VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

PŘEVOD ZÁZNAMU PIANA Z WAV DO MIDI

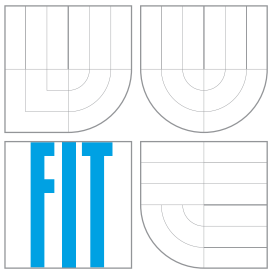
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

JAN BEDNAŘÍK

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

PŘEVOD ZÁZNAMU PIANA Z WAV DO MIDI

CONVERSION OF PIANO RECORDING FROM WAV TO MIDI

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — JAN BEDNAŘÍK
AUTHOR

VEDOUCÍ PRÁCE — Ing. IGOR SZŐKE, Ph.D.
SUPERVISOR

BRNO 2013

Abstrakt

Cílem této práce je návrh systému pro strojový převod polyfonních nahrávek piana z audio formátu WAV do MIDI. Práce popisuje problematiku rozpoznání tónů v hudebních záznamech a předkládá návrh řešení postavený na pravděpodobnostním modelu využívajícím metodu Probabilistic Latent Component Analysis. Pro trénování modelu byly použity nahrávky jednotlivých tónů digitálního piana. Navržený systém byl následně testován na sadě syntetizovaných nahrávek klasické hudby z databáze Classical Piano Midi i na sadě nahrávek piana Korg SP-250 a následně byl vyhodnocen za pomoci odlišných metrik. V závěru jsou výsledky úspěšnosti rozpoznání porovnány s jinými již existujícími systémy.

Abstract

The aim of the thesis is to propose a system capable of automatic conversion of polyphonic piano recordings from the audio format WAV to MIDI. The thesis describes problems related to single tone recognition in music recordings and proposes a solution based on a probabilistic model that uses the Probabilistic Latent Component Analysis method. Recordings of isolated digital piano tones were used to train the system. The proposed system was tested on classical recordings of the Classical Piano MIDI database and on recordings of a Korg SP-250 piano and evaluated using a variety of metrics. The conclusion part contains the results of recognition success rate and their comparison with other existing systems.

Klíčová slova

WAV, MIDI, formát hudebního souboru, piano, rozpoznání tónů, automatický přepis hudby, polyfonie, pravděpodobnostní model, PLCA, analýza hudební nahrávky, MATLAB

Keywords

WAV, MIDI, audio file format, piano, tone recognition, automatic music transcription, polyphony, probabilistic model, PLCA, music recording analysis, MATLAB

Citace

Jan Bednařík: Převod záznamu piana z WAV do MIDI, bakalářská práce, Brno, FIT VUT v Brně, 2013

Převod záznamu piana z WAV do MIDI

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Ing. Igora Szökeho, Ph.D. a uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Jan Bednařík
15. května 2013

Poděkování

Poděkovat bych chtěl především panu Ing. Igorovi Szökemu, Ph.D. za vedení mé práce, ochotu a rady, které mi poskytl. Velký dík patří také panu Ing. Lukáši Burgetovi, Ph.D. za pomoc a konzultace týkající se využité matematické teorie.

Obsah

Kapitola 1

Úvod

Bakalářská práce se zabývá analýzou polyfonních nahrávek piana umožňující rozpoznání jednotlivých tónů a jejich následný export do formátu MIDI. Jedním z využití takové konverze je umožnění okamžitého přenosu hudebního nápadu do notového zápisu. V ideálním případě by mohl skladatel pracovat pouze se samotným hudebním nástrojem, přičemž by mechanickou práci přepisu do not ponechal stroji.

1.1 Formulace problému

Běžně používané digitální audio formáty (jako WAV, MP3, OGG a další) nesou vlastní zvuková data vzniklá sloučením akustických vlnění produkovaných všemi zvukovými zdroji, jež během nahrávání zazněly. Neposkytují však žádnou informaci o tom, jaké zvukové zdroje to byly, kolik jich bylo nebo ve kterých časových intervalech zazněly. Audio formát MIDI naopak zvukovými daty nedisponuje, avšak přesně popisuje, ze kterých hudebních nástrojů nahrávka sestává a v jakých časech se vyskytly jejich tóny.

Převod z MIDI do WAV je tak poměrně přímočarý, neboť pouze syntetizuje zvuky jednotlivých hudebních zdrojů a následně je slučuje do výsledného signálu. Zpětná analýza již však tak snadná není. Ve skutečnosti je vědeckou veřejností rozpoznání tónů v polyfonních nahrávkách obecně považováno za stále nevyřešený problém, a proto se neustále objevují popisy nových metod, jež výsledky zlepšují. Jedním z nejnovějších trendů vykazujícím poměrně dobrou úspěšnost rozpoznání je využití pravděpodobnostních modelů, proto se i návrh systému popsaný v této práci tímto směrem ubírá a staví na metodě Probabilistic Latent Component Analysis (dále PLCA).

Pro testování úspěšnosti byla použita sada syntetizovaných nahrávek dostupných z veřejné databáze Classical Piano Midi i sada reálných nahrávek, přičemž byly výstupy systému porovnány s referenčními MIDI soubory za pomoci dvojice odlišných metrik. V závěru práce je úspěšnost navrženého systému porovnána s jinými již existujícími systémy.

1.2 Rozvržení práce

Kapitola 2 shrnuje základní teoretická fakta a problémy spjaté s rozpoznáváním tónů, kapitola 3 se věnuje principu metody PLCA, a kapitola 4 pak rozebírá vlastní návrh systému a detaily implementace. Kapitola 5 popisuje testovací sadu a použité metriky, kapitola 6 předkládá výsledky testování včetně srovnání s jinými systémy a v závěrečné kapitole 7 jsou nakonec shrnuty možné směry dalšího vývoje.

Kapitola 2

Automatický přepis audio záznamu

Obecným předpokladem kvalitního systému pro převod polyfonních audio záznamů do formátu MIDI je schopnost rozpoznání jednotlivých elementárních zvukových zdrojů, jež se na vzniku výsledného akustického signálu podílely.

K řešení problému lze přistoupit různými způsoby. Objevují se pokusy založené jak na frekvenční tak časové analýze signálu při využití supervizovaného i nesupervizovaného učení modelů a co víc, autoři se často nezaměřují pouze na jeden hudební nástroj nýbrž na obecné hudební nahrávky, čímž se situace dále komplikuje. Nejúspěšnější systémy běžně využívají některý druh klasifikace tónů, ať už za pomoci SVM[1] nebo neuronových sítí, méně úspěšné systémy pak staví na vytváření spektrálních šablon za předpokladu určité struktury spekter tónů.

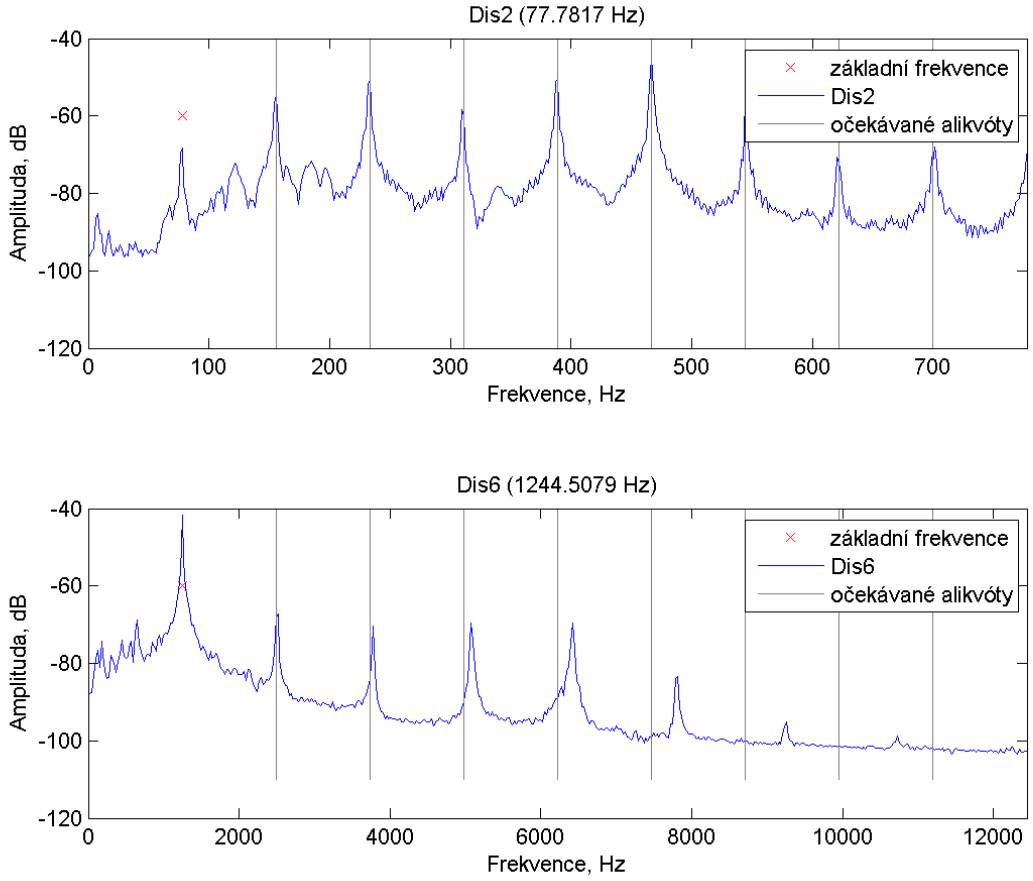
Dalším rozvíjejícím se přístupem je analýza signálu za pomoci pravděpodobnostního modelu Probabilistic Latent Component Analysis (dále PLCA). Tento směr se ukazuje jako poměrně úspěšný a zřejmě také perspektivní vzhledem k tomu, že vědecká komunita původní návrh modelu [9] nadále rozšiřuje [10] [2]. Proto bude i systém navržený v této práci na PLCA postaven. Před vlastním popisem návrhu systému (v následujícím textu označovaném jako WAV2MIDI) budou shrnuty některé známé problémy spjaté s převodem audio nahrávek do symbolické reprezentace a rovněž bude stručně charakterizován samotný výstupní formát MIDI.

2.1 Problémy reálného signálu

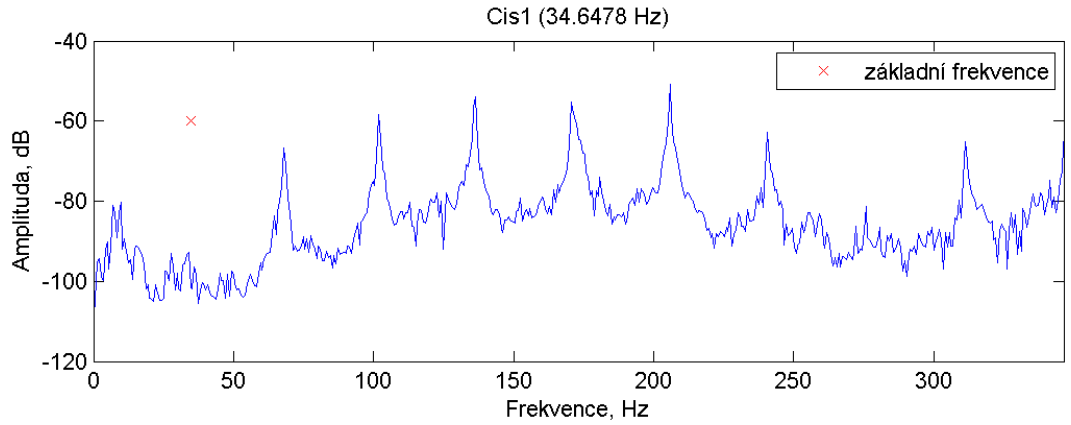
Signály generované reálnými hudebními nástroji se poněkud vymykají teoretickým modelům. Prvním zásadním problémem je neharmoničnost signálů, jev, který lze definovat jako vzdálenost naměřených frekvencí alikvót od očekávaných hodnot ležících na celočíselných násobcích základní frekvence. U piana je tento jev tím citelnější, čím vyšší je základní frekvence a čím vyšší alikvótu sledujeme (viz obrázek 2.1 porovnávající spektra nahrávek tónů *Dis2* a *Dis6* piana Korg SP-250).

Další významnou vlastností reálných signálů je fenomén chybějící základní frekvence (missing fundamental). Jedná se o případ, kdy lidský mozek vnímá výšku tónu o frekvenci, která ve skutečnosti postrádá zdroj vibrací (tedy v signálu fyzicky chybí). V případě piana k tomuto jevu dochází u tónů nižších frekvencí (přibližně u nejnižších osmi tónů), jak je patrné z obrázku 2.2. Kromě zcela chybějící základní frekvence běžně dochází také k jevu, kdy dosahují alikvóty vyšších amplitud, než základní frekvence.

[1] Support Vecoter Machine



Obrázek 2.1: Neharmoničnost tónů Dis2 a Dis6



Obrázek 2.2: Chybějící základní frekvence tónu Cis1

Obě zmíněné skutečnosti znesnadňují proces analýzy výšky tónu a představují limitující faktor pro metody, jež očekávají určitou teoreticky danou strukturu spektra tónů [4]. Naproti tomu modely založené na klasifikátorech nebo pravděpodobnostních modelech jsou

vůči těmto problémům imunní. Od spektra tónů totiž žádnou dopředu pevně danou formu neočekávají a naopak reagují na taková spektra, jaká byla použita k jejich tréninku.

2.2 Formát symbolické reprezentace MIDI

MIDI[2] představuje standard spravovaný asociací MIDI Manufacturers Association a využívaný pro popis hardwarového rozhraní, přenosového protokolu a formátu souboru se symbolickým hudebním zápisem. Formát MIDI je běžně označován jako zvukový formát, přesto sám žádná zvuková data nenese. Obsahuje pouze informace o tom, jaký zvuk má ve kterou chvíli zaznít, přičemž je samotné generování signálu ponecháno na zvukovém zařízení, které protokol MIDI podporuje.

MIDI soubor tak lze označit za symbolickou reprezentaci hudby, neboť má svým uspořádáním velice blízko například k notovému zápisu. Prostřednictvím tzv. MIDI zpráv opatřených časovými značkami je určováno, kdy tón začal a přestal znít, jaký hudební nástroj jej vyprodukoval, zda bylo použito vibrato apod[3].

Protože úkolem systému WAV2MIDI je analýza signálu za účelem rozpoznání jednotlivých tónů, hodí se MIDI formát výborně pro uložení informace o odhadnutých tónech. Formát MIDI je mezi hudebníky výrazně rozšířený a využívaný, díky čemuž se na Internetu pohybuje celá řada amatérských i profesionálních MIDI zápisů. Jednou z takových databází je Classical Piano Midi, jež byla využita pro testování systému WAV2MIDI.

