 Hudebně teoretické vlastnosti

Tato kapitola vychází z [21], kde najdete i historické přístupy, podrobnější vysvětlení myšlenek a detaily implementace.

Dostali jsme se do stádia, kdy už máme reprezentaci, kterou je možné použít k přepisu skladby do MIDI nebo do notového partu: melodii pro různé nástroje, hlasitost not, začátek a délku. Umíme dokonce určit i podrobnosti, jako je azimut, vibrato nebo jemné nepřesnosti ve výšce tónu a v barvě. Hudba pro nás ale znamená mnohem víc, než jenom noty – samotná nota nám ještě nic neřekne. Je třeba se do posloupností not podívat hlouběji a analyzovat jejich vnitřní vztahy. Systém dostane schopnost předvídat, domýšlet si a přidávat významy. Získané informace můžeme zpětně použít pro přesnější určení not.

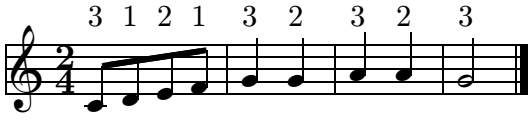
Metody budu popisovat pomocí pravidel. Ve skutečné implementaci se úspěšné splnění pravidel skóruje a skóre se dynamickým programováním pro celou skladbu maximalizuje.

6.1. Metrum, takt a tempo

Události v hudbě jsou kvantizované v čase. Je třeba dodat, že v reálných skladbách bývá kvantizace nepřesná (což přidává nahrávce na lidskosti), nebo se plynule mění (zvláště v expresivní interpretaci klasické hudby).

Začátkům kvantizačních intervalů říkám *doby*. Kvantizaci lze rozdělit do několika úrovní; vyšší úroveň má své doby na každou druhou nebo třetí dobu nižší úrovně. Vyjímečně se mohou objevit i poměry 5:1 nebo 7:1. V některých písních se poměry mezi úrovněmi v průběhu skladby mění.

Nejnižší úroveň kvantizace se nazývá *metrum*, každá nota musí koincidovat s některou jeho dobou. Nejvyšší úroveň, ve které se nějaká doba metra vyskytuje, označuji jako *tíhu* doby (čísla nad notami písně *Kočka leze dírou*).



Pravidla pro určování metra a tíhy:

→ V dobách začínají noty. V těžkých dobách bývají delší a hlasitější.

→ Metrum je lepší, když zůstává konstantní. Musí-li se měnit, probíhají změny plynule (expresivní hra) nebo skokově (změna tempa). Je třeba počítat s nepřesnostmi kvantizace.

→ Na těžkých dobách začínají melodické fráze a mění se v nich akordy.

→ Stejné melodické fráze na různých místech by měly mít stejné rozdělení síly dob.

Interval mezi nejtěžšími dobami se nazývá *takt*. Délka taktu se obvykle pohybuje od 400 ms do 1600 ms. Druhá nebo třetí nejtěžší úroveň tvoří *rytmus*, počet rytmických dob za minutu se označuje jako *tempo*. V hudebním zápisu se vztah rytmu a taktu značí *předznamenáním*, například $\frac{\mathbf{2}}{\mathbf{4}}$ (●·○·), $\frac{\mathbf{3}}{\mathbf{4}}$ (●·○·○·) nebo $\frac{\mathbf{6}}{\mathbf{8}}$ (●··○··).

V moderní hudbě se vyskytuje *synkopace* (noty na těžkých dobách mírně předbíhají rytmus) a *swing* (noty na lehkých dobách se vůči němu opožďují). Budeme-li takové skladby analyzavat, měla by tomu být váhová funkce uzpůsobena.

6.2. Melodické fráze

Na úrovni taktu lze vysledovat opakující se sekvence not, kterým se říká *melodické fráze*. Ústřední melodickou frázi nazýváme *téma* skladby. Lidský mozek si noty melodických frází slučuje do jediného celku.

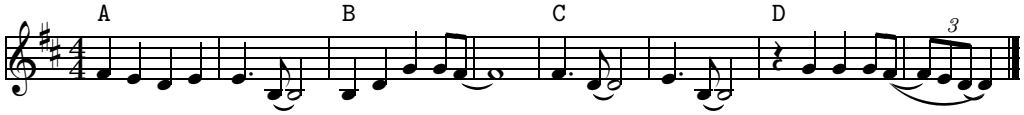
Pravidla pro rozdělení na melodické fráze:

→ Každý výskyt nějaké fráze hraje jediný nástroj; hlasitost, azimut i výška not jsou podobné. Každá nota patří nejvýše do jedné fráze.

→ Fráze začínají po pauzách a po dlouhých notách. Začátky frází a těžké doby jsou korelovány.

→ Délka frází se pohybuje kolem osmi not.

Stejná fráze na různých místech skladby může být transponovaná (posunutá ve frekvenci), hrát jinou rychlostí nebo jiným nástrojem. Vyskytují se i malé obměny frází (přidané noty nebo změny výšek).



