

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

**Fakulta elektrotechniky
a komunikačních technologií**

SEMESTRÁLNÍ PRÁCE

Brno, 2022

Bc. Viktor Slezák



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ

DEPARTMENT OF TELECOMMUNICATIONS

SVĚTELNÉ ANIMACE PRO SYSTÉM SPECTODA NA ZÁKLADĚ ANALÝZY PARAMETRŮ Z HUDEBNÍCH NAHRÁVEK

LIGHT ANIMATIONS FOR THE SPECTODA SYSTEM BASED ON THE ANALYSIS OF PARAMETERS FROM
MUSIC RECORDINGS

SEMESTRÁLNÍ PRÁCE
SEMESTRAL THESIS

AUTOR PRÁCE
AUTHOR

Bc. Viktor Slezák

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. Matěj Ištvanek

BRNO 2022



Semestrální práce

magisterský navazující studijní program **Audio inženýrství**

specializace Zvuková produkce a nahrávání

Ústav telekomunikací

Student: Bc. Viktor Slezák

ID: 203745

Ročník: 2

Akademický rok: 2022/23

NÁZEV TÉMATU:

Světelné animace pro systém Spectoda na základě analýzy parametrů z hudebních nahrávek

POKYNY PRO VYPRACOVÁNÍ:

Vytvořte systém pro výpočet parametrů z hudební nahrávky s důrazem na dynamickou, rytmickou a akordickou strukturu. Využijte nejnovější přístupy založené na metodách strojového učení pro extrakci relevantních parametrů. Získaná data analyzujte a na jejich základě navrhněte a naprogramujte algoritmus generující specifický kód „SpectodaCode“ pro následné vytváření světelných animací. Výstupem práce bude jednoduché webové rozhraní, které po nahrání hudební skladby vygeneruje unikátní světelné animace. Cílem semestrálního projektu je popis parametrů a informací, které lze smysluplně použít pro generování světelných animací. Semestrální práce bude obsahovat implementaci skriptů pro výpočet parametrů a návrh struktury výsledného systému. V budoucí diplomové práci budou parametry využity a optimalizovány pro generování kódu, který bude data převádět na sekvence animací pro ovládání světel.

DOPORUČENÁ LITERATURA:

- [1] MÜLLER, Meinard. Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. Cham: Springer International Publishing, 2015. ISBN 978-3-319-21945-5.
- [2] CARSAULT, Tristan, NIKA, Jérôme, ESLING, Philippe a ASSAYAG, Gérard. 2021. Combining Real-Time Extraction and Prediction of Musical Chord Progressions for Creative Applications. Electronics, vol. 10, no. 21: 2634. DOI <https://doi.org/10.3390/electronics10212634>.

Termín zadání: 1.10.2022

Termín odevzdání: 12.12.2022

Vedoucí práce: Ing. Matěj Ištvanek

doc. Ing. Jiří Schimmel, Ph.D.

předseda rady studijního programu

UPOZORNĚNÍ:

Autor semestrální práce nesmí při vytváření semestrální práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

SLEZÁK, Viktor. *Světelné animace pro systém Spectoda na základě analýzy parametrů z hudebních nahrávek*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2022, 48 s. Semestrální práce. Vedoucí práce: Ing. Matěj Ištvánek

Prohlášení autora o původnosti díla

Jméno a příjmení autora: Bc. Viktor Slezák

VUT ID autora: 203745

Typ práce: Semestrální práce

Akademický rok: 2022/23

Téma závěrečné práce: Světelné animace pro systém Spectoda na základě analýzy parametrů z hudebních nahrávek

Prohlašuji, že svou závěrečnou práci jsem vypracoval samostatně pod vedením vedoucího/závěrečné práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené závěrečné práce dále prohlašuji, že v souvislosti s vytvořením této závěrečné práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

Brno
..... podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucímu diplomové práce panu Ing. Matěj Ištvánek za odborné vedení, konzultace, trpělivost a podnětné návrhy k práci.

Obsah

Úvod	11
1 Teorie	12
1.1 MIR - Music information retrieval	12
1.1.1 Historie	12
1.1.2 Řetězec zpracování - pipeline	13
1.1.3 Současné problémy	15
1.2 Parametrizace hudebních nahrávek	15
1.2.1 Reprezentace audio signálů	15
1.2.2 Časová oblast	16
1.2.3 Frekvenční oblast	17
1.2.4 DFT - Diskrétní Fourierova transformace	19
1.2.5 STFT - Short-time Fourier transform	21
1.2.6 Dynamika hlasitost a intenzita	23
1.2.7 Barva	24
1.3 Detekce nástupů a analýza tempa skladby	25
1.3.1 Využití energie signálu	26
1.3.2 Využití spektra signálu	28
1.3.3 Detekce periodicity	30
1.4 Klasifikace žánrů a nálady	31
1.5 Chromavektory	31
1.6 Dostupná řešení	31
1.6.1 Librosa	31
1.6.2 Madmom	33
1.6.3 Aubio	33
1.6.4 Hodnocení extrakce informací	33
1.7 Systém Spectoda	33
1.8 Hudební signál jako animace	33
2 Výsledky studentské práce	34
2.1 Návrh výsledného systému	34
2.1.1 Uživatelské rozhraní	34
2.1.2 Parametry hudební nahrávky	34
2.1.3 Systém pro generování animací	37
2.1.4 Databáze bloků animací	40
2.2 Výběr vhodných metod pro extrakci vlastností z hudební nahrávky .	40
2.2.1 Detekce dob a tempa	40

2.2.2	Analýza chromavektorů	42
2.2.3	Efektivní hodnota signálu	42
Závěr		43
Literatura		44
Seznam symbolů a zkratek		47
Seznam příloh		48

Seznam obrázků

1.1	Řetězec procesů MIR [20]	14
1.2	Zobrazení časového průběhu signálu	17
1.3	Reprezentace tónu E zahráneného na basovou kytaru. a) Časová oblast b) Frekvenční spektrum	18
1.4	Časově spojitý signál a diskrétní signál	20
1.5	Signál o délce $1s$ s počáteční frekvencí $10Hz$ a koncovou frekvencí $30Hz$ a) Původní signál b) Signál s okénkem od $0,2s$ do $0,5s$ c) Signál s okénkem od $0,35s$ do $0,65s$ d) Signál s okénkem od $0,5s$ do $0,8s$ [16]	22
1.6	Nahrávka piana a) Amplituda nahrávky b) Frekvenční spektrum nahrávky zobrazené pomocí spektrogramu	23
1.7	Tón A5 zahránený na klavír a) Amplituda tónu b) ADSR obálka tónu .	25
1.8	Detkete nástupů perikusivního zvuku a) Amplituda nahrávky b) Lokální energie signálu $E_{xw}(n)$ c) Derivace lokální energie signálu s půlvlnným usměrněním $\Delta_E(n)$	27
1.9	Výpočet spektrálního toku pro nahrávku piana a) Amplituda nahrávky b) Spektrogram nahrávky c) Spektrální tok bez komprese d) Spektrální tok s kompresí spektra $\gamma = 1$	29
1.10	Výpočet spektrálního toku z mel spektrogramu pro nahrávku piana a) Amplituda nahrávky b) Mel spektrogram nahrávky c) Spektrální tok	30
2.1	Blokové schéma postupu uživatele webovou stránkou	34
2.2	Struktura třídy <i>ChromaVector</i>	35
2.3	Struktura třídy <i>Loudness</i>	35
2.4	Struktura třídy <i>Segment</i>	36
2.5	Blokový diagram výběru datasetu	37
2.6	Blokový diagram struktury rozhodovacího procesu	39
2.7	Blokové diagramy tříd <i>Dataset</i> , <i>AnimationBlock</i>	40
2.8	Porovnání metod detekce dob na úryvku skladby Oh-Darling!. a) Melspekrogram b) Detekce dob pomocí Librosa c) Detekce dob pomocí Madmom d) Detekce dob pomocí Aubio	41
2.9	Porovnání metod detekce dob na úryvku skladby Oh-Darling!. a) Melspekrogram b) Detekce dob pomocí Librosa c) Detekce dob pomocí Madmom d) Detekce dob pomocí Aubio	42

Seznam tabulek

1.1 Typické procesy na základně vstupních a výstupních dat. [20] 15

Úvod

V rámci semestrální práce jsou popsány možnosti pro dolování parametrů z hudebních nahrávek a jejich analýzu. Tyto techniky jsou využity pro získání potřebných informací o skladbě. Například data o tempu a rozmístění dob, žánr a tónové či spektrální rozložení skladby. Dále je navržena struktura algoritmu sloužícího pro převod získaných parametrů na sekvence animací kompatibilních se systémem Spectoda.

Práce je rozložena do tří na sebe navazujících cílů. Prvním z cílů je průzkum vědních oborů soustředících se na danou problematiku. Například MIR (Music information retrieval - Obor zabývající se vyhledávání informací v hudebních dílech). Z existujících výzkumů jsou vybrány postupy analýzy hudebních signálů vyhovující pro použití v rámci výsledného algoritmu.

Druhým cílem práce je navrhnout vnitřní strukturu výsledného algoritmu převádějícího získané parametry na sekvence animací pro systém Spectoda. Důležitým úkolem je vymyslet jak bude docházet k takovému přenosu a co dané parametry ovlivní v rámci generování unikátních sekvencí animace.

Poslední třetí cíl se zaobývá vytvořením funkčního systému pro získávání parametrů z hudební nahrávky. Důraz je kladen na využití dostupných moderních metod analýzy hudebních signálu.

1 Teorie

Semestrální práce se zabývá problematikou MIR. Popsánou v kapitole 1.1.

Nabízejí se otázky jak by měla daná animace reagovat na konkrétní děj skladby. Jakým způsobem navrhnout strukturu algoritmů a co by měly získané parametry ovlivňovat při vytváření animací.

V této části jsou popsány následující segmenty: Teorie zpracování hudební nahrávky pomocí známých algoritmů. Například důležitým algoritmem je FT¹ popsaná více v bodě 1.2.3, její varianty pak v bodech 1.2.4 a 1.2.5. Nabízené moderní metody strojového učení s využitím hlubokých neuronových sítí při detekci tempa skladby 1.3 a určení žánru 1.4. Struktura a možnosti systému Spectoda pro generování interaktivních světelných animací jsou podrobně popsány v bodě 1.7.

1.1 MIR - Music information retrieval

Music information retrieval je interdisciplinární vědní obor sestředící se na získávání informací z hudebních nahrávek. Jsou zde kombinovány znalosti mnoha oborů jako jsou muzikologie, psychoakustika, strojové učení, zpracování signálů a další.

Výstupy jeho výzkumu jsou využívány v populárních technologických aplikacích. Jednou z aplikací je personalizované doporučování hudebních skladeb, které se nachází v moderních streamovacích platformách. Další využití je v programech pro mixování hudby používaných diskžokeji k plynulejší práci díky analýze tempa a klíčových částí skladby. Tyto technologie se nachází v mnoha dalších aplikacích a s rozširováním digitálního audia jejich důležitost stále roste.

1.1.1 Historie

V tomto bodě je napsán souhrn historie MIR z knihy [20]. MIR se začíná objevovat na přelomu devatenáctého a dvacátého století s příchodem moderních statistických metod. Začínají se objevovat pokusy o aplikování statistických metod na hudební partitura. Protože ještě nebyly natolik dostupné počítače jednalo se spíše o ruční práci s partiturami a tabulaturou. Z grafických notací se analyzovaly rysy skladeb a specifikovaly charakteristiky hudebního díla. S příchodem počítačů do výzkumných laboratoří se v letech 1960 až 1970 začalo více rozvíjet zpracování signálů a s tím související možnosti analýzy hudebních nahrávek pomocí počítačů. V těchto letech se poprvé začaly objevovat nyní známé termíny jako „computational musicology“ a „music information retrieval“. První oblast výzkumu se soustředila na analýzu tempa skladby. Z důvodu nízké popularity se však výzkum zpomalil. Tento útlum

¹Fourier transform - Fourierova transformace.

pokračoval až do roku 1990 kdy výzkumu MIR pomohly dvě změny. První důležitou změnou byly rostoucí databáze digitální hudby, která se stala lehce dostupná pro výzkumné týmy. Druhým bodem který přispěl k vývoji MIR byl nárůst výpočetního výkonu počítačů a nižší náklady s nimi spojené. Díky těmto změnám se stal výzkum dostupnější a jednodužší na realizaci [20].

Poté v říjnu roku 2000 bylo uspořádáno první mezinárodní symposium soustředící se na MIR. Z této mezinárodní konference se stala tradice a vybudovala se kolem ní velká komunita nazývaná ISMIR². Každoročním vyvrcholením ISMIR je právě výše zmíněná konference, na které vědci z celého světa prezentují pokroky v oblasti výzkumu MIR. Zanedlouho naté v roce 2005 byl v rámci této konference představen model MIREX³ sloužící jako správa zásad pro hodnocení pokroků ve výzkumu MIR[5].

1.1.2 Řetězec zpracování - pipeline

V tomto bodě je popsán postup zpracování dat v palikací MIR. Jedná se o systém, kterým jsou data zpracovávána a určuje standardně využívaný řetězec jak při tvorbě algoritmů postupovat .

Vstupními daty se rozumí zejména hudební informace v digitální podobě. Tyto vstupní data se rozlišují do více typů. Mohou to být obrázky představující digitální formu zápisu hudby pomocí symbolů „not“ [20]. Například digitalizovaná partitura. Dalším možným typem je „digitální hudba“. Jedná se o hudbu čistě v „digitálních notách“ představujících sadu příkazů. Například zápis v MIDI⁴. Nejrozšířenější formou vstupních dat jsou digitální nahrávky představující audio signály.