[2] Musical Instrument Digital Interface

[3] Bližší informace lze získat z oficiálního webu asociace MMA http://www.midi.org/

Kapitola 3

Návrh systému WAV2MIDI

Tato kapitola se věnuje vlastnímu konceptuálnímu návrhu systému WAV2MIDI určenému pro strojový převod nahrávek piana v nekomprimovaném audio formátu WAV do symbolické reprezentace MIDI.

Co navrhovaný systém umí? Zjednodušeně řečeno, WAV2MIDI na vstupu přijímá audio nahrávku piana a na výstupu produkuje co nejvěrnější přepis do formátu MIDI. Jádrem systému je pravděpodobnostní model PLCA, který dovede akustický signál dekomponovat na složky, jež lze s výhodou využít pro detekci výskytu jednotlivých tónů.

Za účelem převodu signálu do formátu MIDI bude hlavní snahou získat tzv. piano-roll[1] matici. Piano-roll matice je dvourozměrná struktura, jejíž řádky odpovídají jednotlivým tónům piana a sloupce (zleva doprava) vývoji času. Matice nese binární hodnoty (1/0) značící, zda se daný tón v daném okamžiku vyskytl či nikoliv. Demonstrační piano-roll matice je uvedena na obrázku 3.4.

V sekci 3.1 je nastíněna problematika prvotní analýzy a vhodné reprezentace vstupního audio signálu. Sekce 3.2 se zabývá vlastním principem metody PLCA. V sekci 3.3 je nastíněn postup pro získání piano-roll matice a sekce 3.4 potom shrnuje způsob trénování systému.

3.1 Mid-level reprezentace signálu

Nekomprimované digitální audio nahrávky formátu WAV nesou vlastní zvuková data ve formě navzorkovaného časového průběhu zaznamenaného signálu, jenž obecně představuje součet všech současně znějících zdrojů akustického vlnění. Pro přehrávání zvukového záznamu je pouhá časová reprezentace vyhovující, avšak pro analýzu vedoucí k separaci elementárních hudebních tónů tvořících výsledný signál naopak nedostačující.

Vyvstává tak potřeba transformace signálu do příhodnější, tzv. mid-level (z angl. middle level - střední úroveň) reprezentace [3], jež při procesu analýzy obecně stojí mezi reprezentací s nízkou a vysokou úrovní sémantické složitosti. V případě systému WAV2MIDI bude mid-level reprezentace stát mezi znázorněním časového průběhu signálu (waveform), z nějž nelze přímo získat samostatné hudební tóny, a symbolickým zápisem MIDI, který již nese přesnou informaci o jednotlivých notách.

Nejpoužívanější reprezentací audio dat pro potřeby analýzy akustického signálu je krátkodobá Fourierova transformace (dále STFT[2]) aplikovaná na jednotlivé po sobě jdoucí

[1] V češtině se neustálil ekvivalentní termín, proto bude v následujícím textu využíván anglický termín.
[2] Z angl. Short Time Fourier Transform.

časové rámce [5]. Datovým i vizuálním produktem STFT je spektrogram, jehož upravenou variantu využívá metoda PLCA (viz. 3.2).

Short Time Fourier Transform Transformaci STFT [1] lze pospat vztahem

$$X[k,m] = \sum_{n=mH}^{mH+N-1} s[n]w[n-mH]e^{-j\omega_k n}, \tag{3.1}$$

kde $s[n]$ představuje vstupní signál, $w[n]$ je okno pro výběr časového rámce o délce N vzorků, $m = 0, ..., M-1$ je index časového rámce, H je velikost posunu okna a k značí index frekvenčního pásma (frequency bin). STFT tedy jednoduše představuje sekvenci diskrétních Fourierových transformací (dále DFT[3]) samostatně vypočítaných nad úseky délky N vstupního signálu. V závislosti na velikosti posunu okna H se mohou úseky překrývat. Výběr konkrétních hodnot parametrů STFT je zdůvodněn v kapitole 4.

Spektrogram Spektrogram je výsledkem transformace STFT provedené nad celým signálem. Při vizuálním znázornění se jedná o třírozměrný graf intenzity signálu, kdy horizontální osa reprezentuje čas, vertikální frekvenci a intenzita každého časového rámce o dané frekvenci je znázorněna příslušnou barvou [11].

Protože spektrogram vzniká aplikací STFT na vstupní signál, sestává z hodnot reprezentovaných komplexními čísly. Metoda PLCA však nepracuje s klasickým spektrogramem nýbrž tzv. magnitude spektrogramem[4], který nese pouze absolutní hodnoty původních komplexních čísel a ztrácí tak zcela informaci o fázi.

3.2 Probabilistic Latent Component Analysis

Metoda Probabilistic Latent Component Analysis (dále PLCA) [9] umožňuje dekompozici obecně N-dimenzionální struktury dat (považované za N-dimenzionální pravděpodobnostní funkci) na volitelné množství bází. Z hlediska systému WAV2MIDI lze využít pouze dvoudimenzionální variantu PLCA, kdy jedna dimenze odpovídá času a druhá frekvenci. Počet bází v obou dimenzích pak bude odpovídat počtu kláves piana. Pro lepší pochopení principu PLCA bude nejprve názorně demonstrováno a neformálně popsáno, jak metoda se signálem pracuje, následně budou definovány potřebné termíny a uveden korektní matematicky popis PLCA.

3.2.1 PLCA neformálně

Vstupem metody PLCA je magnitude spektrogram (dále pouze spektrogram) audio signálu. Spektrogram představuje časově-frekvenční reprezentaci signálu a snahou PLCA bude spektrogram do těchto dvou samostatných dimenzí, časové a frekvenční, rozložit. V čase i frekvenci se navíc bude snažit nalézt stanovený počet složek.

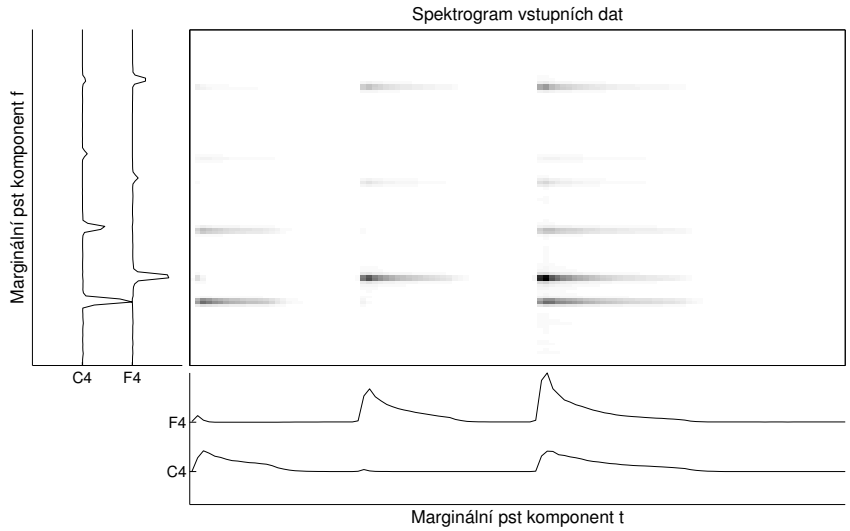
Co si pod tím představit? Zaměřme se v prvé řadě právě na frekvenční dimenzi. Pokud bych chtěl získat frekvenční spektrum signálu, mohl bych využít například DFT. PLCA jde však dále v tom, že takové spektrum rozkládá na jednotlivé složky. To je užitečné proto, že signál může být tvořen více zdroji zvuku (například jednotlivými současně znějícími tóny

[3] Z angl. Discrete Fourier Transform.

[4] Pro termín se neustálil český ekvivalent, proto bude v následujícím textu používána jeho anglická varianta.

piana) a PLCA je pak schopná celé spektrum dekomponovat na složky odpovídající právě těmto samostatným zdrojům zvuku.

Časová doména pak obsahuje právě takový počet složek, kolik jich obsahuje frekvenční doména. Každá z časových složek navíc odpovídá právě jedné frekvenční složce a udává, v jakém čase se daná frekvenční složka vyskytla. Celou situaci demonstruje obrázek 3.1.



Obrázek 3.1: Dekompozice spektrogramu 5s nahrávky piana na 2 frekvenční (levý horní graf) a 2 časové (pravý dolní graf) složky. Vzájemně spjaté frekvenční a časové složky jsou označeny stejným názvem tónu.

Na obrázku je znázorněn spektrogram 5s nahrávky piana, kde zazněl samostatně nejprve tón *C4*, potom samostatně tón *F4* a nakonec oba tóny současně. Tento spektrogram je metodou PLCA dekomponován do frekvenční a časové dimenze, přičemž v čase i frekvenci hledá dvě složky. Celý model používá jisté zjednodušení. Spektrum tónu piana považuje během celého jeho znění za stacionární (v praxi však spektrum během znění tónu kolísá), na základě čehož lze považovat jednu frekvenční komponentu za první tón (*C4*) a druhou frekvenční komponentu za druhý tón (*F4*). Obě jim odpovídající časové složky potom určují, ve kterém čase se jednotlivé tóny vyskytly (pravděpodobnost jejich výskytu).

Pro úplnost zmiňme, že z těchto komponent je navíc možné do určité míry zpětně rekonstruovat původní spektrogram. Představme si frekvenční složky jako sloupcové vektory a časové složky jako řádkové vektory. Pokud vynásobíme vzájemně si odpovídající časový a frekvenční vektor, získáme matici reprezentující jakousi vrstvu rekonstruovaného spektrogramu. Pokud všechny tyto vrstvy sečteme, získáme výsledný rekonstruovaný spektrogram.

3.2.2 PLCA formálně

Metoda PLCA představuje pravděpodobnostní generalizaci metody NMF[5], přičemž dvoudimenzionální varianta PLCA (již systém WAV2MIDI využívá) je numericky identická NMF. Oproti NMF však PLCA jakožto pravděpodobnostní model vyžaduje, aby měly frekvenční a časové složky formu korektních pravděpodobnostních funkcí $P[X = x]$, pro něž platí

[5]Non-negative Matrix Factorization

$\sum_x P[X = x] = 1$ a současně nesou pouze nezáporné hodnoty.

Základní model metody PLCA [9] je definován vztahem:

$$P(\mathbf{x}) = \sum_z P(z) \prod_{j=1}^{N} P(x_j|z), \tag{3.2}$$

kde $P(x)$ je N-dimenzionální pravděpodobnostní funkce náhodné veličiny $\mathbf{x} = x_1, x_2, ..., x_N$, z je latentní proměnná, $P(z)$ je apriorní pravděpodobnost latentní proměnné a $P(x_j|z)$ jsou marginální pravděpodobnosti.

Jak vidno, PLCA obecně pracuje s N-rozměrnými daty, avšak jak již bylo zmíněno, pro potřeby rozpoznání hudebních tónů dostačují dvě dimenze, a sice frekvenční a časová. Základní model PLCA pak lze definovat specifičtěji jako

$$P(\mathbf{x}) = \sum_z P(z)P(f|z)P(t|z), \tag{3.3}$$

kde dvourozměrná pravděpodobnostní funkce $P(x)$ představuje spektrogram akustického signálu a $P(f|z)$ a $P(t|z)$ jsou marginální pravděpodobnosti ve frekvenční a časové doméně odpovídající pouze jedné komponentě z.

Jednotlivé komponenty (říkejme jim frekvenční báze), o nichž byla řeč v sekci 3.2.1, jsou jednorozměrné funkce $P(f|z)$ a je jich právě tolik, kolik si zvolím latentních proměnných z. Každá frekvenční báze $P(f|z)$ je pevně spjatá s jednou časovou bází $P(t|z)$, jež svou hodnotou značí míru příspěvku frekvenční báze do daného časového rámce. V tomto kontextu lze opět jako demonstrační příklad uvést obrázek 3.1 znázorňující dekompozici signálu při volbě $z = 2$. Jsou zde znázorněné dvě frekvenční a dvě časové báze co by marginální pravděpodobnosti frekvenční a časové domény.

V sekci 3.2.1 byla popsána rekonstrukce spektrogramu jako součet jednotlivých "vrstev". Tyto vrstvy vznikají jako součin dvojic odpovídajících si frekvenčních a časových bází vynásobených příslušnou váhou $P(z)$ a lze je považovat za příspěvek dané latentní proměnné do celého rekonstruovaného spektrogramu. Tento proces matematicky popisuje výše zmíněný základní vztah 3.3. $P(x)$ v závislosti na počtu iterací výpočtu a množství latentních proměnných z více či méně odpovídá původnímu (referenčnímu) spektrogramu.

Počet latentních proměnných z odpovídá počtu komponent získaných v každé dimenzi a určuje tak míru detailu analýzy. Je zřejmé, že při vyšších počtech z ponesou komponenty detailnější informaci. To však nutně nemusí být žádoucí jak z hlediska výpočetní náročnosti tak z pohledu následné interpretace významu jednotlivých komponent. Množství z je tak nutné vhodně volit s ohledem na konkrétní aplikaci metody (viz sekce 3.4).

Jádrem analýzy PLCA je tedy odhad maximální věrohodnosti (maximum likelihood estimation) časových a frekvenčních bází $P(t|z)$ a $P(f|z)$ představujících parametry systému a jejich apriorních pravděpodobností (vah) $P(z)$ ze vstupního spektrogramu. Samotné latentní proměnné z zde hrají roli pouhého indexu představujícího pojítko mezi komponentami obou dimenzí. Iterační proces staví na EM algoritmu, kde krok Expectation slouží pro výpočet posteriorní pravděpodobnosti

$$P(z|f,t) = \frac{P(z)P(f|z)P(t|z)}{\sum_{z'} P(z')P(f|z')P(t|z')} \tag{3.4}$$

a v kroku Maximization je na základě nové hodnoty $P(z|f,t)$ stanoven přesnější odhad

věrohodností frekvenčních a časových bází a apriorních pravděpodobností podle vztahů

$$P(z) = \sum_f \sum_t P(z|f,t)P(f,t), \tag{3.5}$$

$$P(f|z) = \frac{\sum_t P(z|f,t)P(f,t)}{P(z)}, \tag{3.6}$$

$$P(t|z) = \frac{\sum_f P(z|f,t)P(f,t)}{P(z)}. \tag{3.7}$$

Iterativní výpočet EM algoritmu vede ke konvergujícímu řešení pro frekvenční a časové báze i apriorní pravděpodobnosti, přičemž vhodný (dostatečný) počet kroků je možné stanovit za pomocí metody KL divergence.