V písni *Eight days a week* je sekvence frází dobře rozpoznatelná: `ABABCCAD`.

6.3. Kontrapunkt

Noty si v hlavě automaticky spojujeme do trajektorií, které nazýváme *melodie*. *Kontrapuntem* rozumíme několik melodií znějících současně – chceme nalézt pravidla jejich oddělení:

→ Jedna melodie upřednostňuje noty s podobnou barvou, hlasitostí a azimutem (jde použít predikce jako v kapitole 5). Každá nota patří aspoň do jedné melodie, počet melodií je přitom co nejmenší. (Pravidlo je podobné pravidlu u melodických frází.)

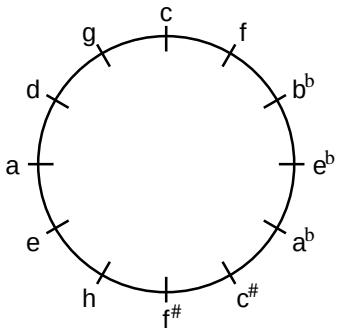
→ Výška melodie by se měla měnit pomalu. V rychlých tempech povolíme menší rozdíly výšek.

→ Melodie by neměly obsahovat moc časových mezer.

→ Není dobré, když se melodie mezi sebou protínají. Spolu znějící melodie by se měly pohybovat podobným způsobem.

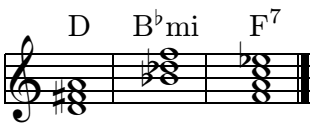
6.4. Harmonie a akordy

Předpokládejme, že výšky not máme reprezentovány dvanáctitónovou logaritmickou stupnicí z druhé kapitoly. Pro zápis výšky použijeme hybrid mezi německou a anglickou notací: oktávu označíme číslem a stupnici v rámci oktávy značkou $\{c\ c^\sharp\ d\ e^\flat\ e\ f\ f^\sharp\ g\ a^\flat\ a\ b^\flat\ h\}$ nebo alternativně $c^\sharp = d^\flat$, $d^\sharp = e^\flat$, $f^\sharp = g^\flat$ a $g^\sharp = a^\flat$. Výška $a5$ odpovídá frekvenci 440 Hz.



Nejdříve si povíme něco málo o harmonických vztazích not. Kritéria „podobnosti“ not pro náš mozek jsou *blízká výška* (např. c-$c^\sharp$, c-d) nebo *jednoduchý poměr zákadních frekvencí* (první zajímavý poměr je *kvinta* 3:2, např. c-g). Vzdálenost výšek v kvintách je dobře vidět na obrázku kvintového kruhu.

Akord je souzvuk tří nebo více tónů. Odvozuje se od *základního tónu* postupně přidanými intervaly, většinou kombinací *velkých* (c-e) a *malých* (c-$e^\flat$) *tercií*. Akordy značíme velkými písmeny a akordovou značkou, která určuje přidané intervaly (příklady: Ddur = d-$f^\sharp$-a, B$^\flat moll$ = $b^\flat$-$d^\flat$-f, F^7 = f-a-c-$e^\flat$). Akordy nás zajímají, protože jsou z velké části zodpovědné za emoce, které v nás hudba vzbuzuje, a protože se jimi řídí melodie doprovodu.



Nalezení základního tónu akordu umožní následující pravidla (akordovou značku určíme až s pomocí tóniny):

→ Akordy se přiřazují celým úsekům skladby (jsou pro všechny noty v úseku společné).

→ Při akordu od základního tónu c nejčastěji hrají *harmonické* noty $c, g, e, e^\flat, b^\flat, g^\flat$ (seřazené podle četnosti). Zbylé noty označíme jako *blue* noty. Pro jiné akordy se výšky not odpovídajícím způsobem posunou (*transponují*).

→ Blue noty bývají obklopeny notami, které jsou jim blízké ve výšce. Blue noty se vyskytují na lehkých dobách a harmonické noty na těžkých.

→ Akordy se nejčastěji mění na těžkých dobách, při změně se základní tón akordu nepohybuje po kvintovém kruhu moc daleko.

V jazzu jsou tato pravidla záměrně porušována, používají se velmi složité akordy a jejich netradiční sekvence. Podrobně se akordickými vztahy v jazzu zabývá [22].

6.5. Tónina

Akordické sekvence získají svou „vyprávěcí“ schopnost až v kontextu *tóniny*. Tónina vymezuje noty, které můžeme v melodii očekávat, noty mimo tóninu se vyskytují jen velmi zřídka.

