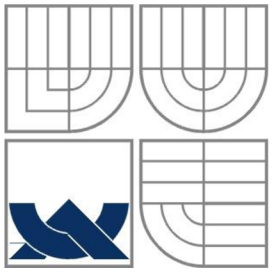
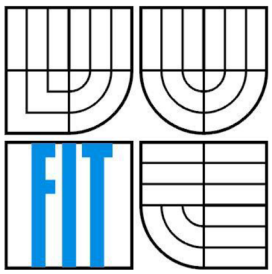


VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ROZPOZNÁVAČ HUDEBNÍHO STYLU Z MP3

MP3 MUSIC STYLE RECOGNITION

DIPLOMOVÁ PRÁCE
MASTER`S THESIS

AUTOR PRÁCE
AUTHOR

BC. DEUTSCHER MICHAEL

VEDOUCÍ PRÁCE
SUPERVISOR

ING. FRANTIŠEK GRÉZL, PH. D.

BRNO 2009

Abstrakt

Práce popisuje návrh rozpoznávače hudebních stylů. Stručně se zabývá digitalizací hudebních dat, způsobem jejich ukládání v počítačích. Dále zmiňuje příznaky používané pro klasifikaci včetně nástinu jejich extrakce. Stěžejní částí je pak srovnání úspěšnosti rozpoznání hudebních žánrů pomocí příznaků získaných přímo z hudebních dat v mp3 formátu a příznaků získaných klasickou analýzou.

Klíčová slova

Rozpoznávač hudebního stylu, hudební styl, mp3, extrakce příznaků, hudba

Abstract

This document describes the concept of music style recognition. It gives a quick reference to the digitalization of music data and storing music data in computers. It also mentions features used for music style recognition and their extraction. The main part of this document compares the successfulness of music genre recognition using features extracted directly from audio data in mp3 format and features extracted by usual analysis.

Keywords

Recognition of music style, music style, mp3, feature extraction, music

Citace

Deutscher Michael: Rozpoznávač hudebního stylu z mp3. Brno, 2009, diplomová práce, FIT VUT v Brně.

Rozpoznávač hudebního stylu z MP3

Prohlášení

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně pod vedením Ing. Františka Grézla, Ph. D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
Michael Deutscher
26.5.2009

Poděkování

Rád bych poděkoval za odbornou pomoc při tvorbě tohoto dokumentu Ing. Františku Grézlovi, Ph.D.

Obsah

1 Úvod

Vyhledávání, klasifikace a porovnávání je problematika provázející lidstvo již od počátku. Vyhledávání se dá nalézt prakticky kdekoliv. Samozřejmě se liší data, ve kterých vyhledáváme. Ovšem pojem data je pro obecnost vhodné brát dostatečně abstraktně. Nemusí se tedy jednat o data v počítači, ale o jakákoliv data jako jsou například inzeráty v novinách, jízdní řády vlaků apod.

Samozřejmě s příchodem digitalizace a jejím neustále probíhajícím rozmachem nabývá vyhledávání nového významu. Další fenomén současné doby, Internet, se na tom podílí nemalou měrou. K velké většině informací z něj získaných se člověk dostává právě vyhledáváním. Pokud je vyhledávání dostatečně efektivní, tak je to nejrychlejší způsob jak požadovanou informaci nalézt. Proč komplikovaně třídit maily v mailové schránce, když se v nich dá snadno vyhledávat?

V současnosti se však vyhledávání nesoustředí pouze na pole Internetu, ale Internet začíná být více chápán jako prostředník. K dispozici jsou nesčetné online databáze různorodých dat a stále vznikají další. Vyhledávání neprobíhá jen v textově reprezentovaných datech, stále většího významu nabývá vyhledávání v mluveném slovu i hudbě a multimediálních datech obecně.