Nutno podotknout, že vstupem pro analýzu PLCA je pouze spektrogram signálu, avšak jak je patrné ze vztahu 3.4, metoda ihned pracuje s hodnotami $P(f|z)$, $P(t|z)$ a $P(z)$. Ty je nutné před zahájením výpočtu vhodným způsobem inicializovat (viz kapitola 4).

Závěrem dlužno zmínit, že využití metody PLCA je poměrně mnohostranné a neomezuje se pouze na automatický přepis hudebních nahrávek do notového zápisu. Mezi možné aplikace spadá také odrušení signálu, extrakce příznaků, rozpoznání řeči nebo separace zvukových zdrojů [9].

3.3 Odhad piano-roll matice

Nyní, když je již znám princip metody PLCA, se nabízí otázka, k čemu je to vše z hlediska návrhu systému WAV2MIDI dobré. Pro rozpoznání tónů budou nejdůležitější časové báze, neboť pakliže by se podařilo zajistit, aby každá dekomponovaná frekvenční báze odpovídala frekvenčnímu spektru právě a pouze jednoho tónu piana (tím se zabývá sekce 3.4), její odpovídající časová báze již pouze ukáže, kdy a na jak dlouho konkrétní tón zazněl.

Jednotlivé časové báze považujme za řádkové vektory, jejichž vertikálním umístěním nad sebe vznikne matice (pojmenujme ji matice časových bází) velice se blížící piano-roll matici. Nenese však binární hodnoty ostře rozhodující o přítomnosti tónu v daném časovém rámci, nýbrž sestává z hodnot značících pravděpodobnost výskytu odpovídajícího frekvenčního spektra (tónu).

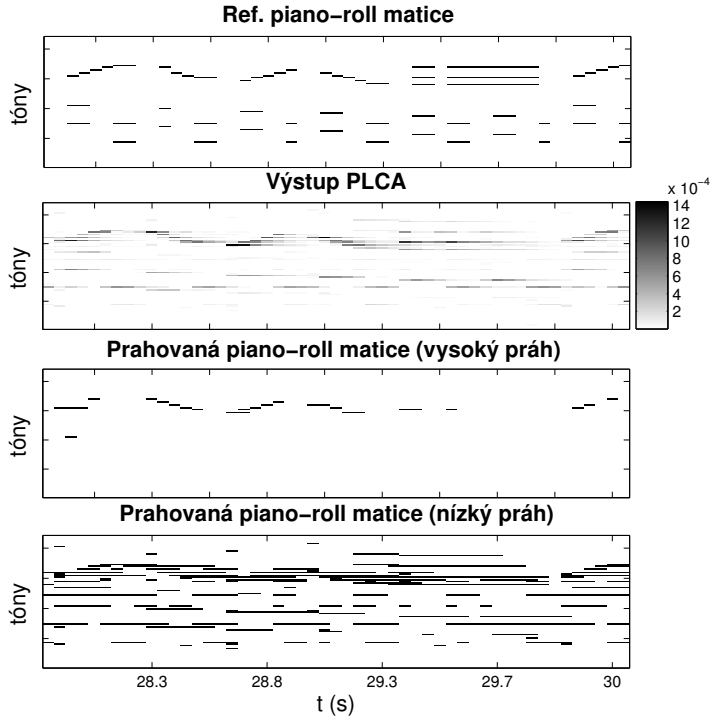
Pro získání binární piano-roll matice je nutné matici časových bází prahovat (viz sekce 3.3.1), zkvalitnění výsledků lze pak dosáhnout post-processingem (viz sekce 3.3.2).

3.3.1 Prahování

Výběr hodnoty prahu pro převod mezi výstupem PLCA (maticí časových bází) a binární piano-roll maticí významným způsobem ovlivňuje správnost rozpoznání jednotlivých tónů. Při volbě příliš vysokého prahu systém některé tóny vůbec nezaznamená, zatímco při volbě příliš nízkého prahu se do piano-roll matice dostanou také tóny, které v originální nahrávce vůbec nezazněly.

Jakým způsobem byla vybrána výchozí hodnota prahu pro systém WAV2MIDI? Protože je PLCA pravděpodobnostní model, musí součet přes všechny buňky matice časových bází nezávisle na jejich počtu dávat v součtu hodnotu 1. Na tuto hodnotu se můžeme dívat jako na energii distribuovanou mezi jednotlivé buňky a pakliže má buňka dostatečnou energii a překročí zvolený práh, dostane se do výsledné piano-roll matice. Práh byl tedy volen jako podíl doposud neznámé hodnoty (označme ji třeba x) a počtu buněk matice časových

bází. Vhodná hodnota pro x (viz kapitola 4) byla určena experimentálně. Vliv velikosti prahu na výslednou binární piano-roll matici demonstruje obrázek 3.2. Dodejme, že tento přístup nezohledňuje hustotu výskytu tónů ani jejich hlasitost, což má v některých případech negativní dopad na úspěšnost rozpoznání (viz kapitola 6).



Obrázek 3.2: Demonstrace vlivu velikosti prahu na úspěšnost rozpoznání.

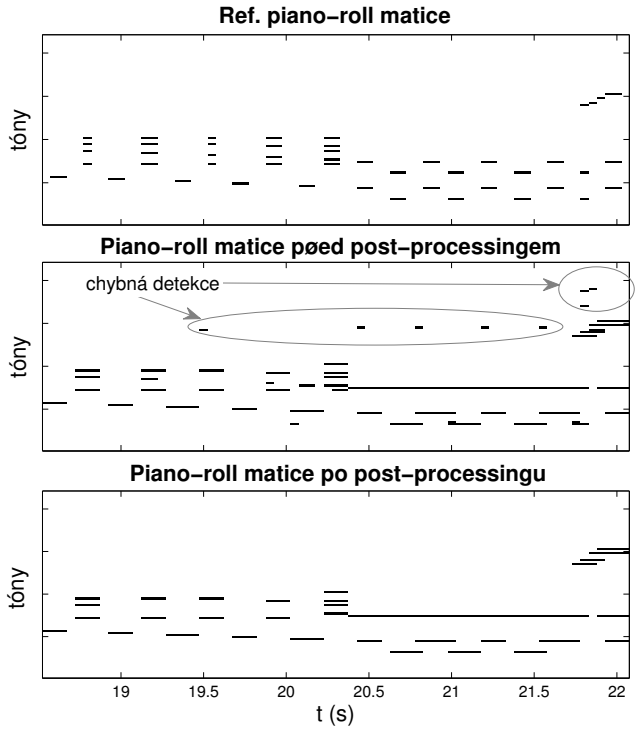
3.3.2 Post-processing

Pro post-processing byla využita velmi jednoduchá metoda, a sice průchod prahovanou piano-roll maticí a odstranění takových detekovaných tónů, jejichž délka je kratší než zvolený práh. Tento přístup, leč svou podstatou triviální, však velkou mírou přispívá ke zvýšení úspěšnosti rozpoznání, neboť částečně řeší problém výskytu neočekávané frekvence při náběhu tónů. Zjednodušeně řečeno se jedná o to, že každý tón při svém náběhu (počátku znění) na velice krátkou chvíli produkuje energii také do jiných frekvenčních pásem, něž do kterých by měl, což vyústí v detekci krátkých tónů, které se v nahrávce nevyskytly. Tyto krátké tóny je post-processing schopen odstranit, což demonstruje obrázek 3.3. Vhodný práh pro délku takového tónu (v dalším textu odkazován zkratkou *dkt*[6]) byl opět určen experimentálně (viz kapitola 4).

3.4 Trénování

Jak docílit toho, aby každá frekvenční báze co nejpřesněji odpovídala spektru právě jednoho tónu piana? Metodu PLCA je nutné nad daty nejprve natrénovat při vhodné volbě počtu

[6] délka krátkého tónu



Obrázek 3.3: Demonstrace vlivu post-processingu na úspěšnost rozpoznání. V prostředním grafu jsou elipsami zaznačeny chybně rozpoznané tóny oproti referenční piano-roll matici (horní graf).

bází (odpovídajícímu počtu latentních proměnných z), na něž má signál dekomponovat.

Trénování lze provádět dvěma způsoby, a sice nesupervizovaně a supervizovaně. V prvním případě metoda PLCA neví nic o tom, jak vypadají frekvenční spektra skutečných tónů piana, které vyprodukovalo vstupní signál, ani o počtu různých tónů, které se v nahrávce naskytly. Musí tak pouze se znalostí počtu očekávaných bází co nejlépe dekomponovat signál jak na frekvenční tak časové báze (a jejich váhy). Přesto je i v tomto případě metoda schopná při volbě vhodného počtu z sama poměrně úspěšně dekomponovat signál na frekvenční báze, které reálným tónům odpovídají. Situace se však ztěžuje při narůstající polyfonii, jak je ostatně demonstrováno v sekci 3.4.1.

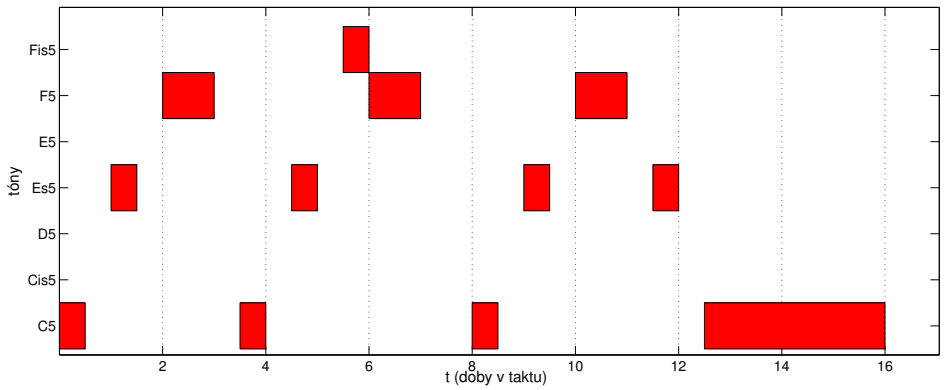
V druhém případě má zajisté metoda PLCA o mnoho výhodnější výchozí pozici, neboť disponuje bankou frekvenčních spekter všech tónů piana a nemusí tyto báze odhadovat. Současně PLCA vlastní informaci o tom, na kolik bází má signál rozložit, neboť tento odpovídá počtu tónů vložených do banky spekter. Její jedinou starostí je správná dekompozice signálu na časové báze (a jejich váhy), jež popisují výskyt předem daných spekter skrze všechny časové rámce. Supervizovanému trénování se věnuje sekce 3.4.2.

3.4.1 Nesupervizované trénování

Pakliže by metoda PLCA dokázala ve vstupním signálu sama rozpoznat frekvenční báze odpovídající spektrům tónů daného piana, bylo by vyhráno. Různá piana však produkují pro stejné tóny různá spektra a pokud je tedy metodě PLCA předložena banka spekter (viz 3.4.2) extrahovaná z jednoho piana, pro jiné piano rozpoznání nemusí dobře fungovat.

Bohužel je PLCA schopná dekompozice na "správná" spektra jen do určité míry. Jak bude demonstrováno dále, úspěšnost v tomto ohledu silně závisí na obsahu vstupního signálu.

Monofonní nahrávka Schopnosti PLCA při nesupervizovaném trénování nejprve ověřme na monofonní nahrávce `plca_mono_smoke.wav` (nahrávka je obsažená na přiloženém CD). Nahrávka sestává ze čtyř různých tónů *C5*, *Es5*, *F5* a *Fis5*, jež dohromady zazní dvanáctkrát. Referenční piano-roll matice nahrávky je uvedena na obrázku 3.4. Pro lepší představu o průběhu signálu a orientaci v následujících grafech čtenáře vyzývám k poslechu nahrávky.



Obrázek 3.4: Referenční piano-roll nahrávky `plca_mono_smoke.wav` získaný z manuálně vytvořeného ekvivalentního MIDI souboru.

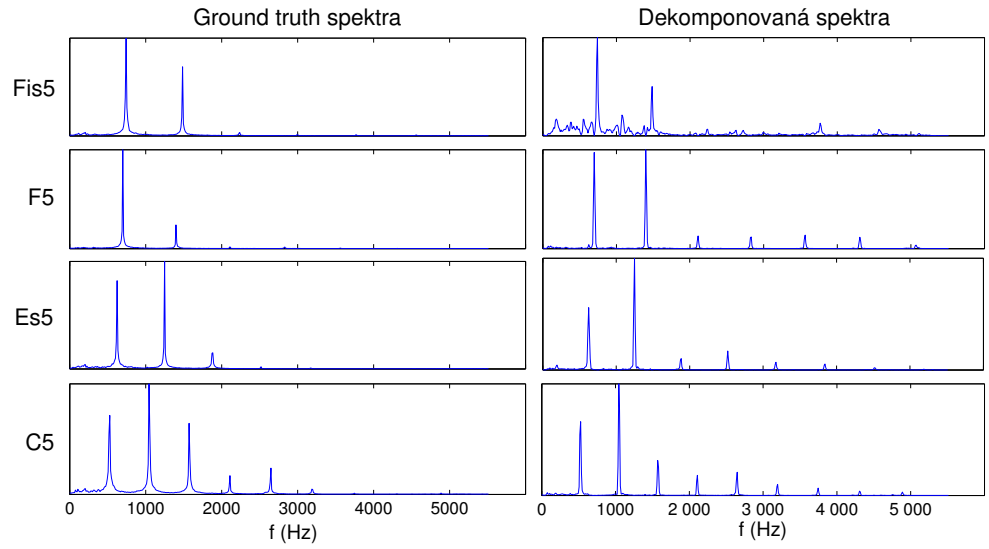
V prvé řadě necháme signál analyzovat metodou PLCA s explicitním požadavkem na nalezení čtyř frekvenčních bází, které by měly dle očekávání odpovídat frekvenčním spektrům jednotlivých tónů. Jak je patrné z obrázku 3.5, dekomponované frekvenční báze se opravdu velmi blíží referenčním bázím (v obrázku označených jako ground-truth[7]) jednotlivých tónů.

Problém nastává pouze u frekvenčního spektra tónu Fis5, kterému PLCA kromě očekávané základní frekvence a případných alikvót přidělila energii také v ostatních frekvenčních pásmech (nejmarkantněji je to patrné v pásmu od 0 do 2000 Hz).

Proč k tomu došlo? Podívejme se, jak celá situace vypadá na spektrogramu. Obrázek 3.6 znázorňuje referenční spektrogram signálu a spektrogram rekonstruovaný pomocí PLCA z dekomponovaných bází.