Pre-processing - předzpracování signálu Na začátku řetězce je zařazen blok předzpracování vstupních signálů. Tento blok se postará o připravení dat do podoby vhodné pro extrakci vlastností. Jedná se například o komprimaci komplexních vstupních signálů popsaných níže. Nebo je signál převáděn z časové do frekvenční oblasti. Více o technikách předzpracování je popsaáno v bodě 1.2.

Feature extraction - extrakce vlastností signálu Podle požadovaných vlastností pro extrakci je využíváno modelů popsaných v bodech 1.3, 1.4 a 1.5. S rostoucí popularitou strojového učení začaly při extrakci vlastností hudební nahrávky převládat kombinace hlubokých neuronových sítí. Tyto kombinace umožňují přesnější

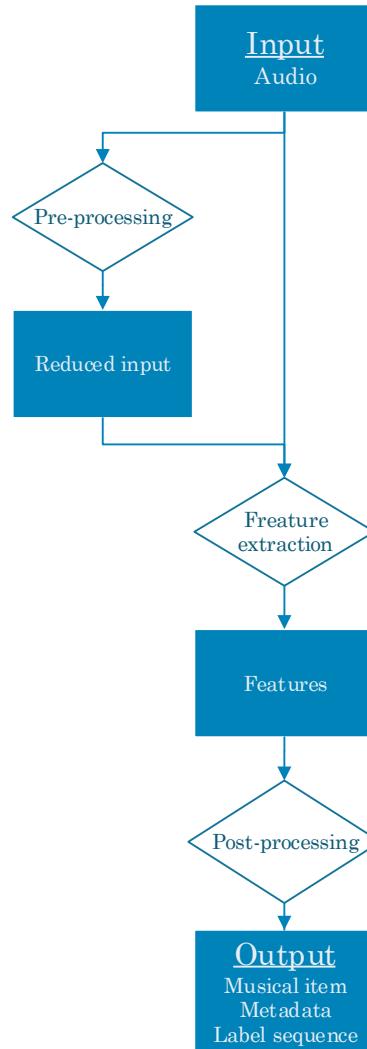
²International Society of Music Information Retrieval - Mezinárodní sdružení pro MIR

³The Music Information Retrieval Evaluation eXchange - komunitní rámec pro hodnocení pokroků výzkumu v oblasti MIR. Obhospodařovaný laboratoří International Music Information Retrieval Systems Evaluation Laboratory sídlící na University of Illinois. [5].

⁴Musical Instrument Digital Interface - Volně dostupný hudební standart specifikující hardwerové a softwarové požadavky pro digitální přenos hudební notace a komunikace mezi nástroji.[26]

parametrizaci a menší chybovost.

Post-processing - konečné zpracování Posledním blokem v řetězci je tzv. „post-procesing“ zajišťující zpracování a optimalizaci získaných dat. Post-procesing zpracuje data do požadované formy. V některých případech také dokáže ovlivnit přesnost zpracování.



Obr. 1.1: Řetězec procesů MIR [20]

Digitální hudební nahrávka se jako forma vstupních dat stala hlavním trendem výzkumu MIR. Je to způsobeno zejména dostupností velkých databází nahrávek ke kterým mají vědecké instituce přístup a nepotýkají se s problémy souvisejícími s autorskými právy [20].

Z důvodu velké komplexnosti vstupních dat se využívá několik technik komprimace signálů. Slučování vícekanálových nahrávek do mono signálu. Převzorkování

signálu na nižší vzorkovací kmitočty, a rozložení na krátké překrývající se úseky, ze kterých mohou být nezávisle extrahovány jejich vlastnosti[11]. Výsledkem je kolekce paralelně složených sekvencí hodnot jednotlivých vlastností, které se následně zpracují na požadovaná výstupní data.

Data	Vyhledávání informací	Klasifikace a odhad	Sekvenční značení
Audio	Identifikace kopie „coverů“, Řazení skladeb, Měření podobnosti, Získání otisku, Generování seznamu skladeb	Identifikace umělce a skladatele, Žánr a nálada, Určení tempa	Extrakce melodie, Odhad akordů, Detekce nástupů, Segmentace

Tab. 1.1: Typické procesy na základně vstupních a výstupních dat. [20]

1.1.3 Současné problémy

1.2 Parametrizace hudebních nahrávek

V této kapitole je popsán audio signál. Jak vzniká, jeho reprezentace v číslicovém zpracování a základní principy práce s audiosignálem. V bodech 1.2.6 a 1.2.7 jsou popsány parametry získávané z audio signálu. Získané parametry slouží pro přesnější popis skladby.

1.2.1 Reprezentace audio signálů

Hudební dílo může být reprezentováno více formami. Jako tradiční médium pro její ukládání ještě před vznikem záznamu sloužily noty a další typy zápisů pomocí symbolů. Výsledné hudební dílo ale představuje mnohem více než počáteční notový zápis. Každý hudebník a hudební nástroj do skladby předává svou unikátnost. Při hře se noty začnou proměňovat v harmonické zvuky, hladké melodie a nástroje vzájemně rezonují. Každý z hudebníků do skladby přináší svou interpretaci. Jinak reagují na tempo zvýrazňují odlišné noty a liší se jejich artikulace. Všechny tyto proměnné ve výsledku způsobují, že dílo není jen mechanické přehrání napsané partitury. Jeho součástí se stává unikátní přednes umělce a použitého hudebního nástroje [16].

Z fyzikálního hlediska při interpretaci díla vznikají zvukové vlny šířící se vzduchem. Tyto vlny jsou reprezentovány kmítáním částic v pružném prostředí. V takovém prostředí jsou částice na sebe vázány a vytvářejí soustavu oscilátorů. Pokud dojde k vychýlení jedné částice ze své rovnovážné polohy, vlivem okolních částic dochází k působení pružných sil a vzniká její kmítání. Zároveň dochází k vzájemnému

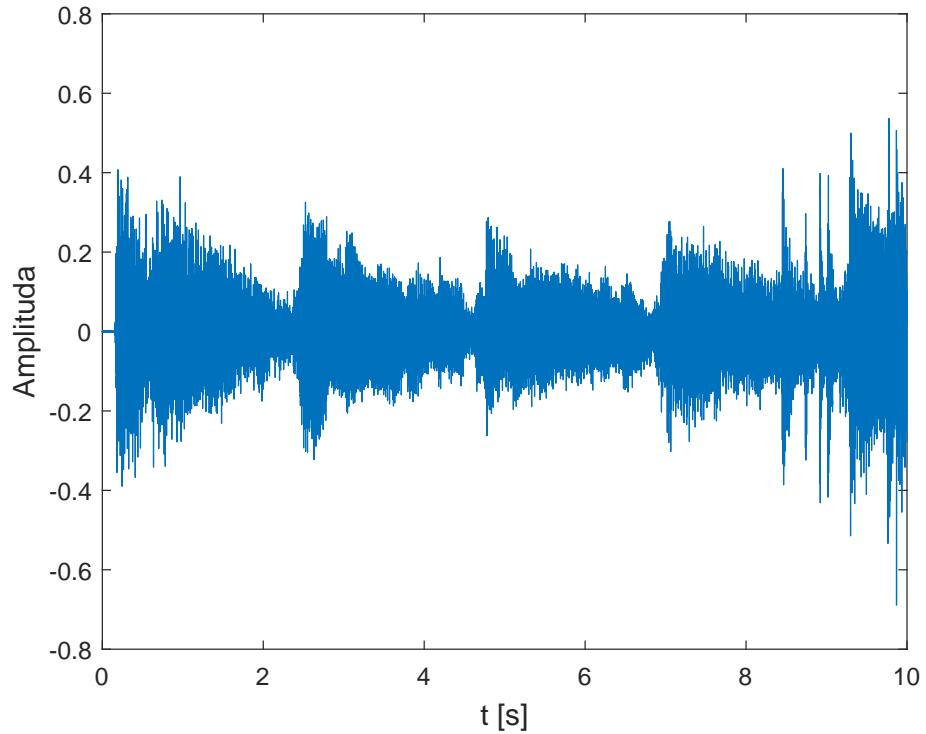
rozkmitání okolních částí a prostředím se začne šířit vlna. Jednotlivé částice kmitají pouze kolem své rovnovážné polohy. Nedochází tak k přenosu látky ale pouze energie a hybnosti[4]. Popsané kmitání jsme schopní zachytit pomocí akustických měničů. Je získán analogový signál šířících se zvukových vln nazývaný jako audio signál. Pojem audio je označován řetězec sloužící k záznamu, přenosu a reprodukci zvuků v mezích lidského slyšení. Avšak v audio signálu se už nenachází přesná reprezentace not a jejich parametrů jako jsou čas nástupu, tón, délka trvání, dynamika. Díky tomu je analýza hudebních signálů obtížným úkolem a je ovlivněna reprezentací interpreta, stavbou nástroje, akustikou prostoru a vnímáním posluchače. Zmíněnými problémy se zabývá samostatný vědní obor s názvem psychoakustika. Nejdůležitějšími parametry audio signálu které jsou podrobně popsány níže definujeme: frekvence, výška tónu, dynamika, intenzita, hlasitost a barva [16].

1.2.2 Časová oblast

Základní reprezentací audio signálu je tzv. zobrazení v **časové oblasti**. V časové oblasti představují číslicový signál vzorky. Jednotlivé vzorky udávají hodnotu signálu v daném čase. Počet vzorků vztažených na jednotku času určuje vzorkovací frekvence signálu. Důležitým pravidlem pro vzorkování signálu je Shannonův-Nyquistův vzorkovací teorém

$$f_{vz} > 2f_{max} \quad (1.1)$$

kde f_{vz} je vzorkovací frekvence a f_{max} je maximální frekvence v audio signálu [1]. Pokud jednotlivé vzorky zobrazíme graficky získáme průběh signálu v čase viz obr.1.2.



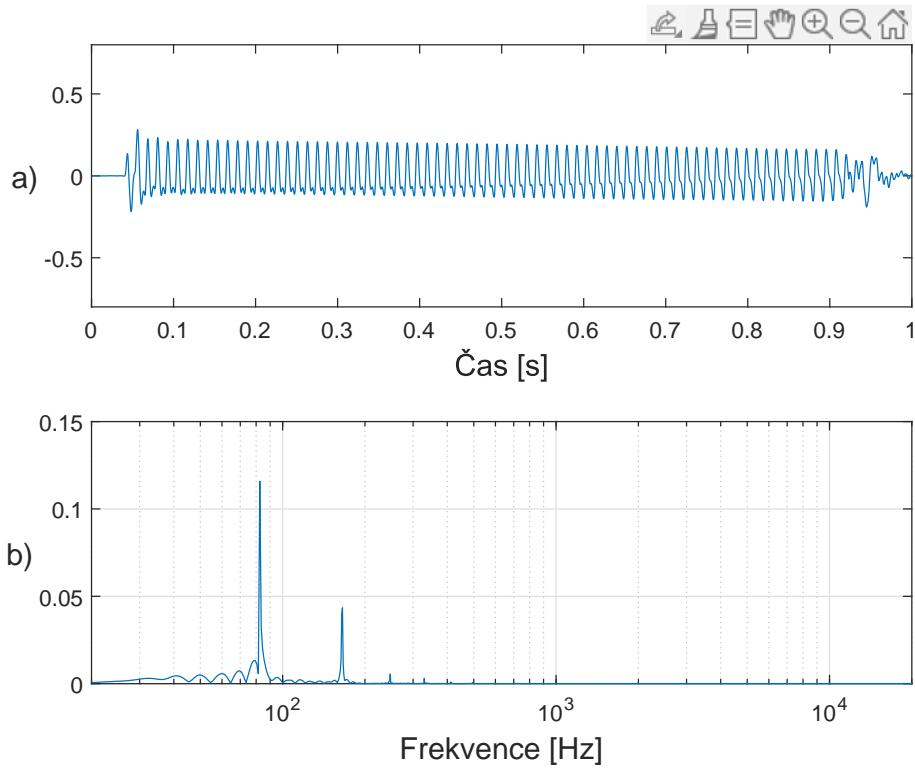
Obr. 1.2: Zobrazení časového průběhu signálu

Tato reprezentace audio signálu poskytuje informace o průběhu amplitudy signálu. Využívá se například pro výpočet energie signálu popsaný v bodě 1.3.1.

1.2.3 Frekvenční oblast

Pro získání dalších informací o hudebním díle se využívá transformace signálu do frekvenční oblasti umožňující odlišné znázornění struktury signálu.

Ve frekvenční oblasti je signál reprezentován jeho frekvenčními složkami popsanými v komplexním tvaru. Spektrum představuje rozložení původní části signálu na jednotlivé frekvenční složky popsané funkcí sinus. Kde reálná složka obsahuje informaci o magnitudě „velikosti“ funkce sinus. Imaginární složka komplexního čísla pak udává počáteční fazu. V grafu jsou poté zobrazeny frekvenční složky se kterých se signál skládá viz obr. 1.3.



Obr. 1.3: Reprezentace tónu E zahráného na basovou kytaru. a) Časová oblast
b) Frekvenční spektrum

Jako názorný důvod proč je transformace do frekvenční oblasti přínosná je dán příklad. Na nástroj je zahrán tón, který je zaznamenán. V časové oblasti je možné určit délku tónu a jeho průběh podle ADSR obálky popsané v bodě 1.2.7. Pokud je ale potřeba zjistit výšku tónu a určit notu, tak se jedná o složitý proces. Díky transformaci do frekvenční oblasti je patrná fundamentální frekvence tónu. Označována také první harmonická. Tato frekvence udává výšku tónu a je tak možné stanovit notu která byla zahrána.