Práce popisuje rozdíl v získávání příznaků pro rozpoznávání pomocí dvou metod. První metoda pracuje s klasickými hudebními daty v nekomprimované podobě a pomocí analýzy těchto dat získá příznaky. Druhá metoda pracuje přímo s formátem mp3 a snaží se získat příznaky uložené přímo mp3 souborech a tedy bez jakékoliv další analýzy. A právě na vyhledávání resp. třídění hudby do jednotlivých hudebních žánrů se snaží tento dokument demonstrovat využití získaných příznaků. Práce je z velké části experimentem a výsledky úspěšnosti rozpoznávání nelze dopředu odhadnout.

2 Hudební data

Hudební data v digitalizované podobě mohou být uchována v různých formátech. Pro lepší pochopení této tématiky je vhodné se seznámit s vlastnostmi zvuku a principy jeho vnímání lidským uchem.

2.1 Zvuk

2.1.1 Zvuk a frekvence

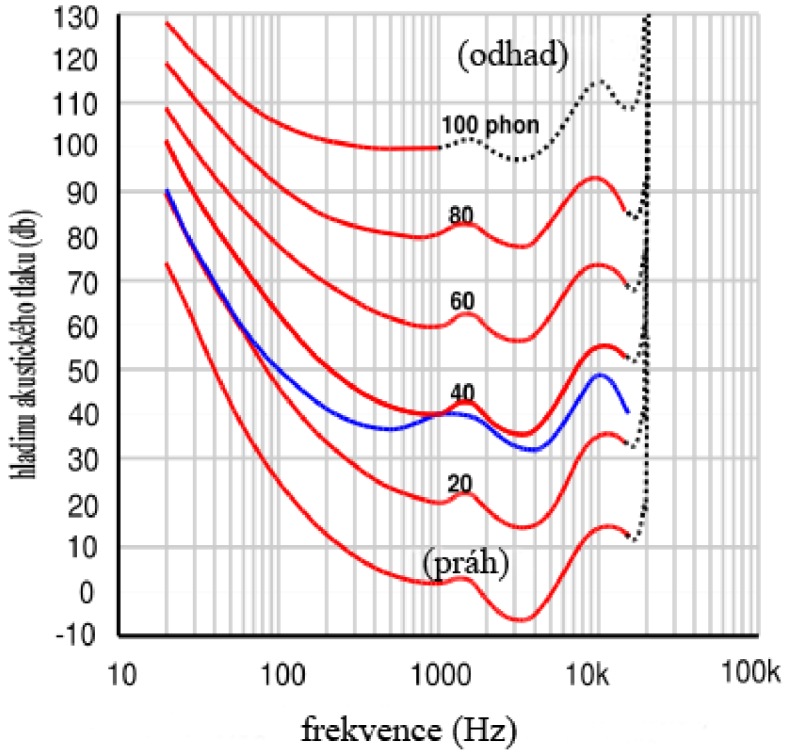
Zvuk je vlnění, vzniká tedy rozechvěním molekul prostředí (např. vzduchu). Zdrojem chvění může být reproduktor, lidské hlasové ústrojí apod. Rychlost šíření zvuku záleží jak na typu prostředí, tak na vlastnostech prostředí. Přibližná rychlost zvuku ve vzduchu je 340 m/s. Záleží ovšem jak na teplotě, tak hustotě a vlhkosti vzduchu.

Jednou z veličin, která popisuje zvuk je frekvence. Frekvence udává počet kmitů (vln) za sekundu. Zdroj zvuku rozechvívá molekuly prostředí určitou frekvencí. V různých prostředích dochází k různým útlumům i v závislosti na frekvenci. Je-li signál periodický, nazýváme jej tónem. Opakem tónu je šum, který tedy není periodický.

V praxi však málokdy slyšíme právě jednu frekvenci. Nejčastěji je základní tón doprovázen harmonickými složkami, tzv. alikvotními tóny, které určují barvu tónu a další vlastnosti. Proto je člověk schopen rozlišit zvuk dvou různých hudebních nástrojů, i když hrají identický tón. Liší se barvou a celkovou charakteristikou.