Z referenčního spektrogramu lze vypozorovat fakt, že každý z tónů během několika prvních časových rámců svého znění produkuje energii nejen do pásma odpovídajícího jeho základní frekvenci, ale také přes poměrně široké spektrum nižších i vyšších frekvencí. Tento fenomén je blíže rozebrán v sekci 3.4.2, nyní si však všimněme faktu, že tato rozložení energie přes široké frekvenční spektrum vzniklá při náběhu (onset) tónu přiřkla PLCA pouze frekvenční bázi tónu *Fis5*. V rekonstruovaném spektrogramu je to dobře znát, tón *Fis5* zde nese v pásmu své základní frekvence menší energii (azurová barva oproti syté červené v referenčním spektrogramu), avšak během celého svého znění distribuuje energii do širokého frekvenčního pásma (vertikální světle modrý pás nad a pod tónem *Fis5* v rekonstruovaném spektrogramu).

[7]Pojem používaný v anglické literatuře pro obecné označení referenčních dat.



Obrázek 3.5: Grafy zachycují porovnání frekvenčních spekter čtyřech tónů piana *C5*, *Es5*, *F5* a *Fis5* získaných pomocí DFT (levý sloupec) a dekomponovaných z monofonní nahrávky (pravý sloupec). Všechny nahrávky pocházejí ze stejného piana. Z grafů je patrné, že si spektra z hlediska umístění základních frekvencí a alikvót poměrně dobře odpovídají.

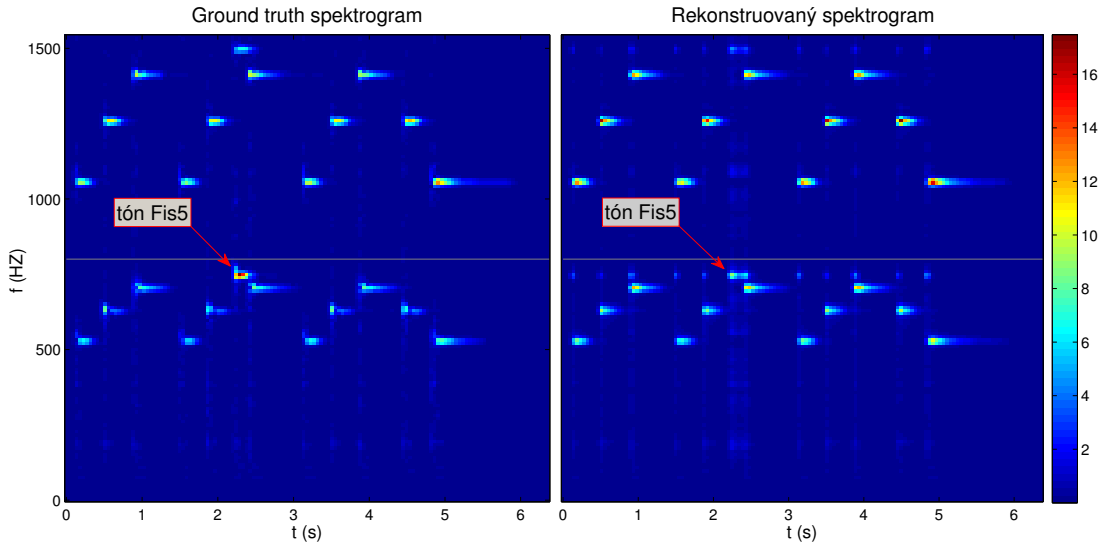
Tento výsledek analýzy si lze představit také tak, že PLCA posbírala spektra odpovídající náběhům všech tónů, nakombinovala a přiřkla do jedné z bází, která v tomto případě odpovídá tónu *Fis5*. To je v rekonstruovaném spektrogramu rovněž patrné v horizontální linii v úrovni tónu *Fis5* ve formě krátkého a nepříliš intenzivního výskytu tónu *Fis5* při každém náběhu kteréhokoliv jiného tónu. Stejně lze tento jev pozorovat v obrázku 3.7 znázorňujícím časové báze jednotlivých tónů.

Možným vysvětlením tohoto jevu je fakt, že v celé nahrávce se tón *Fis5* vyskytuje pouze jednou a PLCA tedy nemusí mít dostatek informace o jeho spektru.

Pakliže je však nahrávka `plca_mono_smoke.wav` prostřednictvím PLCA dekomponována na pět bází, vypadá situace o mnoho přívětivěji. Z obrázku 3.8 je patrné, že PLCA tentokrát správně rozpoznala také tón Fis5 a oddělila zbylé frekvenční spektrum nástupů tónů do samostatné báze.

Polyfonní nahrávka Monofonní nahrávky z předchozího příkladu doplňme o podklad opakující se dvojice akordů *C4-Es4-G4* a *C4-F4-A4* (nahrávka `plca_poly_smoke.wav` opět obsažená na přiloženém CD). Takto v nahrávce vznikne polyfonie až 4 tónů současně, jak je znázorněno v referenční piano-roll matici na obrázku 3.9.

V nahrávce se nyní vyskytuje devět různých tónů, pokusme se tedy signál analyzovat při volbě $z = 9$. Tentokrát však již PLCA selhává. Protože jednotlivé tóny obou v signálu se vyskytnuvších akordů nikdy nezazní samotné, není je PLCA schopná rozpoznat. Na obrázku 3.10 jsou znázorněny čtyři z celkových devíti frekvenčních bází, které lze považovat za samostatné tóny. PLCA byla schopna analyzovat samostatné tóny melodie (nikoliv akordů)



Obrázek 3.6: Referenční (levý graf) a rekonstruovaný (pravý graf) spektrogram signálu `plca_mono_smoke.wav`. Spektrogramy jsou horizontální šedou čárou rozděleny na dvě části, kde spodní část znázorňuje základní frekvence tónů, horní část pak jejich první alikvóty. Z hlediska popisu problematiky nesupervizovaného trénování stačí věnovat pozornost pouze spodním částem, horní části jsou zde uvedeny pouze pro demonstraci vzniku energie napříč celým spektrem během náběhů tónů.

s tím, že frekvenční báze nejvíce odpovídající tónu F5 je znehodnocena výskytem cizího spektra, a sice tónu *A4* (viz obrázek 3.10).

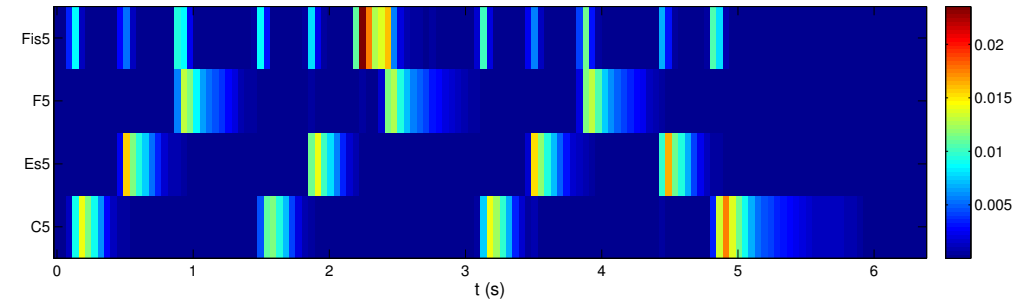
Proč k tomu došlo? Jak je patrné z referenční piano-roll matice, tón *F5* na rozdíl od zbylých tónů melodie v signálu nikdy nezazní zcela sám, nýbrž se vyskytuje společně s akordem *C4-F4-A4*. Z tohoto akordu se během analýzy do frekvenční báze tónu *F5* vkradla základní frekvence tónu *A5*.

Vnucuje se myšlenka, že PLCA nedisponující žádnou apriorní znalostí o spektrech tónů nedekomponuje signál na frekvenční spektra odpovídající jednotlivým tónům, nýbrž na spektra odlišných úseků signálů, které se v celém signálu objevily samostatně. V nahrávce `plca_poly_smoke.wav` se opakuje dvojice akordů, jejichž složky vždy zazní pouze současně. Spustím-li PLCA analýzu pro $z = 6$, kde očekávám čtyři báze pro tóny melodie a dvě báze pro akordy, vskutku se ve výsledku ve dvou frekvenčních bázích objeví společně složky jednotlivých akordů, jak je patrné z obrázku 3.11) .

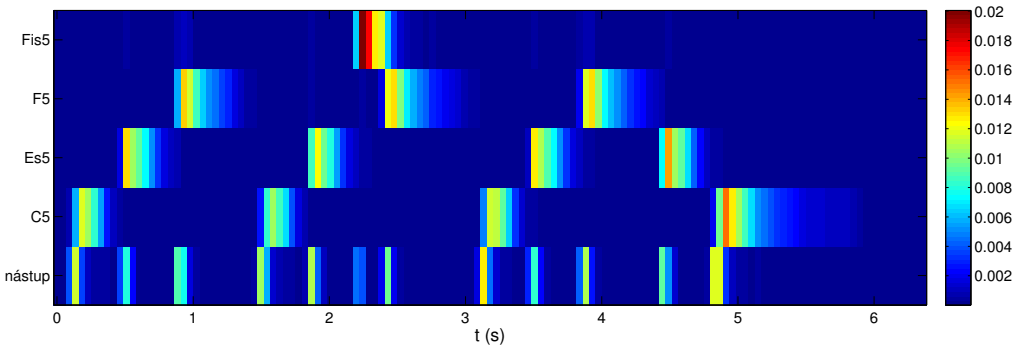
Nutno s povzdechnutím konstatovat, že metoda PLCA je bez dodání apriorních informací o frekvenčních spektrech tónů poměrně naivní. Jednotlivé tóny rozpozná pouze tehdy, když v nahrávce zazní samostatně, což se samozřejmě od reálných hudebních skladeb nedá očekávat. Z toho plyne, že nesupervizované trénování nebude pro účely systému WAV2MIDI dostačovat.

3.4.2 Supervizované trénování

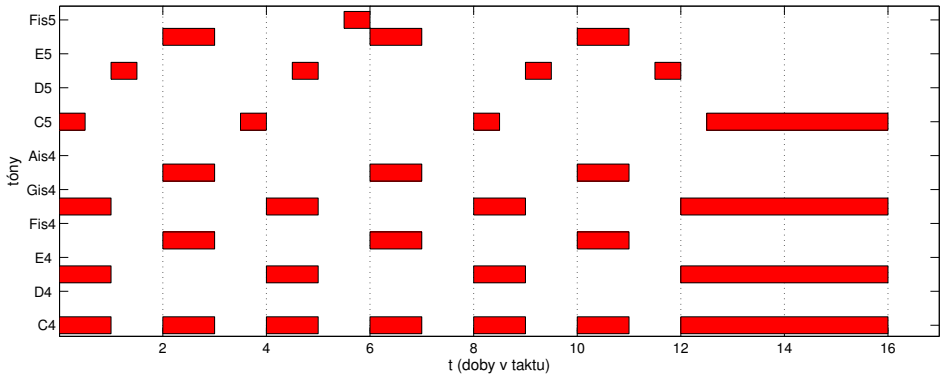
Celá analýza PLCA bude fungovat výrazně robustněji, pokud bude mít k dispozici informaci o tom, jak vypadají frekvenční báze všech jednotlivých tónů piana, které se mohou v nahrávce vyskytnout. Tehdy nebude nucena spektra sama rozpoznávat, nýbrž bude jejím



Obrázek 3.7: Časové báze $P(t|z)$ vzniklé dekompozicí signálu `plca_mono_smoke.wav` při $z = 4$. Graf lze poměrně dobře porovnat s referenční piano-roll maticí na obrázku 3.4, problém zde činí pouze tón *Fis5*, který ve své frekvenční bázi nese části frekvenčních spekter náběhů všech ostatních tónů.



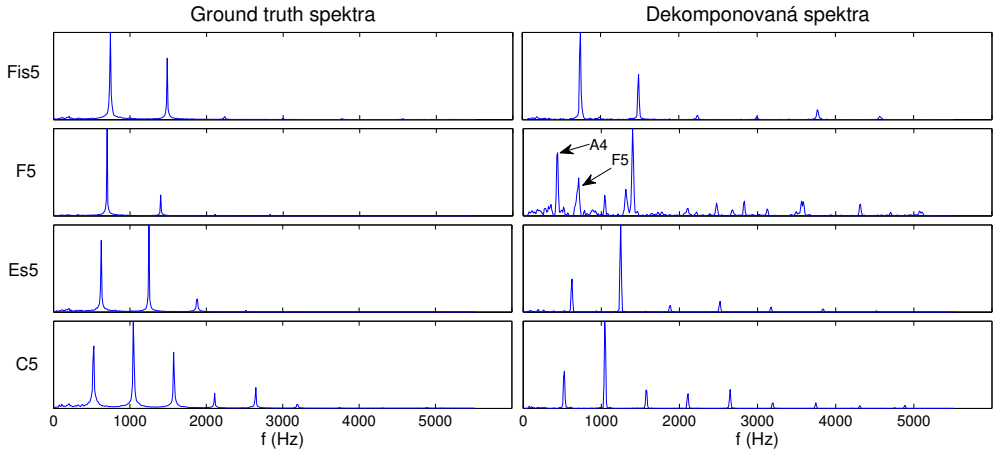
Obrázek 3.8: Časové báze $P(t|z)$ vzniklé dekompozicí signálu `plca_mono_smoke.wav` při $z = 5$. PLCA zde správně rozpoznala jednotlivé tóny a frekvenční spektrum odpovídající nástupům tónů separovala do samostatné báze.



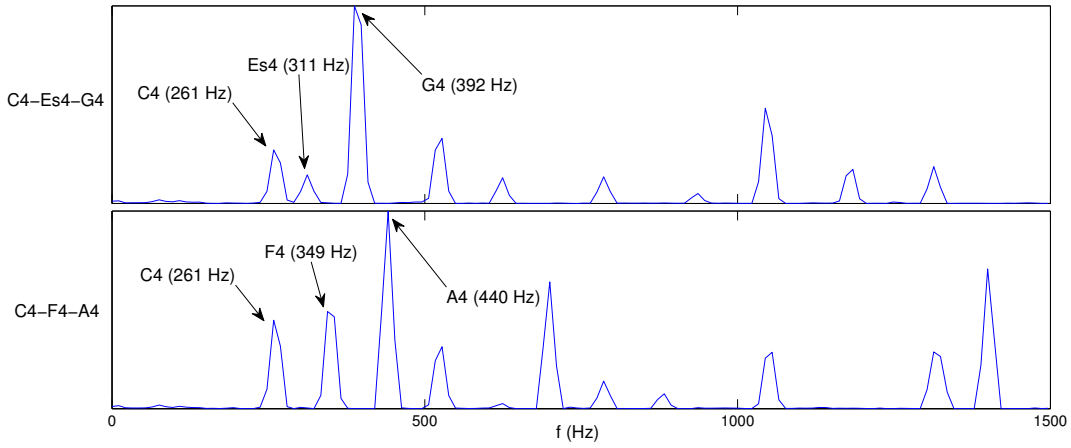
Obrázek 3.9: Referenční piano-roll matice nahrávky `plca_poly_smoke.wav` získaná z manuálně vytvořeného ekvivalentního MIDI souboru.

jediným úkolem odhadnout jejich časové průběhy.