Pro získání frekvenčního spektra signálu je třeba transformovat signál s časové oblasti. K tomu slouží několik úprav Fourierovy transformace podle vlastností vstupního signálu. Tyto metody jsou dále nazývány jako Fourierovy řady, Diskrétní časová Fourierova transformace a Diskrétní Fourierova transformace. V případě audio signálu se vyžívá zejména Diskrétní Fourierovy transformace popsané v bodě 1.2.4.

Fourierovy transformace zkráceně definována jako transformace převádějící signál mezi časovou a frekvenční oblastní pojemci harmonických signálů jež popisují funkce sinus a cosinus [1]. Funkce sinus a cosinus představují komplexní exponenciály.

$$e^{iat} = \cos(\alpha t) + i \sin(\alpha t) \quad (1.2)$$

Fourierova transformace pro **spojitý neperiodický signál** je pak zapsána jako

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i\omega t} dt \quad (1.3)$$

kde $\omega = 2\pi f$ a udává uhlovou frekvenci. Magnituda $|X(f)|$ je potom funkcí sudou [21].

Pro signál který je **spojitý a periodický** se definují Fourierovy řady a integrální funkce je počítaná pouze pro jednu periodu signálu

$$c[f_k] = \frac{1}{T_0} \int_0^{T_0} x(t)e^{-ik\omega_0 t} dt \quad (1.4)$$

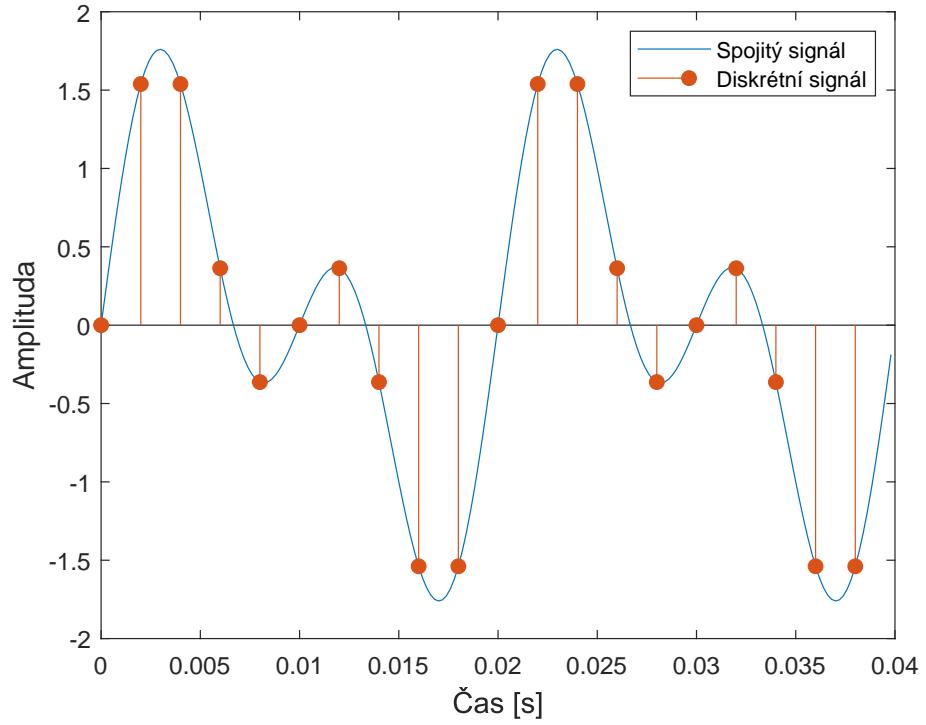
kde $f_k = k \times f_0$ a $k \in (\mathbb{Z}; 0, \pm 1, \pm 2, \dots)$. Vypočítané spektrum je diskrétní a neperiodické [21].

Pokud je vstupní signál diskrétní hovoříme o Diskrétní Fourierově transformaci která je více popsána v následujícím bodě 1.2.4.

Po transformaci signálu do frekvenčního spektra jsou data signálu v komplexním tvaru a jejich magnituda $|X(f)|$ je funkce sudá, tím pádem symetrická kolem nuly a fáze $\varphi_x(f)$ je funkcí lichou čili je středově symetrická. Pro analýzu audio signálů se využívá kladná část spektra.

1.2.4 DFT - Diskrétní Fourierova transformace

Pokud jsou signály zpracovávány pomocí výpočetních procesorů, tak může být uložen pouze omezený počet hodnot signálu. To znamená, že analogový signál spojitý v čase musí být převeden na signál digitální tvz. signál diskrétní, který je není spojitý v čase. Diskrétní signál je potom vhodný pro číslicové zpracování. Proto jsou pospané algoritmy DFT přizpůsobené právě pro zpracování diskrétních signálů nespojitých v čase.



Obr. 1.4: Časově spojitý signál a diskrétní signál

Opět jsou dány odlišné definice pro signál diskrétní neperiodický a diskrétní periodický. Protože se jedná o signál diskrétní, tak zde odpadají integrální funkce. Pokud se jedná o signál **diskrétní neperiodický** jeho výsledné spektrum bude spojité a hovoříme o Fourierově transformaci diskrétní v čase. Matematicky je zapsána v následujícím tvaru

$$X(f) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\Omega n} \quad (1.5)$$

kde

$$\Omega = 2\pi(f/f_s) \quad (1.6)$$

a f_s je vzorkovací frekvence signálu.

Protože v praxi signál není nikdy nekonečně dlouhý, tak je možné jej poskládat za sebe a vytvořit tak signál periodický. Pro periodické signály je výpočet DFT zapsán ve tvaru

$$X[f_k] = \sum_{n=1}^{N_0} x[n]e^{-i\Omega_k n} \quad (1.7)$$

kde

$$\Omega_k = 2\pi \frac{f_k}{f_s} \quad (1.8)$$

$$f_k = \frac{k f_s}{N_0} \quad (1.9)$$

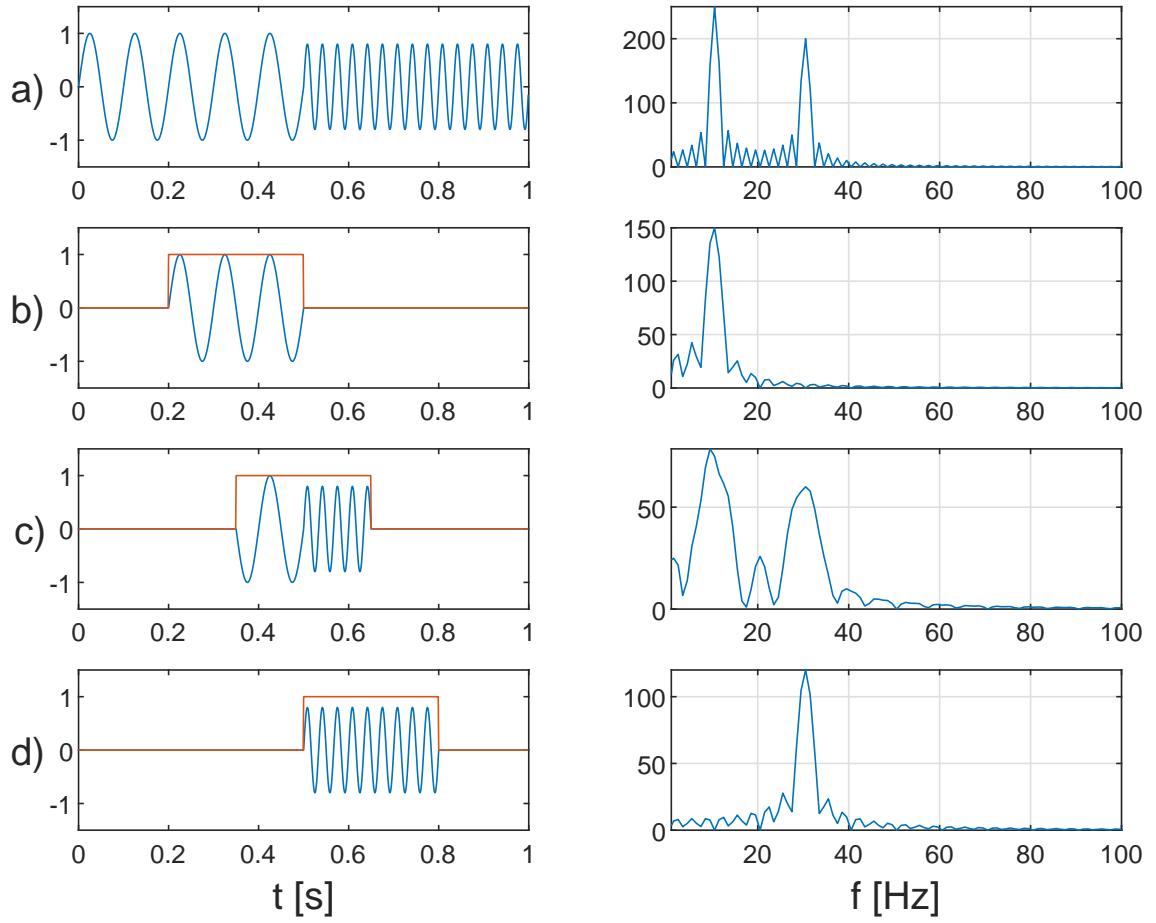
a $k \in (\mathbb{Z}; 0, N_0 - 1)$. N_0 udává počet vzorků v jedné periodě signálu. Hustota spektra K v takovém případě odpovídá $K = N_0$.

Ze strany výpočetní náročnosti je takto definovaný algoritmus neefektivní a výpočetně náročný. Pro výpočet DFT je zapotřebí velkého množství operací složitost algoritmu je pak zapsána jako $O(N^2)$. Proto pokud počet vzorků N dosahuje většího množství je ve spoustě případů tento algoritmus příliš pomalý a neefektivní pro praktické využití.

Počet potřebných operací může být výrazně redukován. Na vývoj efektivního řešení výpočtu DFT se zasloužil Carl Friedrich Gauss a Joseph Fourier zhruba před dvěma sty lety. Tento algoritmus nazýváme Rychlou Fourierovou transformací zkráceně FFT. Počet operací pro výpočet takového algoritmu byl snížen na $O(N \log_2 N)$ [16]. Například při použití vzorků $N = 2^{10} = 1024$. FFT vyžaduje $N \log_2 N = 10240$ operací namísto $N^2 = 1048576$ operací při použití DFT. Jak je vidět snížení výpočetní náročnosti je velké a exponenciálně roste s větším počtem vzorků N . Vynález FFT změnil odvětví zpracování signálů a je dnes využíván v miliardách telekomunikačních zařízeních. Stejně tak i ve zpracování a analýze zvukových signálů hraje důležitou roli [16].

1.2.5 STFT - Short-time Fourier transform

V roce 1946 Dennis Gabor představil STFT jako možnost zařazení frekvenčních složek ke konkrétnímu času signálu [23]. Fourierova transformace umožňovala převod signálu z časové oblasti do frekvenční ale nebylo zřejmé v jakém časovém úseku signálu se získané frekvenční složky nachází. Hlavní myšlenkou STFT je, že namísto analyzování celého signálu je analyzována pouze jeho malá část. Za tímto účelem je definovaná tzv. okénková funkce, která je nenulové pouze v malé části signálu. Analyzovaný signál je následně vynásoben vzniklou okénkovou funkcí a díky tomu vzniká malá nenulová část signálu dle okénkové funkce viz obr. 1.5. Chceme-li analyzovat signál v různých časech je tato funkce po signálu posouvána a následně se počítá DFT pro každý výsledný okénkový signál.



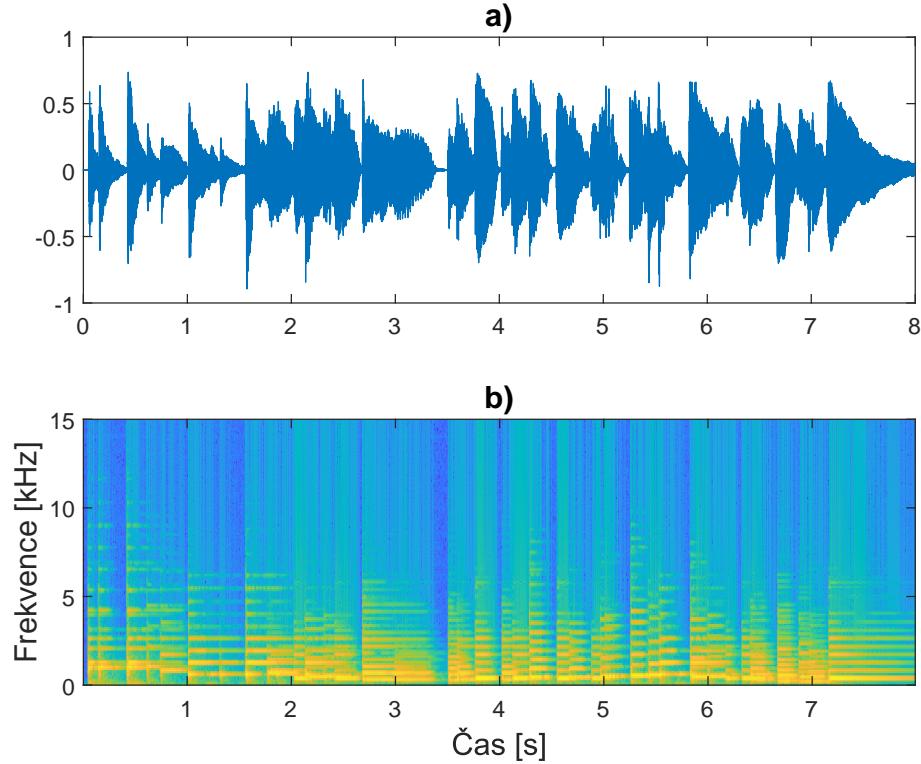
Obr. 1.5: Signál o délce 1s s počáteční frekvencí 10Hz a koncovou frekvencí 30Hz
a) Původní signál **b)** Signál s okénkem od 0,2s do 0,5s **c)** Signál s okénkem od 0,35s do 0,65s **d)** Signál s okénkem od 0,5s do 0,8s [16]

Na obr. 1.5 je graficky znázorněna myšlenka STFT, která ukazuje princip určování frekvenčních složek v čase a jejich výhody. Signál je násoben obdélníkovou okénkovou funkcí ve třech místech. Tyto tři vzniklé signály jsou následně na sebe nezávazně transformovány do frekvenční oblasti. Z výsledků Fourierovy transformace lze vidět, že každá z těchto částí má jiné frekvenční spektrum. Pokud by bylo zapotřebí například určit přesný přechod mezi dvěma frekvencemi nacházejícími se v signálu. Lze zpřesnit časové měřítko analýzy pomocí délky okénka. Tím ale dochází ke zmenšení přesnosti ve frekvenční oblasti.