2.1.2 Lidské ucho a jeho fyzikální vlastnosti

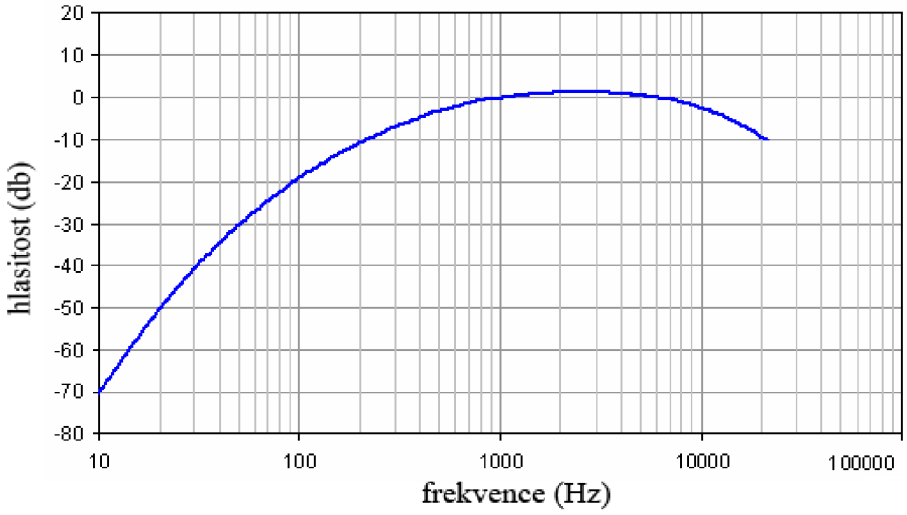
Základní frekvence, kterou dokáže lidské ucho detekovat, je udávána v rozsahu 20 Hz až 20 kHz. Tento rozsah je ale značně individuální a většinou menší. S rostoucím věkem se snižuje jak citlivost lidského ucha, tak rozsah v jakém je schopno zvuk detekovat. Při poslechu hudby hraje velkou roli také rozdílná vnímavost různých frekvencí (viz obrázek 1), což je zcela individuální vlastnost lidského sluchového ústrojí (nezávisí na věku).



Obrázek 1 - ISO 226:2003 (červeně) a Robinson-Dadson (modře) křivky konstantní hlasitosti

Výzkumy bylo zjištěno, že lidské ucho je nejcitlivější na frekvenci v rozsahu 2 až 5 kHz (jiné zdroje uvádějí 700 Hz – 6kHz). Zejména pod 500 Hz citlivost rapidně klesá. V souvislosti s tím je vhodné uvést, že základní rozsah lidské řeči odpovídá 500 Hz – 4 kHz. Lidské sluchové ústrojí je zaměřeno zejména na vnímání lidské řeči nebo zvuků ve stejném frekvenčním rozsahu.

Hlasitost (příklady různé hlasitosti zobrazuje tabulka 1) je tedy subjektivní veličina a závisí na individuálních dispozicích. Pro objektivní hodnocení zvuku se používá veličina intenzita. Pro převod hlasitosti na intenzitu se používají různé korelační křivky (viz obrázek 2).



Obrázek 2 - korelační A-křivka

Hlasitost (B/dB)	Zvuk
0	hranice slyšitelnosti
10	šelest listí, ticho na venkově
20	šum listí, knihovna, tikot hodinek
30	pouliční hluk v tichém předměstí
40	tlumený rozhovor
50	normální pouliční hluk, ruch v kanceláři
60	hlasitý (normální) rozhovor, ruch v davu
70	hluk na silně frekventovaných ulicích velkoměsta, vysavač
80	hluk v tunelech podzemních železnic, křik, symfonický orchestr
90	hluk motorových vozidel
100	maximální hluk motorky, pneumatická vrtačka
110	hlasité obráběcí stroje, rocková kapela
120	startující letadlo
130	hluk působící bolest

Tabulka 1 – příklady hlasitosti

Další vlastností lidského ucha, která plyne z předchozích poznatků, je vliv hlasitosti na vnímání frekvenčního spektra. Převážně basy (hluboké tóny) jsou málo vnímány při nízké hlasitosti. V praxi se to řeší tak, že basy se zeslabují méně než ostatní části spektra. Tato korekce se nazývá fyziologická korekce neboli loudness.