Obecný algoritmus pro dvoudizmenzionální varianty PLCA stanovený v 3.2.2 se zjednoduší tím, že nebude nutné inicializovat frekvenční báze $P(f|z)$ (jsou dopředu známy),



Obrázek 3.10: Srovnání referenčních spekter tónů *C5*, *Es5*, *F5* a *Fis5* se spektry získanými z nahrávky `plca_poly_smoke.wav` pomocí PLCA.



Obrázek 3.11: Dvě z šesti frekvenčních bází, jež metoda PLCA dekomponovala z nahrávky `plca_poly_smoke.wav`. V grafech jsou zaznačeny základní frekvence jednotlivých tónů obou akordů.

a současně odpadne potřeba v kroku M zpřesňovat jejich odhad. V prvé řadě je nezbytné získat frekvenční spektra pro všechny uvažované tóny, jež budou tvořit banku frekvenčních bází metody PLCA.

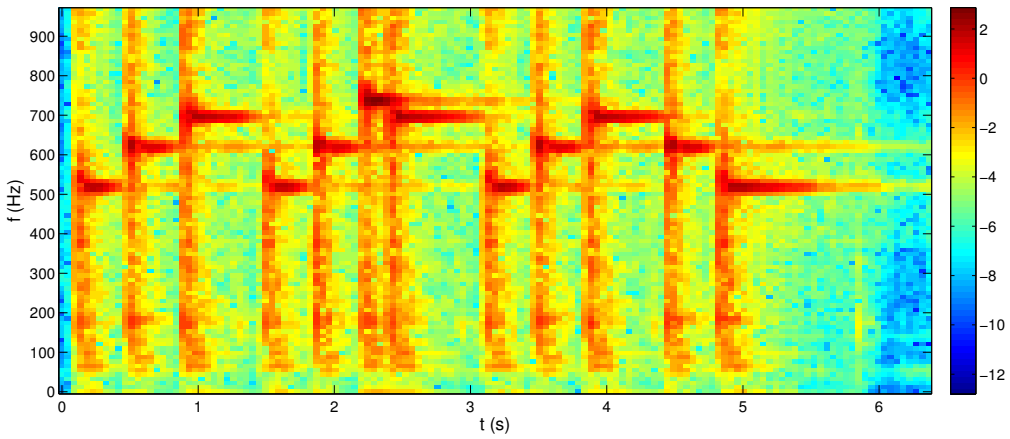
Jaký postup při návrhu banky zvolit? Pro extrakci frekvenční báze tónu je nejvhodnější použít opět PLCA. Pro každý z uvažovaných tónů piana[8] byl pořízen 10s záznam obsahující 6–10 opakovaných úhozů daného tónu při proměnlivých hlasitostech a délkách znění. Pravděpodobně by postačily i kratší záznamy, nicméně delší signál a potažmo větší počet

[8]Pro trénování bylo využito digitální piano Korg SP-250, jež disponuje klaviaturou o plném rozsahu 88 kláves.

úhozů jistě nese o spektru generovaném daným tónem přesnější informaci.

Dalším krokem je spuštění analýzy PLCA nad všemi pořízenými nahrávkami. Nabízí se využít jedinou latentní proměnnou z, neboť v signálu se přeci vždy vyskytuje pouze jeden tón. Vzhledem k vývoji spektra tónu během jeho náběhu a následného znění však lze k problému přistoupit jiným způsobem.

Obrázek 3.12 demonstruje část spektrogramu signálu `plca_mono_smoke.wav`, v němž jsou patrné základní frekvence jednotlivých tónů (graf odpovídá dolní polovině referenčního spektrogramu z obrázku 3.6 po aplikaci logaritmu).



Obrázek 3.12: Graf demonstruje spektrogramu signálu `plca_mono_smoke.wav`. Pro názornější ukázku problematiky měnícího se spektra v rámci průběhu tónu, je spektrogram vyobrazen ve zlogaritmované podobě.

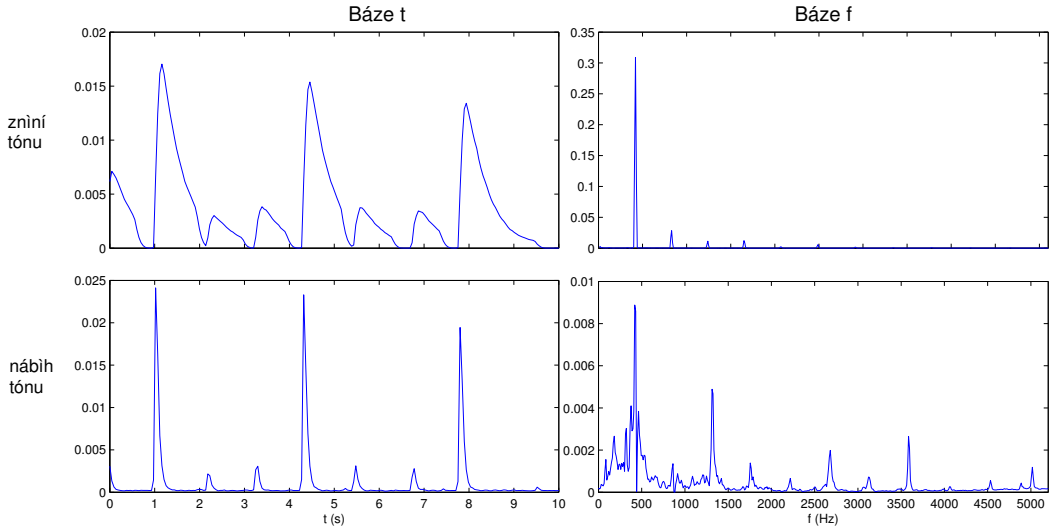
Přestože by měl v ideálním případě každý tón produkovat akustickou energii pouze ve frekvenčním pásmu své základní frekvence a alikvót, v praxi tomu tak bohužel není. Ve spektrogramu je patrné, že během náběhu tónu, neboli v několika prvních časových rámcích znění tónu, vzniká akustická energie téměř napříč celé spektrum (každý tón se chová poněkud odlišně).

Namísto trénování s jednou latentní proměnnou je vhodné pro každý tón stanovit latentní proměnné dvě, díky čemuž se jedna frekvenční báze natrénuje na náběh tónu a druhá na jeho následné znění. To je dobře patrné z obrázku 3.13, který demonstruje PLCA analýzu tónu *A4*. Svrchní řádek zde zobrazuje časovou a frekvenční bázi znění tónu, spodní řádek pak odpovídající báze pro náběh tónu.

Z frekvenčních i časových bází je patrné, že se metoda PLCA zachovala dle očekávání. Časová báze odpovídající náběhu vykazuje velice krátké výskyty energie, jež se při pohledu na druhou časovou bázi vyskytne vždy na začátku znění tónu. Ve frekvenční doméně si pak lze všimnout, že báze pro náběh tónu rozkládá akustickou energii přes široké pásmo, zatímco báze pro znění tónu nese "čisté" špičky na pozicích základní frekvence a alikvót.

Proč se frekvenční báze odpovídající náběhu rovnou nezbavit a nepoužít pouze samotné znění tónu? U velmi krátce znějících tónu (v hudbě označovaných jako *staccato*) nese velkou část energie celého tónu právě spektrum odpovídající náběhu. Pokud bychom jej zanedbali, nemusely by být takové tóny korektně rozpoznány.

Ve výsledné bance frekvenčních bází je přesto pro každý tón použito pouze jedno spektrum. To lze získat součtem obou frekvenčních bází váhovaných apriorními pravděpodob-



Obrázek 3.13: Grafy znázorňují výsledek PLCA analýzy 10s signálu, v němž při různých hlasitostech 8krát zazní tón *A4*. Levý sloupec schraňuje časové báze, pravý pak frekvenční báze. Svrchní řádek odpovídá časově delší části každého tónu zde označované jako znění, spodní řádek pak demonstruje náběh tónu.

nostmi $P(z)$. Vzniklá frekvenční báze se poněkud podobá případu, kdy by byly tóny analyzovány pomocí PLCA s jednou latentní proměnnou. Vzhledem k odlišnému způsobu získání výsledného spektra se nicméně oba případy, leč nepatrně, liší a navrhovaný způsob s využitím dvou z by měl lépe charakterizovat reálné chování tónu.

Kapitola 4

Implementace

Tato kapitola se věnuje vlastní implementaci systému WAV2MIDI a technickým detailům spjatým s návrhem testů a zpracováním zvukových dat. Sekce 4.1 shrnuje použité softwarové technologie, sekce 4.2 popisuje některé důležité podrobnosti spjaté s implementací a během systému a sekce 4.3 nakonec vysvětluje výběr konkrétních výchozích hodnot parametrů systému.

4.1 Použité technologie

Pro implementaci metody PLCA, jež představuje jádro systému WAV2MIDI, její trénování i následné testování a vyhodnocení byl použit programovací jazyk MATLAB. Pro práci s MIDI soubory (především pro jejich načítání do vnitřní programové reprezentace, pro zpětný export a pro vizualizaci) program dále využívá knihovny Midi Toolbox[1] a Midi Tools[2].

Systém byl testován nad sadou syntetizovaných nahrávek a sadou nahrávek reálného piana. Syntéza z MIDI do WAV byla provedena za pomocí softwarového nástroje Direct MIDI to MP3 Converter[3]. Nahrávky reálného piana byly před vlastním testováním předzpracovány filtrem redukujícím šum. Pro tento účel byl využit open-source zvukový editor Audacity[4]. Referenční MIDI soubory pro nahrávky reálného piana byly vytvořeny v prostředí programu FL Studio 11[5].

4.2 Implementační detaily

Aplikace Samotný systém WAV2MIDI je navržen jako konzolová aplikace, již lze spustit jako volání příslušného skriptu, jenž přijímá audio nahrávku ve formátu WAV a produkuje co možná nejvěrnější ekvivalent ve formátu MIDI. Pro detailnější práci s audio daty, vizualizaci výsledků a případné přizpůsobování parametrů je možné volat dílčí specifické skripty samostatně. Součástí přiloženého CD jsou rovněž skripty implementující metriky a testy použité k vyhodnocení úspěšnosti systému. Instalace a práce se systémem i testy je detailně popsána v souboru `README.TXT`, obsaženém na přiloženém CD.

[1] Volně dostupná z https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/miditoolbox

[2] Volně dostupná z http://www.ee.columbia.edu/~csmit/matlab_midi.html

[3] Ve zkušební verzi dostupný z http://www.pistonsoft.com/midi2mp3.html

[4] Volně dostupný z http://audacity.sourceforge.net

[5] Ve zkušební verzi dostupný z http://www.image-line.com/downloads/flstudiodownload.html

Předpoklady Úspěšnost konverze z WAV do MIDI je závislá mimo jiné na kvalitě vstupního signálu, z nějž je vhodné odstranit šum. Tímto problémem se však systém WAV2MIDI nezabývá a předpokládá, že přijímá již předzpracované nahrávky. Rovněž předpokládá, že přijímá signál se stejnou vzorkovací frekvencí s níž byly trénovány frekvenční báze (v současné implementaci je přípustná pouze hodnota 11 kHz.)

Inicializace Jak bylo vysvětleno v kapitole 3, při analýze signálu metodou PLCA je nutné nejprve provést inicializaci dat. Apriorní pravděpodobnosti $P(z)$ jsou inicializovány náhodnými hodnotami (za pomoci funkce `rand()`) a časové báze $P(t|z)$ jsou inicializovány konstantní hodnotou (tedy všechny časové rámce mají stejnou pravděpodobnost). Frekvenční báze $P(f|z)$ jsou již k dispozici díky předchozímu trénování.

Frekvenční báze Natrénované frekvenční báze jsou uloženy jako sloupcové vektory v matici o 88 sloupcích (počet kláves piana) a 513 řádcích (počet frekvenčních pásem daných volbou okna o délce 1024 vzorků). Matice uložená ve formátu MAT-file[6] je přiložena ke zdrojovým kódům (viz A).

Šum Po odstranění šumu mohou nahrávky obsahovat pasáže s nulovou akustickou energií. Při aplikaci STFT nad takovýmto signálem se ve spektrogramu objevují nulové hodnoty. To je však nepřípustné, neboť se hodnoty jednotlivých buněk matice spektrogramu při výpočtu PLCA objevují ve jmenovateli. Proto je vstupního signálu doplněn o Gaussovský bílý šum.

Doba zpracování Na stroji s procesorem Intel Core i5-450M (3M cache, 2,40 GHz), 4GB operační pamětí, nainstalovaným 64bitovým OS Windows 8 a MATLABem verze R2011a trvá převod jedné WAV nahrávky přibližně 1,8násobek její délky.

4.3 Volba parametrů systému

Vlastnosti systému WAV2MIDI lze ovlivnit výběrem hodnot několika parametrů. Z hlediska úspěšnosti rozpoznání je důležité časové a frekvenční rozlišení systému, jež závisí na parametrech metody STFT, počet kroků metody PLCA a výběr vhodného hodnoty pro prahování a post-processing. Přehled hodnot všech parametrů je uveden v tabulce 4.1.

Parametr	**Hodnota**
vzorkovací frekvence	11025 Hz
délka okna	1024
přesah okna	50 %
počet kroků PLCA	75
práh	15
dkt (post-processing)	1

Tabulka 4.1: Souhrn parametrů systému WAV2MIDI.

[6] Formát používaný programovacím jazykem MATLAB pro trvalé uložení proměnných.

STFT Frekvence nejvyššího uvažovaného tónu *C8* má základní frekvenci 4186 Hz. S přihlédnutím k Nyquistovu teorému je tak nutné využít minimální vzorkovací frekvenci $F_s = 2f_{C_8} = 8372$ Hz. Nejbližší vyšší používaná hodnota činí 11025 HZ. Při výběru vhodných parametrů dále vycházím z faktu, že průměrný pianista nezahraje více, než 10 úhozů za sekundu. Pro časové rozlišení by tak měl dostačovat interval 100 ms, tedy okno délky $N = F_s/10 \doteq 1103$ vzorků. Nejbližší běžně používaná hodnota činí 1024. Pro zlepšení časového rozlišení je navíc při posouvání okna využito 50% přesahu posouvajícího se okna Při volbě okna této délky dosahuje systém frekvenčního rozlišení o hodnotě $F_s/N \doteq 10,8$ Hz. To by se mohlo zdát nedostačující vzhledem k faktu, že frekvenční rozdíl mezi dvěma nejnižšími tóny klaviatury, *A0* a *Ais0*, činí $1,635$ Hz. S natrénovanými frekvenčními bázemi je však metoda PLCA schopná i takto nízké tóny bezpečně rozlišit.