Na výsledku přesnosti analýzy pomocí STFT závisí také tvar použité okénkové funkce. V obr. 1.5 je použito obdélníkového okénka které díky svým ostrým hranám zkresluje výsledek o nechtěné frekvenční složky. Existuje více tvarů okénkových funkcí pro odstranění nežádoucích složek. Například to jsou Kaise, Chebyshev, Hann, Haming a další [3].

Pokud jsou analyzované data skrze okénka funkce STFT poskládány zpět za

sebe, tak představují matici průběhu frekvenčního spektra signálu v čase. Takové zobrazení se nazývá **spektrogram** a v případě 2D zobrazení se skládá z časové a frekvenční osy. Magnituda frekvencí je pak zobrazena barevou škálou. Viz obrázek 1.6.



Obr. 1.6: Nahrávka piana a) Amplituda nahrávky b) Frekvenční spektrum nahrávky zobrazené pomocí spektrogramu

1.2.6 Dynamika hlasitost a intenzita

V češtině se pojem hlasitost využívá pro reprezentaci subjektivního vnímání akustického tlaku definovanou například jednotkou phon⁵. Stejně tak je hlasitost využívána, hovoří li se o měřené hlasitosti vyjádřené například hladinou intenzity zvuku popsanou níže nebo efektivní hodnotou signálu. Z důvodu lepší srozumitelnosti jsou dále využívána anglické pojmy „volume“ a „loudness“.

[50 150 250 350 450 570 700 840 1000 1170 1370 1600 1850 2150 2500 2900 3400
4000 4800 5800 7000 8500 10500 13500]

Dynamika popisuje průběh hlasitosti „volume“ interpretovaného hudebního díla. Udává jeden z faktorů jak lze například odlišit stejnou skladbu zahranou různými muzikanty. Interpretací skladby umělec vytváří dynamiku přednášeného díla [16]. V

⁵Phon - logaritmická jednotka vyjadřující individuální vnímání hlasitosti. Vnímání hlasitosti lidského ucha je závislé na křivce prahu slyšení a může se lišit pro každý tón [25].

notovém zápisu je dynamika neboli hlasitost přednesu popsána symboly jako jsou například pianissimo „*pp*“, piano „*p*“, forte „*f*“ a další.

V audio signálu je dynamika brána jako hlasitost „loudness“. Jedná se o změny amplitudy signálu nebo jeho efektivní hodnoty RMS⁶ v čase.

Při měření hlasitosti „loudness“ v akustickém prostoru je pak využíváno pojmu **intenzita** zvuku a **akustický výkon**. Akustický výkon je definován jako množství energie vyzářené akustickým vysílačem ve vzduchu za jednotku času.[7]. Jednotkou je W . A intenzita zvuku je pak definována jako množství energie, které projde jednotkovou plochou kolmou na směr šíření na jednotku času. Jednotkou je Wm^{-2} [27]

Z pohledu vnímání hlasitosti lidským uchem je rozsah vnímané intenzity zvuku v řádech bilionů. Práh slyšení činí $10^{-12} Wm^{-2}$ a práh bolesti je $10 Wm^{-2}$. Pro zmenšení tak velkého řádu je definována hladina intenzity zvuku v decibelech dB . Kde vztažnou hodnotou je práh slyšení $I_0 = 10^{-12} Wm^{-2}$. Hladina intenzity se vypočítá dle rovnice 1.10.

$$L_I = 10 \log\left(\frac{I}{I_0}\right) \quad (1.10)$$

1.2.7 Barva

V hudebním vyjádření se za slovem barva zkrývá komplexní sdružení atributů. Jedná se jak o psychologický tak hudební problém, který je vnímán individuálně[13].

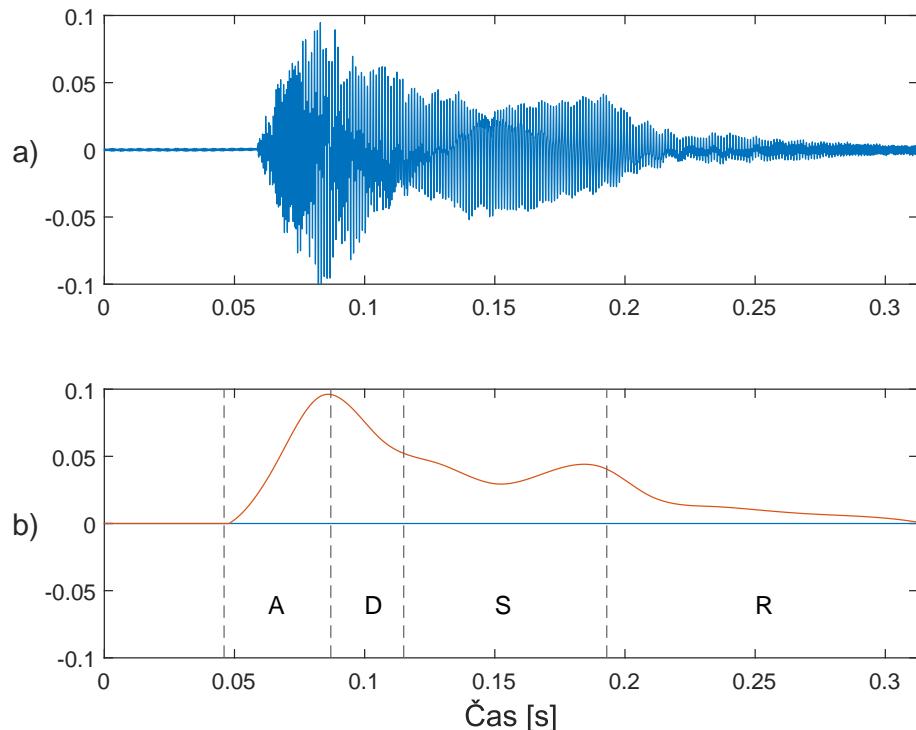
Zjednodušeně se barva definuje jako vlastnosti, díky kterým je možné rozpozнат tón o stejné výšce a hlasitosti zahrnutý na dva různé nástroje[17]. Díky barvě je posluchač schopen rozpozнат odlišné zvuky nástrojů a typů interpretace.

Z důvodu špatné kategorizace barvy fyzikálními veličinami je většinou interpretována přídavnými jmény. Například je barva popisována jako jasná, temná, ostrá, čistá, teplá, pestrá, nazální, tonální a další.

Jedním z možných nástrojů pro analýzu barvy tónu je tzv. obálka tónu „signálu“. Obálku určuje amplituda signálu v čase viz obr. 1.7 Je rozdělena na 4 fáze popsané z knihy Fundamentals of Music Processing [16]. **Attack „náběh“** určující začátek tónu. Například úder paličkou na blánu bubnu. V této fázi se nachází více ruchových složek z daného úderu a má velkou dynamiku. Následuje fáze s názvem **Decay „útlum“**. Po hlasitém úderu amplituda signálu klesá a začíná převládat tonální složka. Decay udává dobu za kterou se signál z jeho maxima sníží na hodnotu sustain. **Sustain „podržení“** je fáze ve které je zřetelný tón a stálá hlasitost. Rezonující blána bubnu. Poslední fází je **Release „uvolnění“** při kterém dochází k poklesu

⁶RMS - udává statistickou hodnotu z měření velikosti veličin. Je využívána u periodických veličin[2].

hlasitosti zdroje zvuku až k uplnému utlumení. Například přiložení tlumítka na rezonující strunu.



Obr. 1.7: Tón A5 zahráný na klavír a) Amplituda tónu b) ADSR obálka tónu

Další informace o barvě signálu se nacházejí v jeho frekvenčním spektru. Tón zahráný na hudební nástroj má svou fundamentální „nosnou“ frekvenci nazývanou první harmonická udávající jeho výšku. Dle konstrukce nástroje se v signálu objevují násobky nosné frekvence. Tyto násobky představují vyšší harmonické složky tónu. Počet a amplituda vyšších harmonických složek má vliv na výslednou barvu tónu a je to hlavní důvod proč je lidské ucho schopné rozeznat stejný tón znějící na různé nástroje.

1.3 Detekce nástupů a analýza tempa skladby

S vývojem nových technologií se mění i přístupy využívané v MIR pro analýzu tempa skladby. V této kapitole je popsán postupný vývoj algoritmů. Od nejjednodušších přístupů po komplexní řešení využívající moderní metody hlubokých neuronových sítí.

Detekce nástupů lze chápat jako detekci začátků not či dalších hudebních událostí, které se vyskytnou v průběhu skladby. Výskyt takových událostí je provázen zvýšením energie signálu zaznamenané ve fázi nástupu dle ADSR obálky popsané v

bodě 1.2.7. Typickým znakem pro fázi nástupu je rychlý nárůst obálky amplitudy signálu. V této fázi se při vytváření tónu vyskytují **Tranzienty**. Tranzienty je možné pozorovat zejména u neperkusivních nástrojů například u dechových nebo smyčcových. Jsou představovány nakmitávajícími a dokmitávajícími pochody. Jedná se o zvuk netónové podoby připomínající hluk s výraznými frekvenčními změnami. U smyčcových nástrojů takový zvuk může být ze začátku způsoben drhnutím smyčce o strunu do chvíle než se ustálí její kmitání. Délka tranzientů pak může dosahovat od jednotek milisekund až po stovky milisekund v závislosti na typu nástroje a technice hry [24]. Například v případě piana tranzient odpovídá počáteční fázi kdy byla zmáčknuta klávesa. Na základě zmáčknutí klávesy dochází ke zvednutí tlumítka, kladívko udeří do struny, struna začíná vibrovat a vibrace se přenáší do těla piana. V této fázi začíná tělo rezonovat a konečně dochází k ustálení tónové složky. Při úderu kladívka dochází k velkému přenosu energie patrném na obálce tónu. Díky tomu je lehké určit začátek tónu podle nárůstu amplitudy obálky [16].

U některých nástrojů je však energie přenášená po celou dobu znění tónu konstantně a dochází k pomalému jemnému náběhu tónu například u určitých technik hry na smyčcové nástroje nebo u dechových nástrojů. Fáze náběhu je pak pomalá a dlouhá a plynule přechází do fáze podržení. Pro tyto jemné zvuky se stává obtížné určit skutečnou pozici začátku noty.

Náročnost detekce nástupů se zvyšuje v případě, že nehraje pouze jeden nástroj. Jedná se například o polyfoní skladbu. Polyfoní skladbou rozumíme skladbu tvořenou více hlasy. Zároveň nemají určenou roli vedoucího a doprovodného hlasu [18]. Takové uspořádání hlasů může vést k překrývání a nástupy mohou být maskovány. Díky jevu maskování je obtížné zaznamenat změny energie signálu. V tomto případě přichází potřeba po komplexnějších metodách detekce nástupů [16]. Například pohled na krátkodobé změny ve spektru signálu pomocí využití STFT popsané v bodě 1.2.5.