Většinou se používají tóniny, které tvoří v kvintovém kruhu souvislý úsek, například pentatónická (*c-d-e-g-a*) z orientální hudby a country. Klasické západní stupnice jsou o dvě výšky bohatší: nejčastěji narazíte na *jónskou* (durovou) stupnici *c-d-e-f-g-a-h*. Je zajímavé, že poloha počátečního tónu v kvintovém kruhu různých obratů této stupnice vyjadřuje náladu: *lydická* stupnice od f (f-*g-a-h-*...) zní až dětinsky, *dórská* od d je melancholická a *aiolskou* od a neboli *přirozenou mollovou* už prostupuje deprese. Přehled dalších stupnic najdete skoro v každé učebnici hry na nějaký hudební nástroj.

Hledání tóniny je celkem jednoduché – všechny uvažované tóniny si pro každý uvažovaný úsek počítají skóre a vezme se maximum. Každá nota z úseku přispěje každé tónině do skóre hodnotou, která odpovídá její relativní výšce (podle tabulky pro durové a mollové tóniny).

	$a^\flat$	$e^\flat$	$b^\flat$	f	**c**	g	d	a	e	h	$f^\sharp$	$c^\sharp$
c dur	4	4	3	8	10	9	7	7	9	8	4	4
c moll	7	9	3	8	10	9	7	4	4	8	4	4

→ Jediné pravidlo zní, že by se tónina neměla měnit moc často.

7 Závěr

V jednotlivých kapitolách jsme postupně prošli jednotlivá stádia vnímání hudby člověkem a nastínili jejich simulaci a algoritmickou implementaci. Z reprezentace zvuku funkcí tlaku v čase jsme prošli několika transformacemi až k notovému zápisu pro různé nástroje. Skladbu jsme nakonec zanalyzovali po globání stránce a identifikovali její hudebně teoretické vlastnosti jako metrum, melodické fráze nebo harmonickou strukturu.

Práce se snažila propojit dva zdánlivě neslučitelné světy umění a techniky. Poodkryla spoustu nápadů a myšlenek, které jen volají po bližším prostudování. Doufám, že čtenáře zaujala a umožnila mu nahlížet na hudbu v novém světle.

Na přiloženém médiu jsou implementace některých popsaných metod z kapitol 3 a 4. Další algoritmy z této práce a nové nápady se budou objevovat na internetové adrese `http://rrrola.wz.cz/icari`.

Literatura

[1] *Online výkladový slovník CoJeCo.* `www.cojeco.cz`

[2] Syrový, V.: *Hudební akustika.* Nakladatelství AMU, 2001

[3] ISO 226: *Normal equal-loudness-level contours.* ISO Acoustics, 2003.

[4] Wolfe, J.: *From idea to acoustics and back again: The creation and analysis of information in music.* Proc. WPAC'03, 2003.

[5] Hainsworth, S. W.: *Techniques for the automated analysis of musical audio.* Ph.D. thesis, Cambridge Univ., 2004.

[6] Cooley, J. W., Tukey, J. W.: *An algorithm for the machine calculation of complex Fourier series.* Math. Comput. 19, str. 297–301, 1965.

[7] Harris, F. J.: *On the use of windows for harmonic analysis with the discrete Fourier transform.* Proc. IEEE 66(1), str. 51–83, 1978.

[8] Smith, S. W.: *The Inner light theory of consciousness.* California Tech. Publishing, 2001.

[9] Daubechies, I.: *Ten lectures on wavelets.* SIAM, 1992.

[10] Johnson, M. K.: *The spectral modeling toolbox: a sound analysis/synthesis system.* M.A. thesis, Dartmouth College, 2002.

[11] Keiler, F., Marchand, S.: *Survey on extraction of sinusoids in stationary sounds.* Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx'02), str. 51–58, 2002.

[12] Lagrange, M. et al.: *Improving sinusoidal frequency estimation using a trigonometric approach.* Proc. DAFx'05, str. 110–115, 2005.

[13] Lagrange, M. et al.: *Enhanced partial tracking using linear prediction.* Proc. DAFx'03, str. 1–6, 2003.

[14] Lagrange, M. et al.: *Tracking partials for the sinusoidal modeling of polyphonic sounds.* Proc. ICASSP'05, 2005

[15] Klapuri, A.: *Signal Processing methods for the automatic transcription of music.* Ph.D. thesis, Tampere Univ. of Technology, 2004.

[16] Yoshii, K. et al.: *AdaMast: A drum sound recognizer based on adaptation and matching of spectrogram templates.* Proc. MIREX, 2005.

[17] Viste, H., Evangelista, G.: *Binaural source localization.* Proc. DAFx'04, str. 145–150, 2004.

[18] Eronen, A.: *Automatic musical instrument recognition.* M.S. thesis, Tampere Univ. of Technology, 2001.

[19] Jensen, K.: *Envelope model for isolated musical sounds.* Proc. DAFx'99, str. 1–5, 1999.

[20] Combrinck, H. P. a Botha, E. C.: *On the mel-scaled cepstrum.* Univ. of Pretoria, 1996.

[21] Temperley, D.: *The cognition of basic musical structures.* MIT Press, 2001.

[22] Velebný, K.: *Jazzová praktika.* Praha 1967, 1983.