PLCA Během testování bylo experimentálně ověřeno, že po proběhnutí zhruba 75 iterací algoritmu metody PLCA se již hodnota KL divergence hodnotící rozdíl mezi referenčním a rekonstruovaným spektrogramem mění tak nepatrně, že nemá smysl nechat algoritmus déle běžet.

Prahování a post-processing Hodnota prahu a hodnota `dkt` pro post-processing zásadně ovlivňuje úspěšnost rozpoznání. Jejich optimální kombinace se pro každou nahrávku liší, nicméně je pro systém WAV2MIDI nutné stanovit konkrétní výchozí hodnoty. Nejvhodnější hodnoty byly určeny z výsledků testů provedených na sadě syntetizovaných nahrávek z hudební databáze Classical Piano Midi Page (viz kapitola 5). Pro každou z nahrávek byla nalezena taková kombinace hodnot prahu a `dkt`, aby byl výsledek porovnání odhadnuté a referenční piano-roll co nejlepší. Průměr těchto hodnot tak tvoří výchozí hodnoty pro systém WAV2MIDI.

Kapitola 5

Testování

Aby bylo možné otestovat úspěšnost systému WAV2MIDI, je nejprve nutné specifikovat, co lze považovat za "úspěšné" rozpoznání. Odborná veřejnost zabývající se problematikou přepisu polyfonních záznamů piana doposud nestanovila standardizované metriky, přesto lze z některých vědeckých článků jisté techniky vysledovat a použít. Pro zevrubné otestování byly zvoleny dvě metriky, jež se liší způsobem porovnání referenčních dat a dat produkovaných systémem (dále odhadnutých dat). Návrhu obou metrik se věnuje sekce 5.1.

Úspěšnost rozpoznání jednotlivých tónů závisí na mnoha faktorech, mezi nimiž lze zdůraznit především rozdílný zvuk různých pian, míru polyfonie, míru variability hlasitosti, nejkratší délka tónu nebo využití sustain pedálu[1]. Návrhem testů reagujících na tyto problémy se zabývá sekce 5.2.

5.1 Metriky

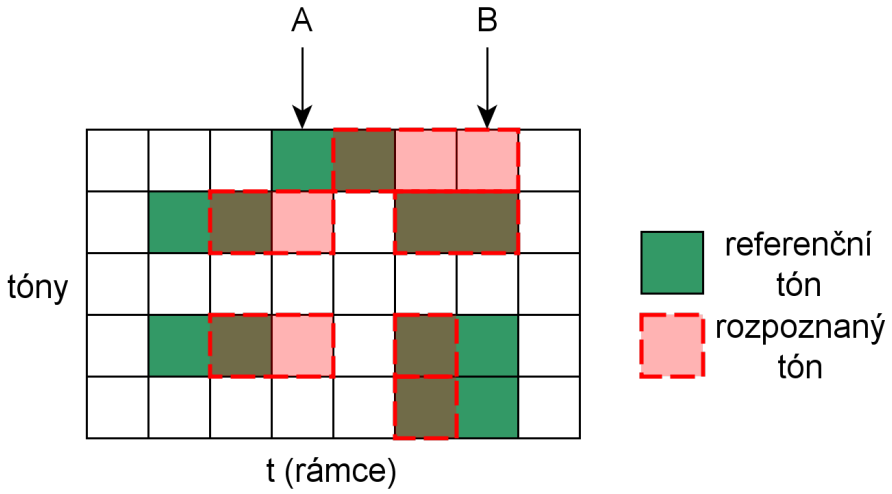
Pro ohodnocení systému WAV2MIDI přebírám tzv. **frame-level metriku** a tzv. **note-level metriku**.

5.1.1 Frame-level

První z metrik lze kategorizovat jako frame-level[2]. Při porovnání referenčních a systémových dat zcela ignoruje informaci o notách přítomných v nahrávce a porovnání provádí po jednotlivých časových rámcích, neboť byla navržena institucí NIST[3] pro účely testování řečových promluv. Článek [7] zabývající se analýzou polyfonních nahrávek piana tuto metriku přesto uplatňuje, tudíž ji pro účely porovnání výsledků s existujícími systémy také využiji.

Protože metrika vyhodnocuje data po časových rámcích, implementačně je prováděna nad piano-roll maticí. Jejím výsledkem jsou čtyři hodnoty E_{tot}, E_{subs}, E_{miss} a E_{fa} značící chybová skóre pro jednotlivá kritéria. Metrika pracuje s funkcemi $N_{ref}(t)$ a $N_{sys}(t)$, jejichž funkční hodnoty pro časový rámec t po řadě odpovídají počtu znějících tónů v referenčním a odhadnutém piano-roll, a s funkcí $N_{corr}(t)$, jež pro dané t vrací průnik mezi $N_{ref}(t)$ a $N_{sys}(t)$. Časový rámec představuje jeden sloupec v piano-roll, časový rámec konkrétního tónu budeme dále značit pojmem *buňka*. Princip metriky demonstruje obrázek 5.1 a následující popis.

[1] Běžně pravý pedál piana, který zamezuje dopadu dusítek zpět na struny a prodlužuje tak délku tónu.
[2] Lze přeložit jako úroveň rámců.
[3] National Institute of Standards and Technology



Obrázek 5.1: Demonstrace frame-level metriky nad piano-roll o osmi časových rámcích a 5 tónech. Zaznačeny jsou dva časové rámce A a B, k nimž se vztahuje vysvětlující text.

Celková chyba E_{tot} značí celkovou chybu reprezentovanou sumou chybějících, přebývajících nebo zaměněných buněk normalizovanou sumou referenčních buněk přes všechny časové rámce. Vypočítat ji lze pomocí vztahu

$$E_{tot} = \frac{\sum_{t=1}^{T} max(N_{ref}(t), N_{sys}(t)) - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)}. \tag{5.1}$$

Záměna E_{subs} značí počet takových buněk v odhadnutém piano-roll, které nemají průnik s buňkami v referenčním piano-roll, avšak maximálně tolika, kolik referenčních buněk nebylo rozpoznáno. Další nadbytečné buňky jsou již považovány za falešný poplach. Pro E_{subs} platí vztah

$$E_{subs} = \frac{\sum_{t=1}^{T} min(N_{ref}(t), N_{sys}(t)) - N_{corr}(t)}{\sum_{t=1}^{T} N_{ref}(t)}. \tag{5.2}$$

Minutí E_{miss} značí počet nerozpoznaných referenčních buněk po odečtení buněk považovaných za chybu záměny. E_{miss} popisuje vztah

$$E_{miss} = \frac{\sum_{t=1}^{T} max(0, N_{ref}(t) - N_{sys}(t))}{\sum_{t=1}^{T} N_{ref}(t)}. \tag{5.3}$$

Falešný poplach E_{fa} nakonec značí počet nadbytečných odhadnutých buněk, které nespadají pod chybu záměny. E_{fa} popisuje vztah

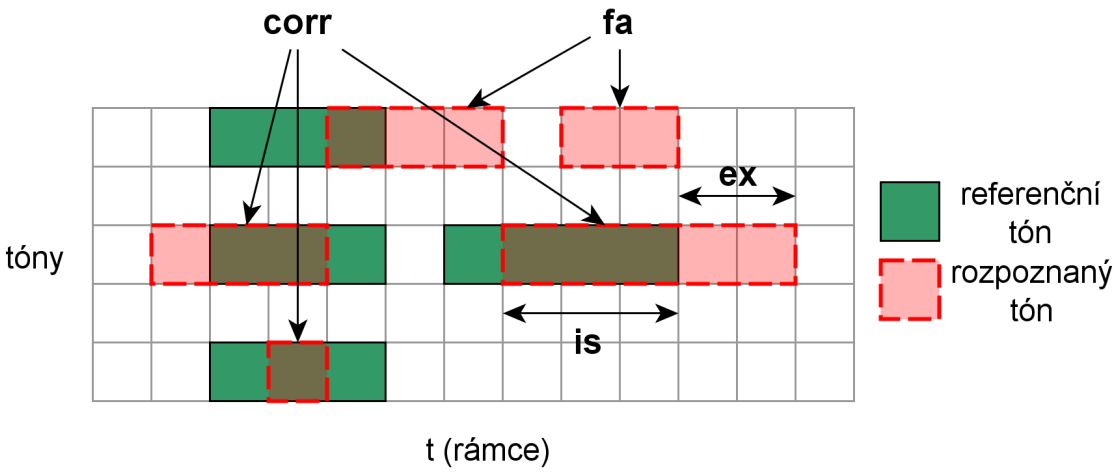
$$E_{fa} = \frac{\sum_{t=1}^{T} max(0, N_{sys}(t) - N_{ref}(t))}{\sum_{t=1}^{T} N_{ref}(t)}. \tag{5.4}$$

Pomocí obrázku 5.1 lze demonstrovat výpočet jednotlivých výsledků. V časovém **rámci A** došlo k situaci, kdy jediná referenční buňka nebyla rozpoznána a současně byly nesprávně odhadnuty dvě jiné buňky. Z pohledu metriky zde došlo k jedné substituci, žádnému minutí a jednomu falešnému poplachu. Jiná situace nastala v časovém **rámci B**, kde byla korektně rozpoznána pouze jedna ze tří referenčních buněk a současně byla nesprávně odhadnuta jedna další buňka. Z pohledu metriky došlo k jedné záměně, jednomu minutí a žádnému falešnému poplachu.

Nutno zmínit, že výsledné hodnoty reprezentují číselné skóre, nikoliv pravděpodobnost, neboť může snadno překročit hodnotu 100 %.

5.1.2 Note-level

Druhá metrika hodnotí úspěšnost rozpoznání z hlediska výskytu tónů (a jejich případného časového posunutí), proto ji lze analogicky označit jako note-level. Pracuje se čtveřicí výsledků *corr* (correct), *fa* (false alarm), *is* (intersection) a *ex* (excess). Princip výpočtu hodnoty *corr* přebírám z článku [6] (autoři ji značí jako *Acc*), výpočet zbývajících tří navrhuji jako rozšíření pro zevrubnější vyhodnocení. Jejich významu je patrný z obrázku 5.2 a z následujícího popisu.



Obrázek 5.2: Demonstrace note-level metriky s na piano-roll matici s časovým rámcem o délce 0.1 s. Zaznačeny jsou různé případy správně rozpoznaného tónu (*corr*), falešného poplachu (*fa*) a na jednom ze správně rozpoznaných tónů je ukázán průnik(*is*) a přesah(*ex*).

Zásah Hodnota *corr* značí poměr mezi počtem správně rozpoznaných tónů a počtem všech referenčních tónů. Za správně rozpoznanou notu se považuje každá taková odhadnutá nota, která má s odpovídající referenční notou nenulový časový průnik a současně se její počátek nachází v rozmezí maximálně 0.1 s od počátku referenční noty. Pokud by došlo k průniku odhadnuté noty s více referenčními, za zásah se považuje pouze první z nich.

Falešný poplach Hodnota *fa* udává poměr počtu chybně odhadnutých not (takových, jež mají nulový průnik s referenčními notami) ku počtu referenčních not.

Průnik Hodnota *is* značí poměr délky průniku správně odhadnutého a referenčního tónu ku délce referenčního tónu.

Přesah Hodnota *ex* nakonec značí poměr délky přesahu správně odhadnutého tónu ku délce referenčního tónu.

Pro hodnocení úspěšnosti mají největší význam položky *corr* a *fa*. Nelze zohlednit pouze jednu z nich, neboť při úpravě parametrů systému WAV2MIDI (především prahu a maximální délky zanedbaných krátkých tónů, viz sekce 4) jejich hodnoty rostou či klesají současně. Tedy pakliže je například snížena hodnota prahu, systém sice odhadne více tónů, čímž stoupne hodnota *corr*, avšak rovněž vznikne více chybně odhadnutých tónů a stoupne hodnota *fa*.

5.2 Testy

Pro ověření úspěšnosti systému byla navržena dvojice testů. Jak bude vysvětleno níže audio nahrávky jsou získávány jednak syntézou z MIDI a jednak přímým záznamem zvuku piana, přičemž byla v obou případech použita vzorkovací frekvence $Fs = 11025$ Hz. Na výsledky má základní vliv několik faktorů, jmenujme ty nejdůležitější, jimž byly testy uzpůsobeny:

Zvuk piana
: Banka frekvenčních bází byla extrahována z jednoho konkrétního piana. Výrazně odlišný zvuk piana použitého pro testování znesnadňuje rozpoznání tónů.

Sustain pedál
: Sustain pedál prodlužuje délku tónu. To však v MIDI souborech není promítnuto v délkách tónů nýbrž v samostatné MIDI zprávě. Při zachování sustain pedálu by výsledný syntetizovaný zvuk neodpovídal piano-roll matici získané z MIDI souboru a nebylo by možné úspěšnost systému řádně vyhodnotit.

Dynamika
: Změna hlasitosti v průběhu skladby představuje problém, neboť je pro celou skladbu volen jediný práh, nad který se příliš slabě znějící tóny nemusí dostat.

Polyfonie
: Více současně znějících tónů vyústí ve složitější spektra a překrývající se základní frekvence a alikvóty. To omezuje schopnost rozpoznání tónů.

Délka tónu
: Přítomnost příliš krátkých tónů negativně ovlivňuje úspěšnost rozpoznání, neboť mohou být ve výsledném MIDI souboru vlivem post-processingu a nedostatečného časového rozlišení systému zanedbány.