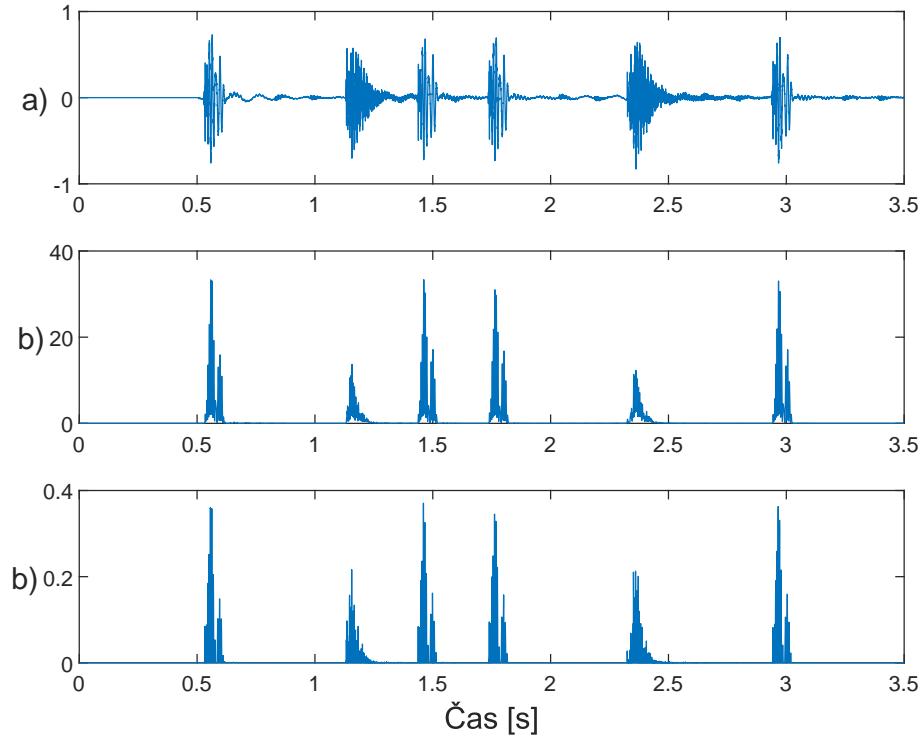
1.3.1 Využití energie signálu

Jak již bylo zmíněno při hraní dochází k přenosu energie od hřáče na hudební nástroj. Tento přenos je ve značné míře provázen rychlým nárůstem amplitudy ve fázi náběhu na začátku tónu. Například při úderu kladívka piana na strunu nebo úderu paličky na blánu bubnu. Ná základě tohoto jevu je možnost detektovat nástup tónu pomocí funkce pro výpočet energie signálu v daném místě. Náhlé změny signálu v takto definované funkci ukazují potencionální místa nástupů. Matematicky pak funkci popisujeme. Stejně jako u STFT popsané v bodě 1.2.5 je definována okénková funkce $w(n)$ ve tvaru zvonu „bell-shaped function“ která je posouvaná po diskrétním signálu $x(n)$. Okénková funkce je symetrická podle počátku a potom platí že $m \in [-M : M]$

a $M \in \mathbb{N}$. Funkce lokální energie signálu je pak zapsána

$$E_{wx}(n) = \sum_{m=-M}^M |s(n+m)w(m)|^2 \quad (1.11)$$

pro $n \in \mathbb{Z}$ [16].



Obr. 1.8: Detkete nástupů perkusivního zvuku **a)** Amplituda nahrávky **b)** Lokální energie signálu $E_{wx}(n)$ **c)** Derivace lokální energie signálu s půlvlnným usměrněním $\Delta_E(n)$

Pro názornější zobrazení se následně vypočítá derivace funkce lokální energie. V případě diskrétního signálu se derivace realizuje jako rozdíl dvou po sobě jdoucích vzorků. Protože pro detekci nástupů je důležitá zejména pozitivní změna energie nikoliv její pokles jsou ponechány pouze pozitivní rozdíly a negativní jsou zapsány jako nulové. V anglické literatuře je tento proces známý také jako půlvlnné usměrnění „half-wave rectification“. Vzorce jsou zapsány následovně

$$r = E_{wx}(n+1) - E_{wx}(n) \quad (1.12)$$

$$\Delta_E(n) = \frac{r + |r|}{2} = \begin{cases} r, & \text{if } r \geq 0 \\ 0, & \text{if } r < 0 \end{cases} \quad (1.13)$$

kde $r \in \mathbb{R}$ a $n \in \mathbb{Z}$. Na obr. 1.8 lze vidět zobrazení takto vypočítané lokální energie signálu a její derivace. Analyzovaný signál se skládá z několika perkusivních úderů. Pro výpočet bylo použito okno typu „bell-shaped function“ o velikosti 201 vzorků.

1.3.2 Využití spektra signálu

Díky rozložení signálu na jeho spektrum pomocí Fourierovy transformace popsané v bodě 1.2.3 je možné lépe rozeznat strukturu nahrávky a zmírnit efekt maskování který nastává při metodách založených na energii signálu popsaných v bodě 1.3.1. V polyfonní hudbě mohou interpretace o nižší hlasitosti být maskovány hlasitějšími projevy. Kdy například jeden nástroj v tónové fázi podržení ADSR obálky, popsané v bodě 1.2.7, může být hlasitější než fáza náběhu druhého nástroje. Takový nástup je pak maskován a je obtížná jeho detekce. V případě maskování některých z nástrojů v časové oblasti je možné spektrum signálu zaměřit na frekvenční oblast spektra ve které se maskovaný nástroj nachází. Díky tomu je možné nástup tónu takového nástroje detektovat pomocí spektra signálu. Každý hudební nástroj obsahuje jiné frekvenční spektrum [16]. Díky tomu se různé nástroje nachází na odlišných místech spektra, jak spektrum daného nástroje vypadá určuje také barvu popsanou v bodě 1.2.7. S takovým jevem je důležité počítat při detekci nástupů založené na spektru signálu.

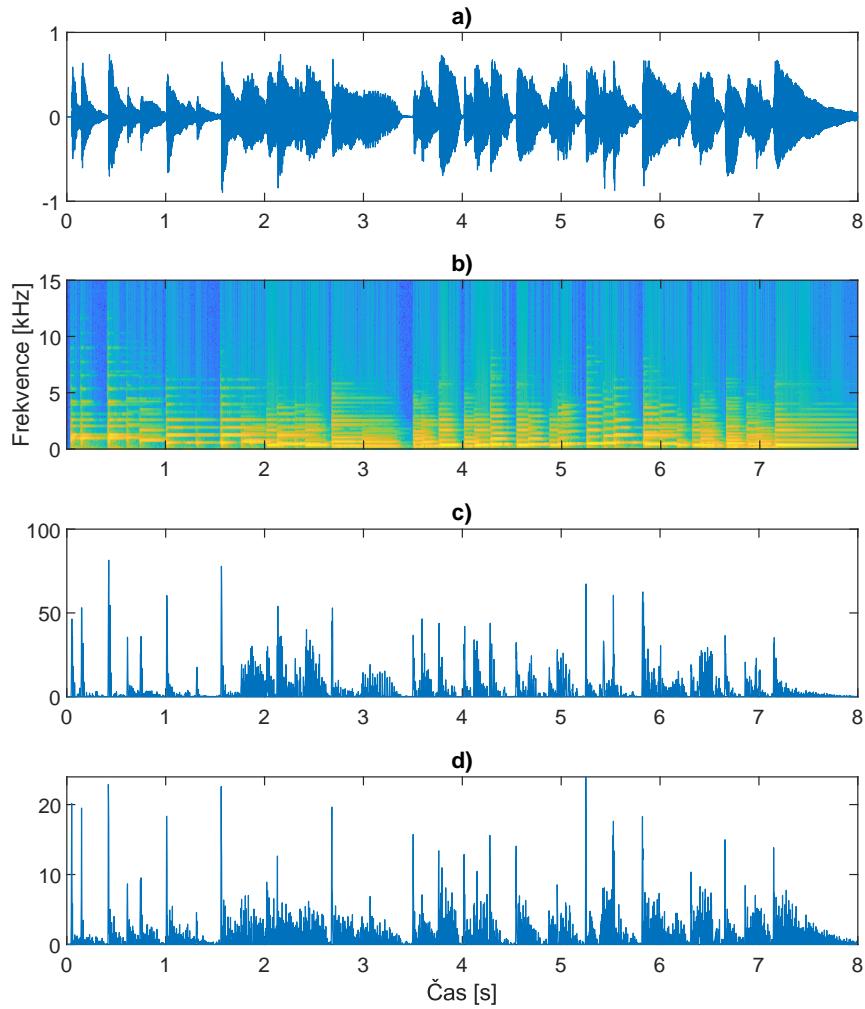
Jedním ze spůsobů analýzy nástupů pomocí spektra je detektovat změny ve spektru v průběhu času. Při zobrazení časového průběhu spektra nazývaného spektrogram popsaný v bodě 1.2.5. Jsou za sebou v čase poskládány vektory nesoucí informaci o spektru. Pokud jsou počítány rozdíly mezi dvěma po sobě jdoucimi vektory spektrogramu jsou získány informace o změně spektra v čase. Protože tranzienty se při fázi náběhu skládají z části z ruchové složky rozléhající se přes velkou část slyšitelného frekvenčního spektra je tak možné detektovat nástupy. Popsané metodě porovnávání vektorů se také říká **spektrální tok** [16]. Pro výpočet spektrálního toku existuje více druhů přístupů lišících se předszpracováním dat a konečným zpracováním výsledků. Níže jsou popsány dvě metody výpočtu ze spektrogramu a následně z mel spektrogramu.

Výpočet spektrálního toku se provede derivací vstupního diskrétního signálu který zde představuje vektory spektrálních složek. Výpočet probíhá pro každou spektrální složku a výsledek je sečten. Popsáno rovnici

$$r(n) = \sum_{k=0}^K |X|(n+1, k) - |X|(n, k) \quad (1.14)$$

$$\Delta S_t(n) = \frac{r(n) + |r(n)|}{2} = \begin{cases} r, & \text{if } r \geq 0 \\ 0, & \text{if } r < 0 \end{cases} \quad (1.15)$$

kde $n \in \mathbb{Z}$ a K je počet spektrálních složek v jednom vektoru.



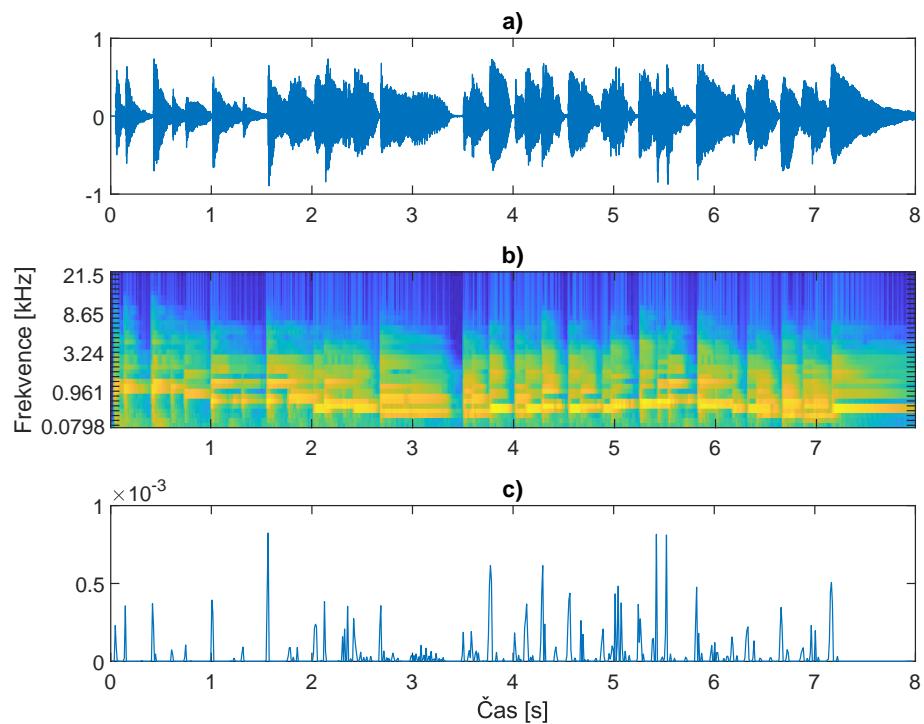
Obr. 1.9: Výpočet spektrálního toku pro nahrávku piana **a)** Amplituda nahrávky **b)** Spektrogram nahrávky **c)** Spektrální tok bez komprese **d)** Spektrální tok s kompresí spektra $\gamma = 1$

Pro dosažení přesnějších výsledků je možnost komprese spektrogramu. Funkce pro logaritmickou kompresi singálu je zapsána jako

$$X_c = \log(1 + \gamma|X|) \quad (1.16)$$

kde pro správnou kompresi je potřeba aby $\gamma \geq 1$. Míra komprese je určena velikostí γ . Nižší hodnoty jsou vhodné pro lepší detekci perkusivních zvuků. Při vyšších hodnotách γ dochází k větší komprezi spektrogramu a jsou zřetelnější zvuky o nižší intenzitě. Při velké komprezi dochází také k zvýraznění ruchových složek signálu [16].

Například v článku z ISMIR 2021 [12] je spektrální tok počítán s Mel Spectrogramu popsaného níže. Protože lidské ucho vnímá výšku frekvence logaritmicky je jednodužší rozeznat rozdíl mezi frekvencemi u nižší části spektra než u vyšších frekvencí. Například rozdíl mezi 500 Hz a 1 kHz je pro lidské ucho dobře detekovatelný a zřetelný. Odlišné rozlišovací schopnosti výšky v různých frekvenčních pásmech je způsobeno biologickým rozvržením vnitřního ucha. Proto pro lepší rozlišování na základě lidského vnímání zvuku vznikla Melova stupnice. **Melova stuopnice** je vjemová škála výšek, které posluchač posoudil jako od sebe stejně vzdálené. Referenčním bodem je pak definováno 1000 mel které odpovídají frekvenci 1000 Hz[22]. Výpočet spektrálního toku je následně stejný jak pří výpočtu ze spektrogramu. Výslednou křivku je možné vidět na obr. 1.10.



Obr. 1.10: Výpočet spektrálního toku z mel spektrogramu pro nahrávku piana **a)** Amplituda nahrávky **b)** Mel spektrogram nahrávky **c)** Spektrální tok

1.3.3 Detekce periodicity

Detekce periodicity je nezbytnou součástí procesu při detekci dob skladby. Důležitou technikou pro detekci periodicity je autokorelační funkce. Autokorelační funkce slouží pro detekci periodických vzorů v datech. Tato metoda dokáže detektovat tempo, ale neumožňuje určit o jakou dobu se jedná. Základní použití je výpočet autokorelační funkce s energie signálu z bodu 1.3.1. Matematicky je pak popsána dle rovnice 1.17. [8]

$$r(n, i) = \sum_{u=-(T_w/2)+1}^{T_w/2} E(n+u)E(n+u-i) \quad (1.17)$$

Při detekci periodicity je možné využít také banky rezonátorů hřebenového filtru s konstantním poločasem [19] nebo fázově blokující rezonátory [9].