Experimenty byla zjištěna i vnímavost ucha na změnu hlasitosti. Získané hodnoty se liší výzkum od výzkumu, ale hodnoty se obecně pohybují v rozsahu 0,5 – 1 dB. Jinak řečeno, lidské ucho nepozná drobné odchylky v hlasitosti.

Vlastnost nazývaná časové maskování popisuje situaci, kdy po hlasitém tónu následuje stejný tón s nižší hlasitostí. V tomto případě dochází k potlačení vnímání tónu s nižší hlasitostí. Dokonce může dojít k potlačení, i pokud tichý tón hlasitému tónu předchází.

Podobně funguje i spektrální maskování. Pokud znějí dva frekvenčně blízké tóny současně, může lidské ucho potlačit vnímání jednoho z nich. Maximální úroveň maskovaného signálu je závislá na frekvenční vzdálenosti obou tónů a úrovni maskujícího signálu. Tento jev ovlivňuje také frekvence tónu, jelikož jak už bylo zmíněno výše, vnímavost lidského sluchu závisí i na frekvenci.

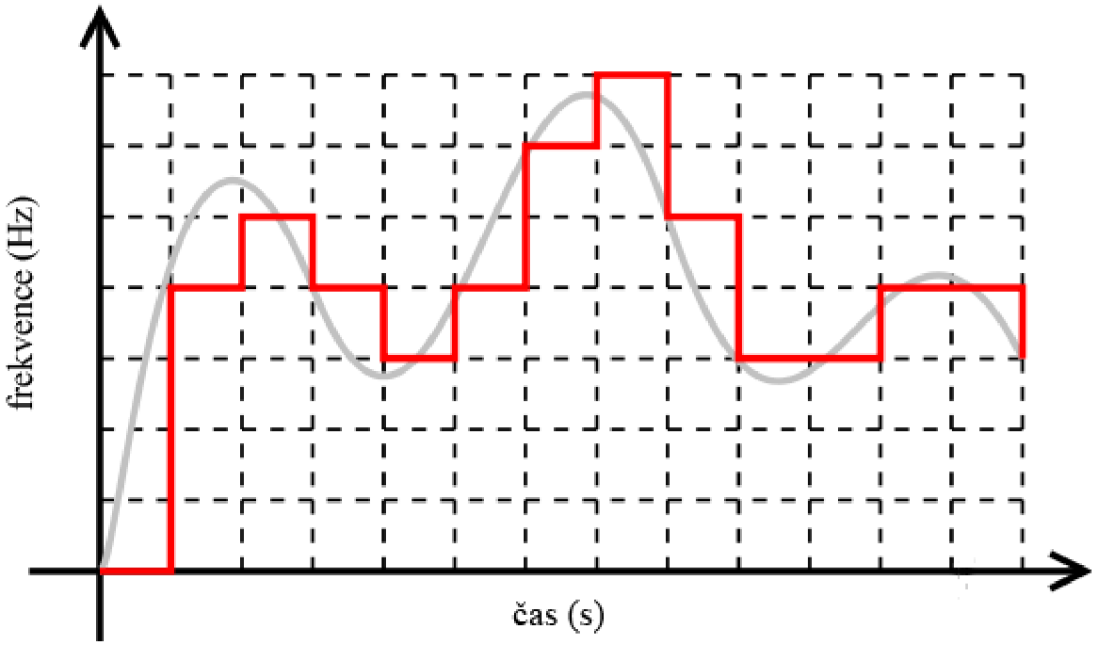
Vlastnosti lidského ucha se často používají při návrhu kompresních algoritmů používaných různými kodeky.