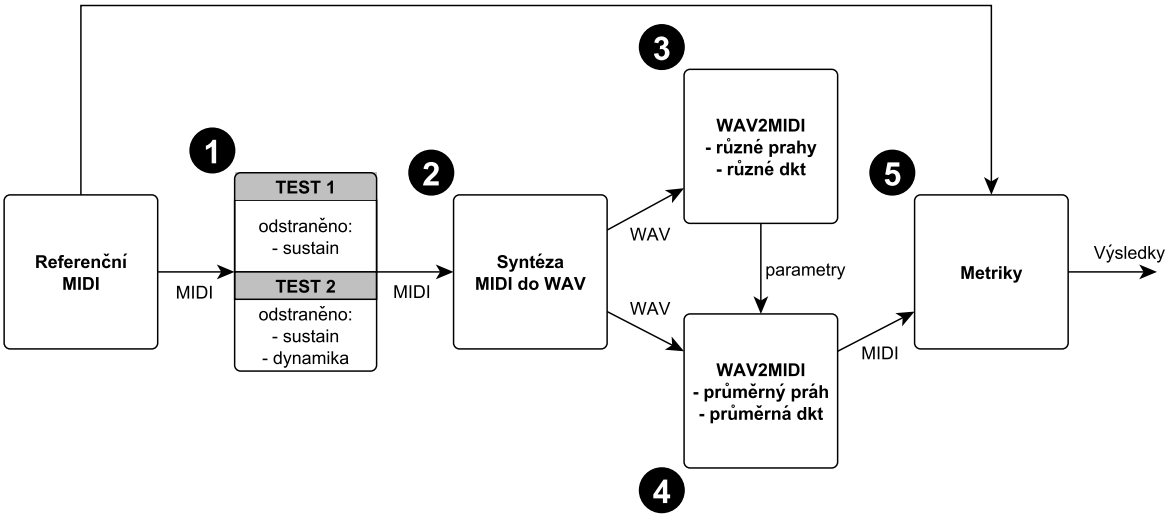
5.2.1 Syntetizované nahrávky

Pro první z testů byla využita volně přístupná hudební databáze Classical Piano Midi německého autora Bernda Kruegera[4], jež skýtá okolo 300 přepisů kompozicí skladatelů klasické klavírní hudby, jako jsou J.S.Bach, W.A.Mozart, L. van Beethoven a další, do formátu MIDI.

[4]Databáze je dostupná na http://www.piano-midi.de.

Dle tvrzení autora databáze byly nahrávky tvořeny ručně pomocí MIDI sequenceru a kromě samotných not nesou také informaci o síle úhozů, dynamice, změně tempa či použití sustain pedálu. To pro testování představuje problém (generovaný zvuk přesně neodpovídá délkám tónů určených midi zprávami Note on/Note off značícím začátek a konec znění tónu).

Celý proces přípravy souborů a testování je schematicky znázorněn na obrázku 5.3.



Obrázek 5.3: Schematické znázornění procesu testování syntetizovaných nahrávek v pěti krocích.

1. Aby bylo možné zjistit, jaký vliv má změna dynamiky v průběhu skladby na úspěšnost rozpoznání, byl celý test proveden dvakrát pro MIDI soubory se zachovanou a odstraněnou dynamikou. V obou případech je odstraněn výskyt sustain pedálu.

2. MIDI soubory byly syntetizovány do audio formátu WAV. Tento krok snižuje úspěšnost rozpoznání, neboť zvuk jednotlivých tónů produkovaný syntézou jistě neodpovídá zvuku tónů piana, jež bylo použito pro tvorbu banky frekvenčních bází metody PLCA.

3. Úkolem tohoto kroku je získat nejvhodnější hodnotu prahu a `dkt` pro všechny testované skladby. Každá nahrávka je analyzována systémem WAV2MIDI a je vytvořena množina odhadnutých MIDI souborů při použití různých kombinacích hodnot prahu a dkt. Každý odhadnutý MIDI soubor je porovnán s referenčním MIDI souborem a na základě výsledků metrik je vybrána nejlepší hodnota prahu a `dkt`. Seznam nejlepších hodnot prahů a `dkt` všech 274 skladeb je zprůměrován, čímž je získána jediná hodnota pro práh a `dkt`.

4. Syntetizované WAV soubory produkované krokem 2 jsou opět analyzovány systémem WAV2MIDI, avšak tentokrát při volbě jednotného prahu a `dkt` obdrženého z kroku 3. Výstupem jsou odhadnuté MIDI soubory.

5. Odhadnuté MIDI soubory produkované krokem 4 jsou nakonec porovnány s referenčními MIDI soubory za pomocí frame-level a note-level metrik a jsou produkovány výsledky.

5.2.2 Reálné nahrávky

Druhý test již nepracuje se syntetizovaným audiem, nýbrž jsou vstupem testu nahrávky reálného hudebního nástroje. Pro pořízení nahrávek bylo vyžito stejné piano, které posloužilo pro tvorbu banky frekvenčních bází PLCA tak, aby se snížil negativní dopad rozdílných zvuků pian na úspěšnost rozpoznání.

Nahrané byly tři skladby a to v monofonní i polyfonní variantě. K těmto skladbám byly pro účely porovnání ručně vytvořeny referenční MIDI soubory. Pro omezení vlivu přítomnosti sustain pedálu a změny dynamiky na výsledky testů byly nahrávky pořízeny bez sustain pedálu a se snahou o úhozy ve stejné dynamice.

Kapitola 6

Zhodnocení dosažených výsledků

Tato kapitola se věnuje výsledkům testování úspěšnosti systému WAV2MIDI. Sekce 6.1 rozebírá vyhodnocení metrik nad syntetizovanými nahrávkami a současně prezentuje porovnání výsledků s jinými již existujícími systémy. Sekce 6.2 popisuje výsledky testů s nahrávkami reálného piana.

6.1 Syntetizované nahrávky

Pro testování bylo z hudební databáze Classical Piano Midi náhodně vybráno 11 skladatelů a pro každého 8–10 skladeb v celkovém počtu 100 MIDI souborů. Jak bylo vysvětleno v sekci 5.2.1, pro testování systému byly MIDI soubory syntetizovány do audio formátu WAV a to jak ve variantě s měnící se dynamikou tak ve variantě s dynamikou konstantní.

Pro větší přehlednost jsou výsledky testování kategorizovány podle skladatelů, přičemž je pro každý test uveden celkový výsledek. Výsledky note-level metriky obdržené z testů provedených nad nahrávkami s měnící se hlasitostí, respektive s konstantní hlasitostí se potom nacházejí v tabulce 6.1, respektive tabulce 6.2.

Z obou tabulek lze vypozorovat, že měnící se hlasitost v rámci skladby má na úspěšnost rozpoznání zásadní vliv. V případě nahrávek s konstantní hlasitostí se procentuální hodnota správně rozpoznaných tónů *corr* v kategoriích některých skladatelů zvýšila o více než 10 % a současně nepatrně klesla hodnota *fa*. Díky přizpůsobení hlasitosti všech nahrávek byl systém WAV2MIDI schopen správně rozpoznat téměř 84 % tónů při 11,77% falešném poplachu, což by byl při srovnání s jinými systémy (viz tabulku 6.4) velmi dobrý výsledek, kvalitu systému však podle těchto hodnot úspěšnosti rozpoznání nelze hodnotit. V reálném prostředí živý hráč nikdy nezahraje všechny tóny zcela stejně silně a ani to není jeho záměrem. Test pro nahrávky s konstantní hlasitostí byl tedy proveden pouze pro demonstraci vlivu tohoto parametru na úspěšnost rozpoznání a následující analýza výsledků staví na první testovací sadě nahrávek s měnící se hlasitostí tónů.

Nejhorší výsledky s hodnotami $corr = 64,21$ % a $fa = 10,44$ % vykazují nahrávky skladatele Liszta. Nejlepších výsledků naopak dosáhly nahrávky skladatele Mozarta s hodnotami $corr = 85,33$ % a $fa = 15,38$ %. Proč k tomuto došlo? Jak již bylo uvedeno v sekci 5.2, dalšími významnými faktory ovlivňujícími úspěšnost rozpoznání jsou míra polyfonie a průměrná délka znějícího tónu. Tyto charakteristiky jsou pro jednotlivé skladatele uvedeny v tabulce 6.3.

Nahrávky skladatele Liszta dosahují výrazně vyšší průměrné polyfonie a nižší průměrné délky tónu ve srovnání s nahrávkami skladatele Mozarta. Demonstrujme vliv obou faktorů

Skladatel	***% corr***	***% fa***	***% is***	***% ex***
Albéniz	77,30	7,72	77,41	54,24
Bach	84,15	9,98	87,99	46,45
Beethoven	71,36	14,34	80,95	74,76
Chopin	67,18	5,64	83,29	63,93
Debussy	81,17	16,20	82,70	57,46
Grieg	69,54	16,98	84,26	107,87
Haydn	79,34	11,07	88,19	95,96
Liszt	64,21	10,44	83,94	85,00
Mozart	85,33	15,38	84,47	69,24
Schubert	67,50	18,58	74,02	55,65
Tchaikovsky	74,86	13,17	74,11	36,09
CELKEM	73,66	12,77	81,59	69,74

Tabulka 6.1: Výsledky note-level metriky pro nahrávky s měnící se hlasitostí.

Skladatel	***% corr***	***% fa***	***% is***	***% ex***
Albéniz	86,22	7,38	66,62	36,58
Bach	88,83	6,85	77,68	25,27
Beethoven	83,13	13,85	73,13	41,31
Chopin	79,67	5,02	68,41	33,97
Debussy	90,64	14,19	76,70	40,62
Grieg	83,58	14,77	76,52	60,76
Haydn	86,64	9,41	83,20	62,49
Liszt	73,24	10,54	71,10	51,84
Mozart	92,22	13,90	79,84	38,57
Schubert	80,66	18,17	65,04	34,15
Tchaikovsky	82,87	13,04	60,18	20,23
CELKEM	83,86	11,77	72,23	41,64

Tabulka 6.2: Výsledky note-level metriky pro nahrávky s konstantní hlasitostí.

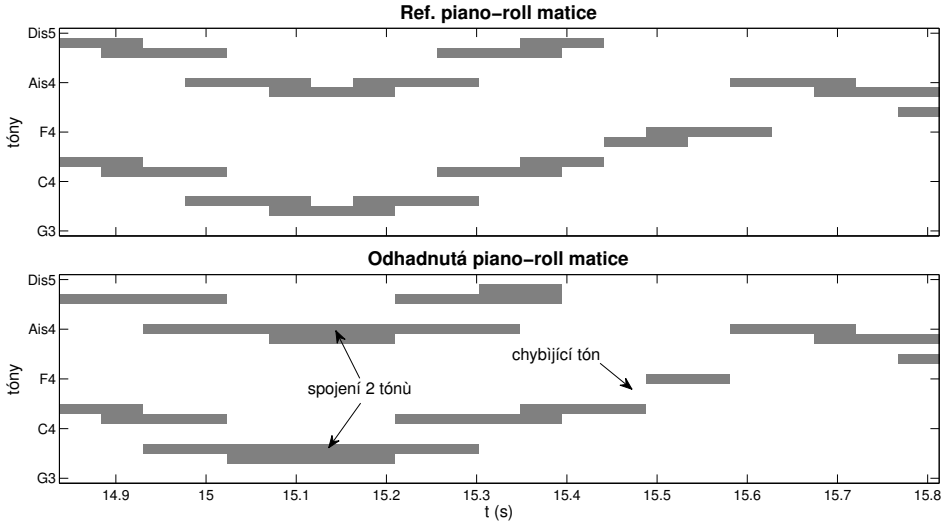
na úspěšnosti rozpoznání v nahrávce Mazzepa, jež mezi všemi nahrávkami skladatele Liszta dosáhla vůbec nejhoršího výsledku $corr = 53,90$ % s nicméně poměrně dobrým výsledkem $fa = 2,12$ %.

Obrázek 6.1 demonstruje problém výskytu příliš krátkých tónů, jejichž délka hraničí s rozlišovací schopností systému WAV2MIDI. Na 1s úseku piano-roll matice jsou zde patrné

Skladatel	Maximální polyfonie	Průměrná polyfonie	Průměrná délka tónu (s)
Albéniz	8	3,04	0,32
Bach	10	2,20	0,27
Beethoven	8	2,65	0,38
Chopin	8	2,72	0,22
Debussy	11	2,64	0,46
Grieg	6	2,23	0,33
Haydn	6	2,00	0,25
Liszt	9	2,71	0,23
Mozart	7	2,12	0,36
Schubert	9	3,27	0,45
Tchaikovsky	8	3,31	0,46

Tabulka 6.3: Charakteristiky nahrávek kategorizované dle skladatele.

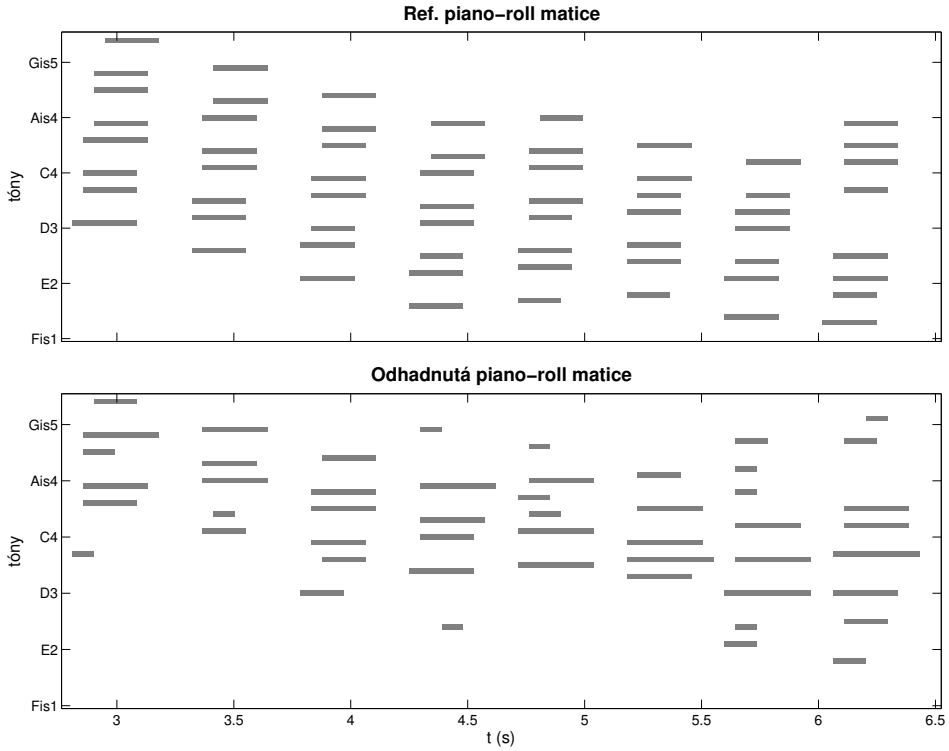
dva zdroje vzniklých chyb. Jeden výskyt tónu *E4* byl zanedbán vlivem post-processingu a dvojice tónů *Ais4*, *D4* vyskytující se opakovaně po sobě v příliš krátkých intervalech byla v obou případech systémem rozpoznána jako jediný dlouze znějící tón. Oba případy snižují hodnotu *corr*, avšak nemají žádný vliv na hodnotu *fa* (proto dosahuje v případě nahrávky Mazzepa *fa* velmi nízké hodnoty).