Autokorelační funkce zde přináší několik výhod. Jednoduchost a efektivita. Je to jednoduchá metoda pro zjistění periodicity v signálech a vyžaduje menší množství paměti než jiné výše zmíněné metody. Autokorelace může odhalit ne zcela zřejmé periodické vzory, které nemusí být při poslechu nebo pozorování okamžitě zřejmé. Zároveň je robustní vůči šumu. Při analýze také nedochází ke ztrátě informace. [10]

Autokorelace má také své nevýhody. Není vhodná pro analýzu v reláném čase protože potřebuje dostatečné množství dat a je potřebné aby v analyzovaném datovém okně byla obsažena celá perioda. V opačném případě nedojdete k jejímu nalezení. Autokorelace není schopná pracovat s fází signálu. Pro rozpoznání fáze je zapotřebí dalších metod. Nekonstantní tempo v analyzovaném signálu představuje pro autokorelační analýzu problém.[10]

1.4 Klasifikace žánrů a nálady

1.5 Chromavektory

1.6 Dostupná řešení

Díky vědecké základně v oblasti music information retrieval vznikají open source knihovny umožňující snadné použití v programovacím jazyce python. Tyto knihovny jsou šířeny za účelem usnadnění práce v oblasti MIR. Cílem některých knihoven je usnadnit přechod výzkumných týmů do programového jazyka Python a zakomponovat moderní praktiky softwarového vývoje. Díky tomu se staly tyto metody dostupné širší komunitě vědců.[14] V této kapitole je popsáno několik rozšířených knihoven poskytujících metody z oblasti MIR.

1.6.1 Librosa

Knihovna Librosa vznikla v roce 2015 na základě potřeb vědecké komunity pro snadné použití metod z oblasti MIR v jazyce python. Publikován a popsána byla článkem Audio and Music Signal Analysis in Python [14] na SciPy 2015. Jedná se o otevřenou knihovnu jejíž další vývoj probíhá zapomoci vědecké komunity na GitHubu kde je kladen důraz na kompatibilitu, řádnou dokumentaci obsahující popisy

funkcí s příkladovými kódy, a testování funkčnosti. Knihovna obsahuje pokročilé metody pro detekci dob.

Librosa obsahuje funkce pro extrakci audio vlastností. Pro tuto práci jsou důležitými funkcemi zejména Detekce tempa a dob skladby, rozpoznání začátků tónů. Následně jsou v knihovně obsaženy nástroje pro vizualizaci získaných vlastností. Například zobrazení spektrograma či grtafické zobrazení detekovaného tempa a dob. Funkce pro detekci tempa a dob skladby pomocí dynamického programování je v knihovně zapsána:

```
tempo, beats = librosa.beat.beat_track(y=y, sr=sr)
```

Kde y je analyzovaný signál a sr je vzorkovací frekvence signálu. Detekce dob je ve funkci $beat_track()$ realizována pomocí dynamického programování. Proces je rozdělen na 3 etapy popsané Danielem P.W. Ellis ve článku Beat Tracking by Dynamic Programming [6]. Jedná se o tyto kroky:

1. Výpočet obálky síly nástupů
2. Globální odhad tempa
3. Porovnání obálky síly nástupů s globálním tempem

Výpočet obálky síly nástupů: Při výpočtu obálky síly nástupů je vstupní signál převzorkován na 8 kHz. Poté je vypočítána STFT s délkou okna 32 ms a 4 ms mezerou mezi okny. Vypočítaný spektrogram je převeden na mel spektrogram o 40 mel pásmech. Mel stupnice je využívána ve snaze vyrovnat důležitost každého frekvenčního pásma v návaznosti na logaritmické vnímání frekvencí lidským uchem. Mel spektrogram je převeden na dB a následně jsou počítány diferenciální rovnice prvního řádu podle času pro každé mel pásmo. Principem půlvlnného usměrnění jsou výsledné negativní hotnoty změněny na nulové. Hodnot kladné jsou v daném čase sečteny napříč všemi pásmi. Výsledný signál je filtrován horní propustí s mezní frekvencí 0,4 Hz a vyhlazen pomocí konvoluce s Gaussovy obálky s šírkou okolo 20 ms. Pomocí výše popsaného procesu je získána jednorozměrná obálka síly nástupů v závislosti na čase reprezentující proporcionální nárůst energie přibližně sečtené ve mel pásmech.

Globální odhad tempa: Globální odhad tempa je určen z obálky síly signálu $O(t)$ získané v předchozím kroku. Je počítána autokorelace mezi původní obálkou signálu $O(t)$ a jejími zpožděnými verzemi. Pro zpoždění ve kterých se potká velké množství vrcholů obálky nastává velká korelace. Tato korelace může nastat v celočíselných násobcích dané hodnoty zpoždění. Díky tomuto jevu je těžké určit který čas zpozdění představuje správné tempo skladby. Avšak lidmi vnímané tempo skladeb má sklon být kolem 120 BPM jak bylo zjištěno ve výzkumu popsaném v článku Ambiguity in Tempo Perception: What Draws Listeners to Different Metrical Levels? [15]. Proto je na výslednou autokorelacii

aplikováno váhové okno snižující hodnotu autokorelace se vzdáleností od 120 BPM. Čas ve kterém autokorelace dosahuje největší hodnoty je čas hledaného tempa skladby.

Porovnání obálky síly nástupů s globálním tempem: Matematicky je tento krok formulován rovnicí 1.18 kde $\{t_i\}$ je pole nalezených dob, $O(t)$ je obálka síly signálu, α je váhový koeficient určující důležitost mezi obálkou síly signálu a detekovaným tempem. Funkce $F(\Delta t, \tau_p)$ zapsaná rovnicí 1.19 měří konzistenci mezi skutečným časem od sebe vzdálených dob Δt a ideálním časem mezi dobami τ_p určeným cílovým tempem.

$$C(\{t_i\}) = \sum_{i=1}^N O(t_i) + \alpha \sum_{i=2}^N F(t_i - t_{i-1}, \tau_p) \quad (1.18)$$

$$F(\Delta t, \tau) = -(\log \frac{\Delta t}{\tau})^2 \quad (1.19)$$

1.6.2 Madmom

1.6.3 Aubio

1.6.4 Hodnocení extrakce informací

1.7 Systém Spectoda

Systém spectoda využívá pro generování animací několik základní bloků animací. Tyto bloky lze spolu skládat, kombinovat, odečítat a sčítat mezi sebou. Díky tomu je možné následně vytvářet komplexní animace.

1.8 Hudební signál jako animace

2 Výsledky studentské práce

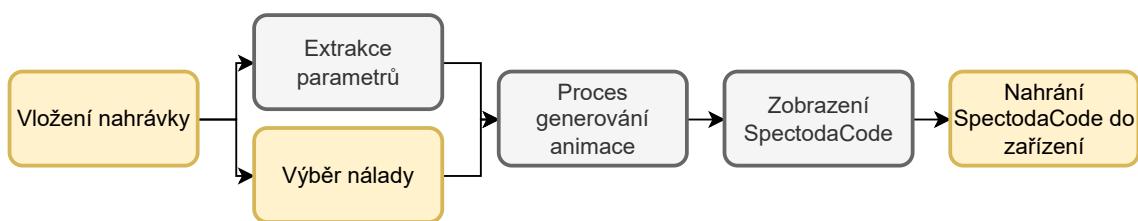
2.1 Návrh výsledného systému

Návrh popisuje komplexní systém skládající se z několika částí, uživatelské rozhraní algoritmů pro získání pamrametrů hudební nahrávky a algoritmu generujícího SpectodaCode na základě získaných parametrů. V této kapitole je podrobně popsán návrh jednotlivých částí systému.

2.1.1 Uživatelské rozhraní

Uživatelské rozhraní je reprezentováno webovou stránkou a je naprogramováno pomocí značkovacího jazyka HTML spolu s formátováním v jazyce CSS. Funkčnost webové stránky je zajištěna funkcemi jazyce JavaScript. Javascript také vytváří propojovací můstek pro komunikaci s vnějším systémem v jazyce Python.

Jedná se o jednoduché webové rozhraní ve kterém uživatel nahraje hudební skladbu ve formátu .wav. Rozhranní obsahuje pole pro vložení cesty k hudební skladbě umožňující výběr ze souborů v uživatelské uložišti. Níže je posuvník s 4 základními hodnotami pro výběr nálady „mood“. Jedná se hodnoty „chill“, „hang out“, „feeling happy“ a „dancing“, které jsou v tomto pořadí na posuvníku. Uživatel může pomocí posouvání posuvníku vybrat pro jakou náladu chce vytvořit animaci. Pod posuvníkem pro výběr nálady se nachází tlačítko pro spuštění procesu generování SpectodaCode. Poslední částí webového rozhraní je textové pole ve kterém se zobrazí vygenerovaný SpectodaCode.



Obr. 2.1: Blokové schéma postupu uživatele webovou stránkou

Na blokovém schématu 2.1 je zobrazen proces postupu uživatele skrze webové rozhraní.

2.1.2 Parametry hudební nahrávky

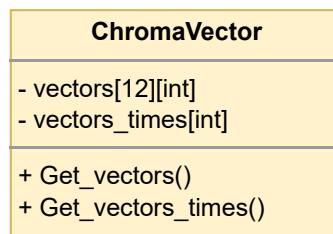
Systém popsaný v bodě 2.1.3 vyžaduje vstupní data o hudební nahrávce. Tyto data jsou rozdělena 7 odlišných objektů. Každý z těchto objektů představuje určitou

vlastnost analyzované nahrávky. Tyto vlastnosti jsou získány pomocí technik popsaných v bodě 2.2. Jednotlivé vlastnosti a jejich datové struktury jsou shrnutý v následujících bodech.

Detekce dob představuje pole hodnot jehož délka je závislá na době trvání nahrávky. Jednotlivé hodnoty pak udávají časy hudební nahrávky, ve kterých se nacházejí doby.

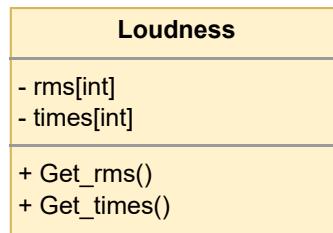
Tempo skladby je číslo typu float s jednotkou BPM vyjadřující počet úderů za minutu. Hodnota BPM se vztahuje k počtu čtvrtových not za minutu. Vybraný algoritmus s knihovny Librosa ovšem nedetectuje jestli se jedná o noty čtvrtové. Postup zjištění tempa je popsán v bodě 1.6.1.

Chromavektory jsou získány v podobě pole jehož počet řádků udává 12 půltónů rozdělujících 1 oktávu. Délka pole je závislá na délce nahrávky a velikosti okna při výpočtu STFT. K matici chromavektorů je přidáno pole o stejné délce. Hodnoty v poli udávají čas konce okna, ve kterém jsou počítány chroma vlastnosti. Tyto dvě proměnné jsou zadány jako parametry třídy s názvem *ChromaVector* jehož struktura je zobrazena v blokovém schématu 2.2.



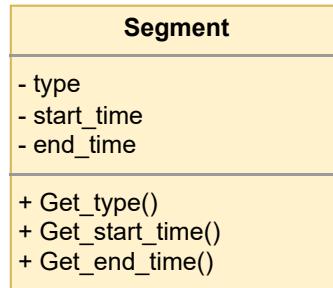
Obr. 2.2: Struktura třídy *ChromaVector*

Efektivní hodnota signálu je zapsána třídou *Loudness* obsahující dva atributy. Prvním z nich je pole *rms* jehož délka je závislá na délce signálu a obsahuje efektivní hodnoty signálu v časech uložených v druhém atributu. Druhý atribut *times* je pole o stejně délce jako pole s hodnoty *rms* a obsahuje časy skladby ve kterých je hodnota *rms* počítána.



Obr. 2.3: Struktura třídy *Loudness*

Segmentace je zapsána jako pole objektů třídy *Segment*. Tato třída obshahuje 3 atributy *type*, *start_time* a *end_time*. Argument *type* obsahuje statickou hodnotu označující o jakou část skladby se jedná. Tyto hodnoty jsou vypsány v úryvku kódu 2.1. Argumenty *start_time* a *end_time* označují začátek a konec segmentu v nahrávce.



Obr. 2.4: Struktura třídy *Segment*

```

# Song segments variables
SILENT      = 0
REFRAIN     = 1
STROPHE     = 2
BRIDGE      = 3
  
```

Výpis 2.1: Hodnoty proměnné *type*

Žánr představuje proměnou *genre* typu short vekteré je zapsáno číslo označující žánr skladby. Statické hodnoty žánrů jsou zapsány v úryvku kódu 2.2.

```

# Song genre variables
CLASSIC     = 0
FOLK        = 1
POP          = 2
ROCK         = 3
METAL        = 4
ELECTRONIC   = 5
  
```

Výpis 2.2: Hodnoty proměnné *genre*

Nálada představuje proměnou *mood* typu float s přednastavenými statickými hodnotami zobrazenými v části kódu 2.3. Nálada může být i v rozmezí přednastavencích hodnot. V tomto případě je výsledná hodnot lineráně závislá na vzdálenosti od přednastavených hodnot.

```

# Song mood variables
CHILL       = 0
  
```

HANG_OUT	= 1
HAPPY	= 2
DANCING	= 3

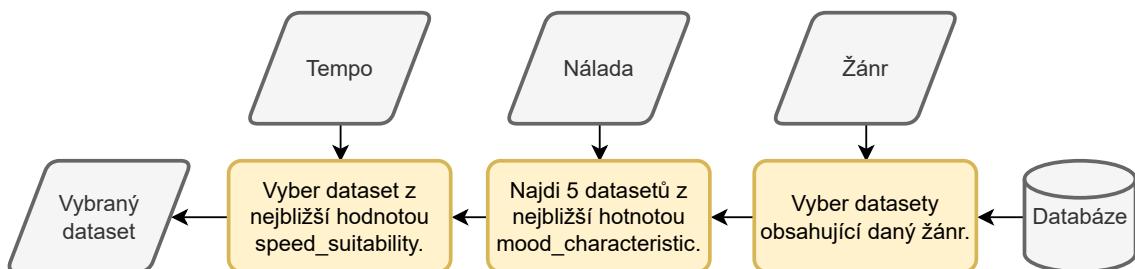
Výpis 2.3: Hodnoty proměnné *mood*

2.1.3 Systém pro generování animací

Systém pro generování animací představuje nejdůležitější část práce. Jeho struktura udává vizuální kvalitu animací a schopnost přizpůsobit se daným skladbám různých žánrů. V této kapitole je popsána základní struktura systému.

První částí systému je vstupní rozhraní, ve kterém jsou přijímány data obsahující parametry o hudební nahrávce. Struktura přijímaných dat je popsána v bodě 2.1.2. Každý ze zmíněných parametrů plní důležitou funkci v rozhodovacím procesu skládání bloků animace. Níže jsou popsány rozhodovací funkce pro jednotlivé parametry.

Žánr a nálada jsou používány pro výběr vhodného balíčku animací. Tyto balíčky jsou nazývány datasety a jejich datová struktura je popsána v bodě 2.1.4. Žánr i nálada jsou zaznamenány jako předdefinovaná celočíslena hodnota. Postup výběru datasetu je následující. Každý dataset obsahuje seznam žánrů pro který je vhodný. Na základě proměnné *genre* dochází k výběru všech datasetů vhodných pro daný žánr. Následně na základě proměnné *mood* ve které je uložená nálada je vabráno 5 datasetů s nejbližší hodnotou *mood_characteristic*.



Obr. 2.5: Blokový diagram výběru datasetu.

Tempo skladby je posledním parametrem při výběru vhodného datasetu. Hodnota proměnné *speed_suitability* každého z 5 vybraných datasetů je porovnána s tempem skladby. Finální dataset je vybrán ten jehož hodnota *speed_suitability* je nejbližší hodnotě tempa skladby. Blokový diagram procesu je znázorněn na obrázku 2.5. Tempo skladby je také použito pro výpočet délky jednotlivých animací na které je přímo úměrná rychlosť animace. Tento postup je více popsán níže.

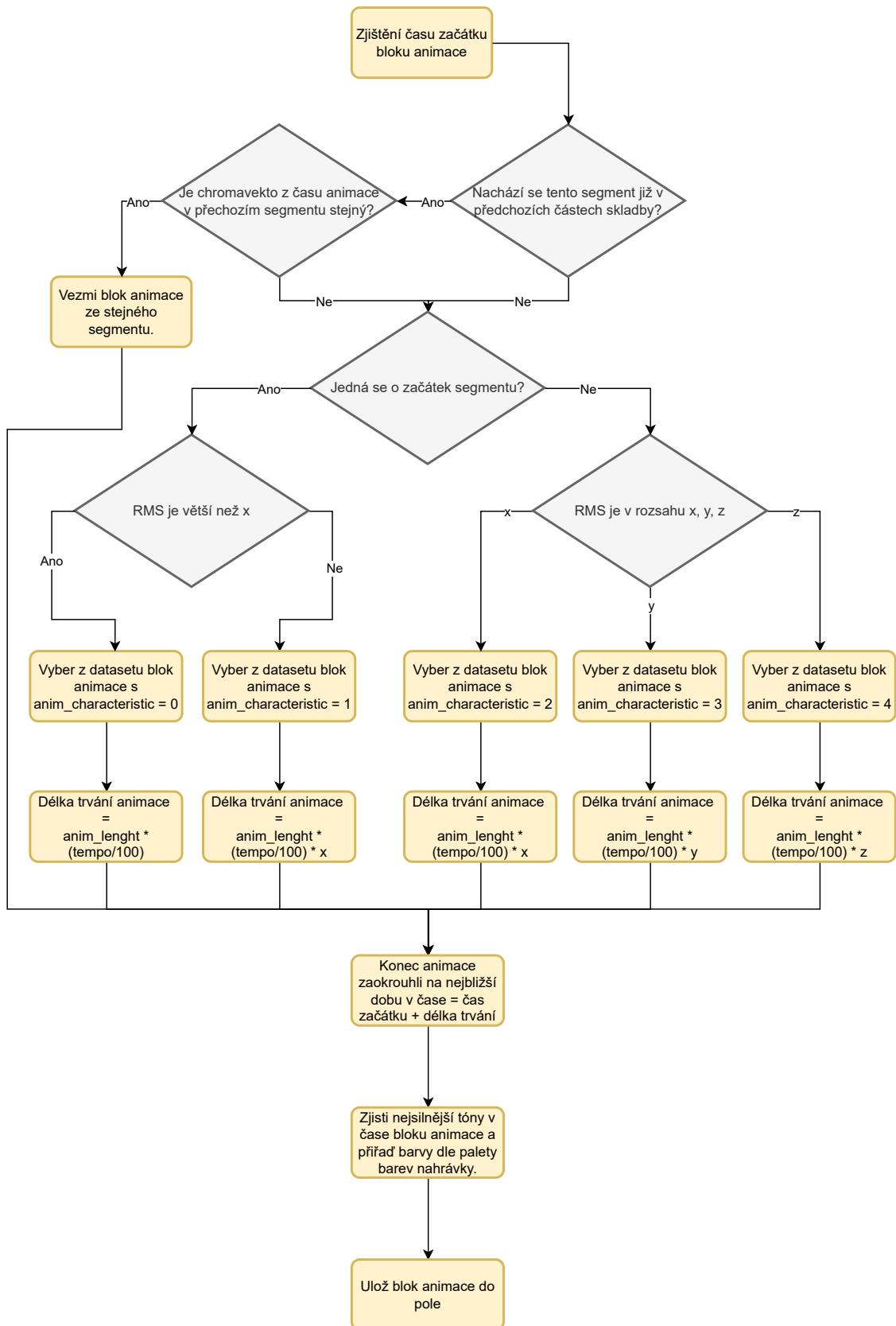
Segmentace slouží pro detekci opakujících se částí nahrávky. Například nahrávka obsahuje více než jednu sloku či refrémů. Je kladen důraz aby v opakujících se segmentech nahrávky byla animace stejného typu. Z toho plyne, že v každém refrému bude podobná animace.

Detekce dob je základní parametrem na základě kterého se nastaví začátky a konce animací tak, aby odpovídaly rytmu nahrávky.

Chroma vektory udávají tónovou strukturu skladby v průběhu času. Tento parametr je využit pro nastavení barevné škály animací. Chromavektory se analyzují a jsou vybrány 4 hlavní tóny nejčastěji se vyskytující v nahrávce. Keždému tónu je přiřazen barevný odstín. Poté je animaci v dané části skladby přiřazena barva dle tónu určeném chromavektorem daného času nahrávky.

Efektivní hodnota signálu pomáha procesu segmentace. Refrém a sloka se mohou lyšit hlasitostí která je způsobena efektivní hodnotou. Zároveň je použita pro výběr vhodného typu animace. Na základě velikosti efektivní hodnoty signálu v čase je vybrán typ animace. Vyšší hodnoty představují rychlejší a údernější animace. Nižší hodnoty naopak pomalejší a klidnější animace. Tento proces je zajištěn porovnáváním proměněných *anim_characteristic* obsažené v každém bloku animace s efektivní hodnotou mapovanou na určený rozsah.

Druhá část systému tvoří samotnou logiku skládání blkoků animací. Tato rozhodovací struktura je zobrazena na blokovém diagramu 2.6.

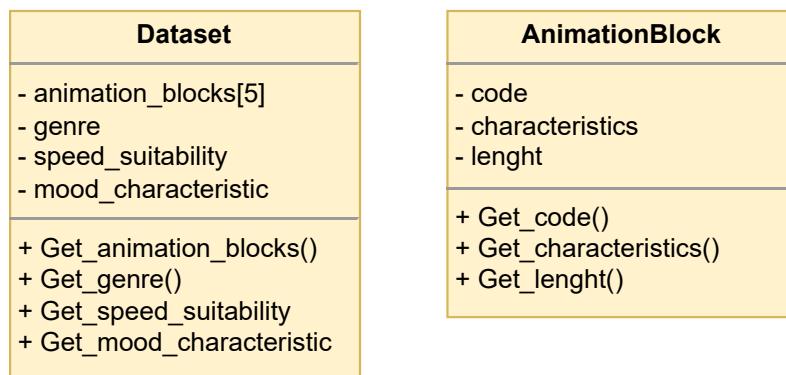


Obr. 2.6: Blokový diagram struktury rozhodovacího procesu

2.1.4 Databáze bloků animací

Jak je popsáno v bodě 1.7... jsou základní animace které jsou skládány dosebe a tak je možné vytvořit komplexní animace. Systém využívá již předsložených komplexních animací do bloků animací. Těmto blokům jsou přiřazeny parametry *code*, *characteristic* a *length* a jsou uloženy jako objekt datové třídy *AnimationBlock*.

Z objektů tipu *AnimationBlock* jsou dále tvořeny datasety zapsané datovou třídou *Dataset*. Tyto datasety obsahují 5 bloků animací a parametry *genre*, *speed_suitability* a *mood_characteristic*. UML diagramy těchto tříd jsou zobrazeny na obrázku 2.7.



Obr. 2.7: Blokové diagramy tříd *Dataset*, *AnimationBlock*

2.2 Výběr vhodných metod pro extrakci vlastností z hudební nahrávky

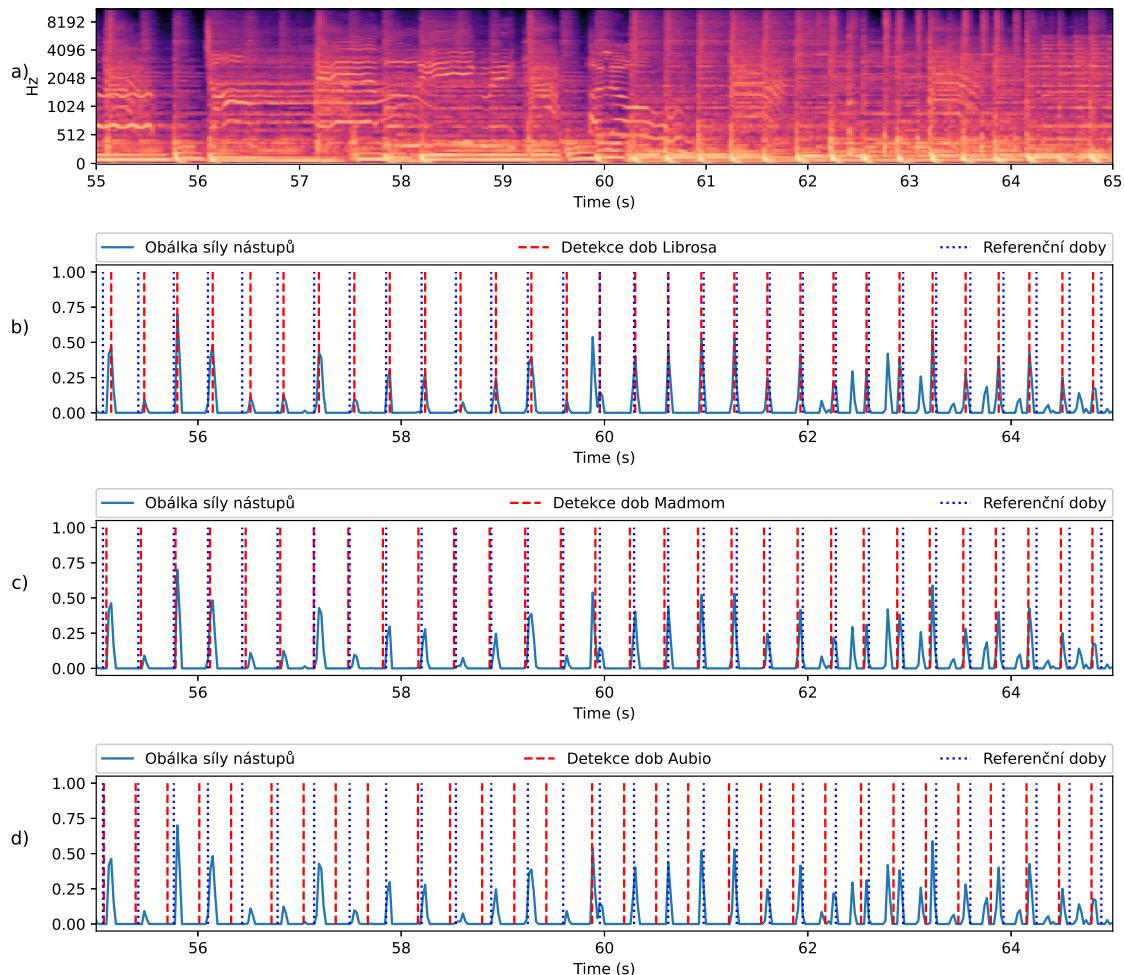
Vědecká komunita vytvořila volně šířících knihovny obsahující techniky z oborů MIR. V této části práce jsou prozkoumány 3 knihovny zmíněné v bodě 1.6. Jsou použity jejich funkce pro získání parametrů z hudební nahrávky potřebných pro navazující diplomovou práci. Tyto funkce jsou mezi sebou porovnány z hlediska přesnosti výsledků, rychlosti výpočtů, jednoduchosti použití a možnosti využití pro komerční účely.