Obrázek 6.1: Srovnání 1s úseku referenční a odhadnuté piano-roll matice skladby Mazzepa skladatele Liszta, na němž je patrný chybný odhad tónu *Ais4* a zanedbání tónu *E4*.

Na obrázku 6.2 je potom patrný negativní dopad příliš vysoké polyfonie na úspěšnost rozpoznání. Ve zhruba 4s úseku skladby se vyskytlo osm akordů sestávajících ze 7–8 tónů.

v odhadnuté piano-roll matici je znát, že byl systém WAV2MIDI schopen každý akord rozpoznat pouze částečně (další snížení hodnoty *corr*), zatímco u některých z nich dochází ke špatnému odhadnutí tónů, které se v referenční nahrávce nevyskytly (zvýšení hodnoty *fa*).



Obrázek 6.2: Srovnání necelých 4 s referenční a odhadnuté piano-roll matice skladby Mazzepa skladatele Liszta.

Nutno zmínit, že proměnlivost hlasitosti, míra polyfonie a délka znějícího tónu nejsou jedinými faktory majícími vliv na úspěšnost rozpoznání. Lze si všimnout například faktu, že nahrávky skladatele Chopina vykazují ve srovnání s Lisztem nepatrně vyšší průměrnou polyfonii a nepatrně nižší průměrnou délku tónu, tudíž by měl systém dle stanovených předpokladů vykázat nižší úspěšnost, k čemuž však nedošlo. Důvodem mohou být pro nahrávky Liszta nevhodně stanovené parametry analýzy (pro práh a post-processing), jež byly jednotně odhadnuty ze všech testovaných nahrávek.

Jak značí výsledná hodnota *is*, správně odhadnuté tóny mají vůči referenčním tónům takovou časovou pozici a délku, že se s nimi téměř z 82 % protínají. Na vysokou hodnotu přesahu správně rozpoznaných *ex* o velikosti téměř 70 % má kromě nepřesností při rozpoznání vliv také fakt, že se syntéza z MIDI do WAV snaží co nejvěrněji simulovat reálný zvuk piana a tónům přidává i dozvuk (reverb), kvůli kterému zvuk nahrávek zcela přesně neodpovídá zápisu v MIDI.

V závěru této sekce uvádím srovnání úspěšnosti systému s ostatními již existujícími systémy [6]. Autoři pro testování využili 25 náhodně vybraných MIDI souborů z databáze Classical Piano Midi, přičemž pro porovnání používají hodnoty *corr*, E_{tot}, E_{subs}, E_{miss} a E_{fa}.

Systém navržený autory Polinerem a Ellisem používá pro klasifikaci tónů metodu SVM[1] supervizovaně trénovanou nad daty se správně vyznačenými tóny a pro post-processing HMM[2]. Systém autorů Ryynänena a Klapuriho výskyt hudebních tónů odhaduje za pomocí HMM z frekvenčního spektra podle výskytu špiček, jež považuje za možné základní frekvence nebo alikvóty tónů. Systém autora Marolta pak pro rozpoznání používá neuronové sítě. V tabulce 6.4 je uvedeno srovnání všech tří zmíněných systémů se systémem WAV2MIDI.

Systém	***% corr***	% $\boldsymbol{E_{tot}}$	% $\boldsymbol{E_{subs}}$	% $\boldsymbol{E_{miss}}$	% $\boldsymbol{E_{fa}}$
WAV2MIDI	73,7	60,2	8,6	34,1	17,5
Poliner a Ellis	67,7	34,2	5,3	12,1	16,8
Ryynänen a Klapuri	46,6	52,3	15,0	26,2	11,1
Marolt	36,9	65,7	19,3	30,9	15,4

Tabulka 6.4: Srovnání systému WAV2MIDI s jinými již existujícími systémy.

Nutno zmínit, že porovnání není zcela přesné. Na výsledky má vliv mj. fakt, že autoři Poliner a Ellis použili pro testování svého i zbývajících dvou systémů 25 nahrávek (ve srovnání se 100 nahrávkami použitými pro testování WAV2MIDI) a syntézu z WAV do MIDI prováděli za pomocí softwarového nástroje Apple iTunes, který pro syntézu využívá v porovnání s nástrojem Direct MIDI to MP3 Converter rozdílnou banku zvuků. Rovněž využili vzorkovací frekvenci 8 kHz (oproti 11 kHz v případě systému WAV2MIDI). Výsledky navíc uvádějí souhrnně pro testovací sadu syntetizovaných nahrávek společně s vlastními reálnými nahrávkami piana. Tyto však nebyly pro testování systému WAV2MIDI k dispozici. Dále nelze opomenout fakt, že systémy autorů Ryynänena a Klapuriho a autora Marolta byly navrženy pro automatický přepis libovolného hudebního nástroje a nebyly tak přizpůsobeny přímo pianu.

Z pohledu hodnoty *corr* note-level metriky systém WAV2MIDI dle očekávání dosahuje lepšího skóre, než oba systémy pro přepis obecné hudby (ne pouze piana). Překvapivě však dosahuje rovněž lepšího výsledku než systém Polinera a Ellise. Za touto skutečností však jednoznačně stojí fakt, že při testování systému WAV2MIDI byl z referenčních MIDI nahrávek odstraněn sustain pedál. Jednotlivé výsledky frame-level metriky však již odhalily nedostatky systém WAV2MIDI. Pro hodnoty E_{miss} a E_{fa} dosahuje ze všech systémů nejhorších výsledků, což koresponduje s vysokou hodnotu přesahu *ex* note-level metriky. Celková chyba E_{tot} pak systém řadí na třetí místo. Nepoměr výsledků note-level a frame-level metriky lze interpretovat tak, že tón sice rozpoznán byl, ale svůj referenční vzor délkou výrazně přesahuje (hodnota *ex*), což odpovídá problému nízkého časového rozlišení znázorněného na obrázku 6.1. Za neuspokojivé výsledky frame-level metriky lze rovněž vinit využití příliš triviálního post-processingu.

6.2 Reálné nahrávky

Pro testovací účely byl ručně vytvořeny tři polyfonní a tři monofonní MIDI zápisy skladeb Comptine d'un autre ete skladatele Yanna Tiersena, River Flows in You skladatele Yiruma

[1] Support Vector Machine
[2] Hidden Markov Model

a klavírní aranž skladby I Want It That Way hudební skupiny Backstreet Boys.

Následně byly pořízeny audio záznamy reálného piana (Korg SP-250), na nějž hrál reálný hráč s tím, že se snažil dodržovat tempo jednotlivých not tak, aby nahrávky co nejlépe odpovídaly referenčním MIDI souborům. Nahrávky pořízeny do nekomprimovaného formátu WAV při vzorkovací frekvenci 11025 Hz a před vlastním testování z nich byl odstraněn šum. Pro další porovnání byly rovněž vytvořeny nahrávky syntetizované z referenčních MIDI souborů (podobně jako v předchozích testech). Všechny nahrávky i referenční MIDI soubory jsou součástí přiloženého CD. Výsledky obou testů jsou uvedeny v tabulkách 6.5 a 6.6.

Skladba	*% corr*	*% fa*	*% is*	*% ex*
Tiersen mono	98,93	14,44	90,56	76,76
Yiruma mono	92,97	8,59	92,31	62,91
BB mono	94,81	0,94	93,35	74,86
CELKEM	95,57	7,99	92,07	71,51
Tiersen poly	68,66	5,78	75,61	28,10
Yiruma poly	65,72	9,39	73,54	23,92
BB poly	84,21	3,22	85,16	36,02
CELKEM	72,86	6,13	78,10	29,35

Tabulka 6.5: Výsledky testů provedených nad nahrávkami pocházejícími ze záznamu zvuku reálného piana, na nějž hrál živý hráč.

Skladba	*% corr*	*% fa*	*% is*	*% ex*
Tiersen mono	89,30	6,42	91,56	95,82
Yiruma mono	92,19	19,19	91,37	51,31
BB mono	94,34	1,42	94,90	82,40
CELKEM	91,94	9,01	92,61	76,51
Tiersen poly	77,05	2,24	73,93	24,23
Yiruma poly	69,87	1,09	74,23	24,40
BB poly	93,57	4,39	82,50	31,71
CELKEM	80,16	2,57	76,89	26,78

Tabulka 6.6: Výsledky testů provedených nad nahrávkami vzniklými syntézou referenčních MIDI do WAV.

Celkové výsledky pro monofonní nahrávky, kdy hodnota *corr* přesahuje v obou případech 90 %, dotvrzují fakt, že míra polyfonie má na úspěšnost rozpoznání výrazný vliv. V případě polyfonních nahrávek jsou výsledky srovnatelné s výsledky testů provedených nad databází Classical Piano Midi.

Ačkoliv byly reálné nahrávky pořízeny s využitím stejného piana, jež bylo použito pro

trénování metody PLCA, syntetizované nahrávky dosahují vyšší úspěšnosti. Na této skutečnosti se však do vysoké míry podílí měnící se hlasitost jednotlivých tónů v reálných nahrávkách, zatímco v syntetizovaných nahrávkách mají všechny tóny stejnou hlasitost. Reálné nahrávky referenčním MIDI zápisům přesně neodpovídají také časově, což je dáno nedokonalostí hry živého hráče případně akustikou místnosti.

Kapitola 7

Závěr

Závěr je věnován shrnutí odvedené práce, využitých metod a postupů i dosažených výsledků. Současně je předkládán návrh možných směrů dalšího vývoje systému WAV2MIDI.

7.1 Shrnutí

V práci jsem se věnoval studiu metody PLCA určené k časově-frekvenční dekompozici audio signálu a jejímu přizpůsobení pro účely rozpoznání tónů v polyfonních nahrávkách piana. Metoda PLCA byla supervizovaně natrénována nad nahrávkami tónů reálného piana a použita jako jádro navrženého systému WAV2MIDI, jenž slouží pro převod nahrávek z formátu WAV do symbolické reprezentace (piano-roll matice) a pro jejich výsledný export do formátu MIDI. Pro ověření úspěšnosti byla využita sada testovacích nahrávek z volně přístupné databáze Classical Piano Midi obsahující skladby autorů klasické klavírní hudby a rovněž byla pořízena sada nahrávek reálného piana.

Výsledky analýzy byly vyhodnoceny za pomoci dvojice rozdílných metrik, přičemž frame-level metriku zcela přejímám z literatury a note-level metriku doplňuji o možnost měření dalších vlastností, jako je průnik a přesah mezi správně rozpoznanými a referenčními tóny. Výsledky testů byly v závěru porovnány se třemi již existujícími systémy. Podle note-level metriky systém dosahuje nejlepšího skóre s hodnotou téměř 74 % správně rozpoznaných tónů, zatímco z pohledu frame-level metriky má systém WAV2MIDI oproti nejúspěšnějšímu systému (viz tabulka 6.4) téměř dvojnásobnou chybovost E_{tot}, což připisuji využití příliš triviálního post-processingu.

7.2 Možnosti dalšího vývoje

Analýza výsledků testování odhalila mnohé nedokonalosti systému, jež si tak žádají další práci. V tomto kontextu je možné opět zmínit nepříliš dobré výsledky frame-level metriky, jež poukazují na nedostatečné časové rozlišení systému i nedostatečnou sílu post-processingu. Vliv na neuspokojivé výsledky má pravděpodobně také způsob volby prahu, který nezohledňuje hustotu tónů v nahrávce ani jejich hlasitost, proto by bylo vhodné navrhnout vhodnější přístup.

Časově frekvenční analýzy s vyšším časovým rozlišením při zachování frekvenčního rozlišení lze dosáhnout například využitím metody Constant-Q Transform (dále CQT), jež zakládá na STFT, avšak pro každou oktávu využívá stejný počet frekvenčních pásem. Vyššího časového rozlišení zde dosahuje tím, že pro analýzu nižších oktáv signál podvzorkuje,

a může si tak dovolit využít kratší okna [8]. Pro post-processing a odhadnutí výsledné piano-roll matice by pak bylo vhodné uplatnit některou ze sofistikovanějších metod, jako jsou například HMM [2].

Literatura

[1] Bay, M.: *Methods for multiple pitch tracking and instrument separation from monaural polyphonic recordings.* Dizertační práce, University of Illinois at Urbana-Champaign, 2012.

[2] Benetos, E.; Dixon, S.: Multiple-f0 estimation and note tracking experience. *Music Information Retrieval Evaluation eXchange (MIREX)*, říjen 2011.

[3] Ellis, D.; Rosenthal, D.: Mid-level representations for Computational Auditory Scene Analysis. *Readings in Computational Auditory Scene Analysis*, 1995-08.

[4] Foo, S. W.; Wong, P. L.: Recognition of piano notes. *IEEE International Conference on Information, Communications and Signal Processing (1999:Singapore)*, 1999.

[5] Klapuri, A.; Davy, M.: *Signal Processing Methods for Music Transcription.* Springer, 2006, ISBN 0-387-30667-6.

[6] Poliner, G. E.; Ellis, D. P.: A Discriminative Model for Polyphonic Piano Transcripton. *EURASIP J. Advances in Signal Processing*, , č. 8, 2007-01: s. 154–162.

[7] Poliner, G. E.; Ellis, D. P.: Improving Generalization For Polyphonic Piano Transcription. *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007-10.

[8] Schorkhuber, C.; Klapuri, A.: Constant-Q Transform Toolbox for music processing. *7th Sound and Music Computing Conf.*, 07-2010.

[9] Smaragdis, P.; Raj, B.; Shashanka, M.: A Probabilistic Latent Variable Model for Acoustic Modeling. *Neural Information Processing Systems Workshop*, 2006-12.

[10] Smaragdis, P.; Raj, B.; Shashanka, M. V. S.: Sparse and shift-invariant feature extraction from non-negative data. In *ICASSP*, IEEE, 2008, ISBN 1-4244-1484-9, s. 2069–2072.

[11] Smith III J. O.: *Mathematics of the discrete Fourier transform (DFT) with audio applications second edition.* 2007, ISBN 978-0-9745607-4-8.

Příloha A

Obsah CD

Na přiloženém CD se nachází následující adresáře a soubory:

- **doc** - elektronická verze tohoto textu v PDF a zdrojové soubory v LaTeXovém formátu
- **data** - hudební data odkazovaná z textu práce
- **src** - zdrojové soubory systému WAV2MIDI včetně souboru MAT-file nesoucího natrénovanou matici frekvenčních bází, použitých knihoven a skriptů pro ukázku vyhodnocení testů
- **test** - hudební data použitá pro ukázku testů
- **README.txt** - textový soubor obsahující detailní informace o obsahu CD, instalaci systému, použití skriptů a práci s testy