2.2.1 Detekce dob a tempa

Pro porování detekce dob jsou vybrány 3 funkce. Pro každou knihovnu jedna funkce. První je z knihovny Librosa funkce *brat_track*. Z knihovny Madmom je použit *BeatTrackingProcessor* a z knihovny aubio funkce *tempo*.

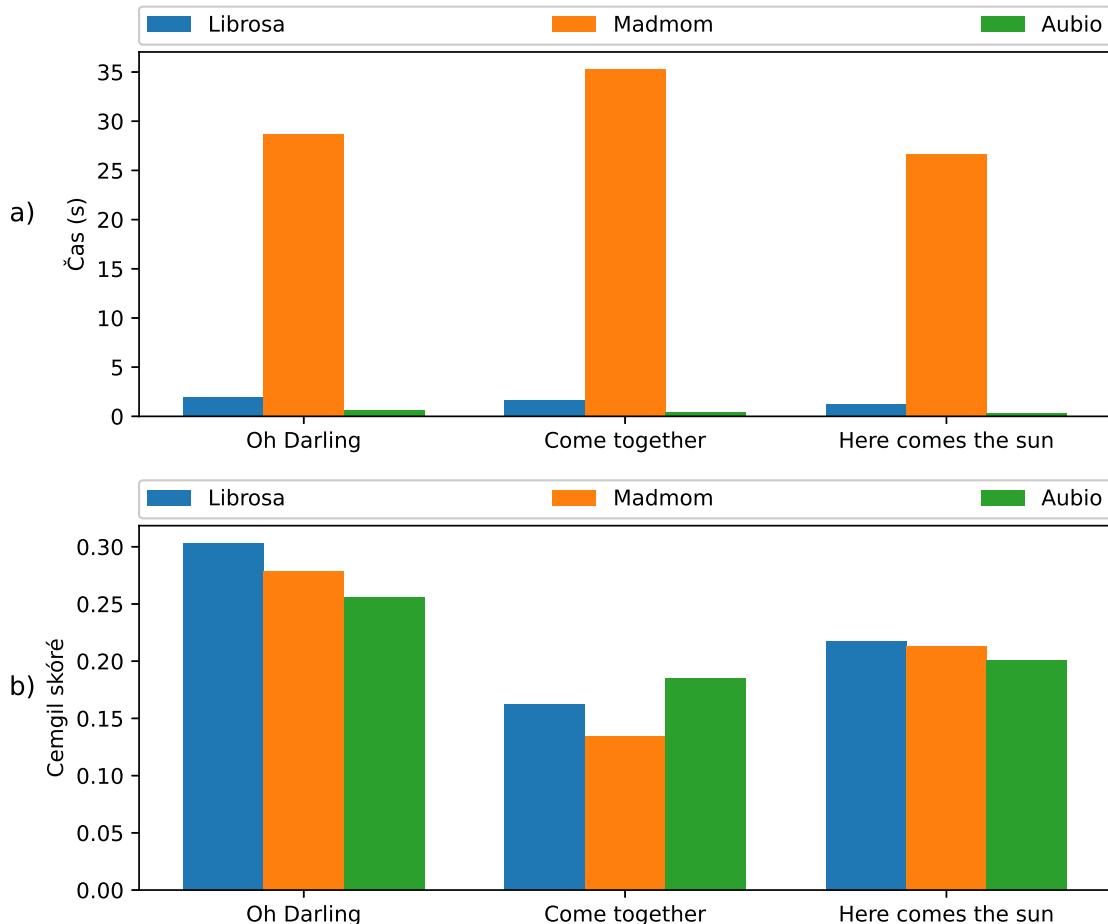
Funkce jsou porovnány na třech skladbách skupiny Beatles. Níže v grafu 2.8 Je vidět úryvek skladby Oh-Darling!. Pro lepší zobrazení jsou osy grafu v rozsahu 55 - 65 s nahrávky. Na prvním z grafu je vidět melspekrogram vypočítán pomocí

funkce *melspectrogram* z knihovny Librosa. Grafy b) - d) zobrazují v pozadí obálku síly nástupů a vertikální pruhované čáry znázorňují detekované doby dané funkce a vertikální modré tečkované čáry jsou referenční doby zaznamenané institucí Centre for digital music na univerzitě Queen Mary, University of London. tato databáze je dostupná online na Isophonics.net.



Obr. 2.8: Porovnání metod detekce dob na úryvku skladby Oh-Darling!. a) Mel-spekrogram b) Detekce dob pomocí Librosa c) Detekce dob pomocí Madmom d) Detekce dob pomocí Aubio

Z grafu lze vidět, že funkce z knihoven Librosa a Madmon se nejvíce blíží referenčním dobám. Pro přesné hodnocení funkcí je využita knihovna Mir_eval poskytující funkce pro hodnocení přesnosti detekce dob. Pomocí této knihovny je počítáno Camgil skóre. Popis knihovny a výpočtu je zmíněn v bodě 1.6.4. Posledním hodnoceným parametrem je doba výpočtu. Výsledky Camgil skóre, a doby výpočtu pro jednotlivé skladby jsou zobrazeny na obrázku 2.9.



Obr. 2.9: Porovnání přesnosti a času metód detekce dob na skladbách Oh-Darling!, Come Together a Here Comes The Sun. **a)** Čas výpočtu **b)** Cemgil skóré

2.2.2 Analýza chromavektorů

2.2.3 Efektivní hodnota signálu

Vstupními parametry systému jsou získané parametry jež jsou popsány v bodě 2.1.2

Závěr

Shrnutí studentské práce.

Literatura

- [1] Bracewell, R.: *The Fourier Transform and its Applications*. Tokyo: McGraw-Hill Kogakusha, Ltd., druh vyd n , 1978.
- [2] Cartwright, K.: Determining the effective or RMS voltage or various waveforms without calculus. ro n k 8, 01 2007.
- [3] Cohen, L.: Time-frequency distributions-a review. *Proceedings of the IEEE*, ro n k 77, . 7, 1989: s. 941–981, doi:10.1109/5.30749.
- [4] Crocker, M.: *Handbook of Acoustics*. A Wiley-Interscience Publication, Wiley, 1998, ISBN 9780471252931.
URL https://books.google.cz/books?id=1x_RvffW-hcC
- [5] Downie, J. S.; Ehmann, A. F.; Bay, M.; aj.: *The Music Information Retrieval Evaluation eXchange: Some Observations and Insights*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ISBN 978-3-642-11674-2, s. 93–115, doi: 10.1007/978-3-642-11674-2_5.
URL https://doi.org/10.1007/978-3-642-11674-2_5
- [6] Ellis, D.: Beat Tracking by Dynamic Programming. *Journal of New Music Research*, ro n k 36, 03 2007: s. 51–60, doi:10.1080/09298210701653344.
- [7] Acoustics — Determination of sound power levels and sound energy levels of noise sources using sound pressure — Engineering methods for an essentially free field over a reflecting plane. Standard, International Organization for Standardization, B ezen 2010.
URL <https://www.iso.org/obp/ui/#iso:std:iso:3744:ed-3:v1:en>
- [8] Klapuri, A.; Davy, M.: *Signal Processing Methods for Music Transcription*. 01 2006, ISBN 978-0-387-30667-4, doi:10.1007/0-387-32845-9.
- [9] Large, E. W.; Kolen, J. F.: Resonance and the Perception of Musical Meter. *Connection Science*, ro n k 6, . 2-3, 1994: s. 177–208, doi:10.1080/09540099408915723, <https://doi.org/10.1080/09540099408915723>.
URL <https://doi.org/10.1080/09540099408915723>
- [10] Lartillot, O.; Grandjean, D.: Tempo and Metrical Analysis by Tracking Multiple Metrical Levels Using Autocorrelation. *Applied Sciences*, ro n k 9, 11 2019: str. 5121, doi:10.3390/app9235121.

- [11] Lidy, T.; Rauber, A.: Music Information Retrieval. In *Handbook of Research on Digital Libraries: Design, Development, and Impact*, IGI Global, 2009, ISBN 978-1-59904-879-6, s. 448–456.
- [12] Matthew E. P. Davies, M. F., Sebastian Bock: *Tempo, Beat and Downbeat Estimation*. <https://tempobeatdownbeat.github.io/tutorial/intro.html>, 2021.
URL <https://tempobeatdownbeat.github.io/tutorial/intro.html>
- [13] McAdams, S.; Giordano, B. L.: 113The Perception of Musical Timbre. In *The Oxford Handbook of Music Psychology*, Oxford University Press, 01 2016, ISBN 9780198722946, doi:10.1093/oxfordhb/9780198722946.013.12, https://academic.oup.com/book/0/chapter/292611024/chapter-ag-pdf/44515461/book_34489_section_292611024.ag.pdf.
URL <https://doi.org/10.1093/oxfordhb/9780198722946.013.12>
- [14] McFee, B.; Raffel, C.; Liang, D.; aj.: librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, ro n k 8, 2015.
- [15] McKinney, M. F.; Moelants, D.: Ambiguity in Tempo Perception: What Draws Listeners to Different Metrical Levels? *Music Perception: An Interdisciplinary Journal*, ro n k 24, . 2, 2006: s. 155–166, ISSN 07307829, 15338312.
URL <http://www.jstor.org/stable/10.1525/mp.2006.24.2.155>
- [16] Müller, M.: *Fundamentals of Music Processing*. Springer International Publishing, 2015, doi:10.1007/978-3-319-21945-5.
URL <https://doi.org/10.1007%2F978-3-319-21945-5>
- [17] Müller, M.; Klapuri, A.: Chapter 27 - Music Signal Processing. In *Academic Press Library in Signal Processing: Volume 4, Academic Press Library in Signal Processing*, ro n k 4, editace J. Trussell; A. Srivastava; A. K. Roy-Chowdhury; A. Srivastava; P. A. Naylor; R. Chellappa; S. Theodoridis, Elsevier, 2014, s. 713–756, doi:<https://doi.org/10.1016/B978-0-12-396501-1.00027-3>.
URL <https://www.sciencedirect.com/science/article/pii/B9780123965011000273>
- [18] Salamon, J.; Gomez, E.: Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, ro n k 20, . 6, 2012: s. 1759–1770, doi:10.1109/TASL.2012.2188515.
- [19] Scheirer, E. D.: Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, ro n k 103, . 1, 01 1998: s.

- 588–601, ISSN 0001-4966, doi:10.1121/1.421129, https://pubs.aip.org/asa/jasa/article-pdf/103/1/588/8083614/588_1_online.pdf.
URL <https://doi.org/10.1121/1.421129>
- [20] Schreibman, S.; Siemens, R.; Unsworth, J. (edito i): *A new companion to Digital Humanities*. West Sussex, England: John Wiley & Sons Ltd, 2016, ISBN 9781118680599.
- [21] Sneddon, I.: *Fourier Transforms*. Dover books on mathematics, Dover Publications, 1995, ISBN 9780486685229.
URL <https://books.google.cz/books?id=jhpsLpRRerwC>
- [22] Stevens, S. S.; Volkmann, J.; Newman, E. B.: A Scale for the Measurement of the Psychological Magnitude: Pitch. *The Journal of the Acoustical Society of America*, ro n k 8, . 3, 1937: s. 185–190, ISSN 0001-4966.
- [23] Strichartz, R.: *A Guide To Distribution Theory And Fourier Transforms*. World Scientific Publishing Company, 2003, ISBN 9789813102293.
URL <https://books.google.cz/books?id=YfA7DQAAQBAJ>
- [24] Syrový, V.: *Hudební akustika*. Akustická knihovna Zvukového studia Hudební fakulty AMU, Akademie múzických umění, 2013, ISBN 9788073312978.
URL <https://books.google.cz/books?id=ikrmoAEACAAJ>
- [25] Tumarkin, A.: *The Decibel, The Phon and the Sone*, ro n k 64. Cambridge University Press, 1950, 178–188 s., doi:10.1017/S0022215100011919.
- [26] Wikipedie: Musical Instrument Digital Interface — Wikipedie: Otevřená encyklopédie. 2022, [Online; navštívěno 1. 12. 2022].
URL https://cs.wikipedia.org/w/index.php?title=Musical_Instrument_Digital_Interface&oldid=21081530
- [27] WikiSkripta: Vlastnosti zvuku —. 2022, [Online; navštívěno 21. 11. 2022].
URL https://www.wikiskripta.eu/index.php?title=Vlastnosti_zvuku&oldid=458442

Seznam symbolů a zkratek

MIR	Music information retrieval - Obor zabývající se vyhledávání informací v hudebních dílech
MIDI	Musical Instrument Digital Interface - Digitální rozhraní hudebních nástrojů
ISMIR	International Society of Music Information Retrieval - Mezinárodní združení pro MIR
MIREX	The Music Information Retrieval Evaluation eXchange
FT	Fourier transform - Fourierova transformace
FFT	Fast Fourier transform - Rychlá Fourierova transformace
DFT	Discrete Fourier transform - diskrétní Fourierova transformace
STFT	Short-time Fourier transform - krátkodobá Fourierova transformace
RMS	Root mean square - efektivní hodnota
BPM	Beats per minute - doby za minutu
HTML	Hypertext markup language - hypertextový značkovací jazyk

Seznam příloh