2.1.3 Digitální zvuk

V minulých kapitolách byl popisován zvuk analogový. Analogový zvuk je spojitý signál, jehož převedením do digitální podoby vzniká nespojitý signál. V digitální reprezentaci se hodnoty vyjadřují čísly a jejich počet je konečný. Analogově-digitální převodník převádí analogový signál na určitý počet hodnot (hladin), tento proces se nazývá kvantizace. Protože počet hladin je konečný, převodník vždy vybírá tu nejbližší hladinu k aktuální hodnotě vstupního analogového signálu. V praxi se nejčastěji používá 65 536 (tzn. 2 byty, resp. 16 bitů) hladin. V profesionálních zařízeních se pracuje i s větším rozlišením, ale výsledek se zpravidla převádí na 16 bitů.

Logickým důsledkem konečného počtu hladin a následného přiřazování je, že kvantovaný signál neodpovídá přesně původnímu analogovému signálu. Rozdíl mezi hodnotou analogového signálu v daném čase a kvantovanou digitální hodnotou je dobře patrný z obrázku číslo 3 a nazývá se kvantovací chyba.

Během procesu kvantizace se část informací ztrácí. Jak velká část informací to bude, záleží na šířce hladiny. Šířka hladiny určuje i poměr mezi nejmenším a největším zaznamenatelným signálem. Nazývá se dynamický rozsah. Šířka hladiny nemusí být obecně stejně velká, ale nejčastěji se používá právě lineární rozložení hladin (jednodušší zpracování signálu).



Obrázek 3 – analogový signál (šedě) a jeho digitální reprezentace (červeně)

Dynamický rozsah pro 16-bitovou kvantizaci je 96dB, přičemž lidský sluch zvládá přibližně 120 dB (viz předchozí kapitola). Jedná se ale o teoretickou hodnotu, kdy pro její dosažení je potřeba

absolutní ticho, které je v reálném světě těžko dosažitelné. Na druhou stranu různé další převodníky, použitá aparatura apod. snižují dynamický rozsah reprodukovaného zvuku.

Další podstatnou roli hraje vzorkovací frekvence. Ta určuje, jak často dochází k převedení analogové hodnoty signálu na odpovídající hodnotu digitálního signálu. Pro určení správné frekvence se využívá Shannonův teorém (v literatuře se vyskytuje pod dalšími názvy jako např. Nyquistův teorém). Ten zjednodušeně říká, že vzorkovací frekvence má být minimálně dvojnásobkem frekvenčního rozsahu daného signálu. Vzorkování s větší frekvencí je zbytečné a při vzorkování s menší frekvencí může vznikat tzv. aliasing - jev, kdy dochází k překrytí frekvenčních spekter vzorkovaného signálu, což znamená ztrátu informace. Tedy pokud se vyjde z předpokladu, že lidský sluch je schopen vnímat zvuk v rozsahu 20Hz až 20kHz, je potřeba vzorkovat frekvencí alespoň 40 kHz.

Nejpoužívanější vzorkovací frekvencí je 44,1 kHz. Hodnota vznikla tak, že se hledala frekvence vyšší jak 40kHz společná pro dvě různé normy televizního vysílání (NTSC a PAL). Mezi další používané frekvence patří 48, 96 a 196 kHz.

Výše popsaná digitální reprezentace zvuku se nazývá „pulse code modulation“ (PCM). Především z důvodu úspory ukládacích kapacit se používají různé kompresní metody, z nichž některé budou vybrány v následující kapitole.

2.2 Hudební formáty

Hudebních formátů existuje velké množství a zde budou zmíněny jen ty nejznámější. Jednotlivé kapitoly jsou rozděleny dle toho, jak dané hudební formáty na zvukový signál nahlížejí. MIDI v první kapitole pracují s databází vzorků různých nástrojů. V druhé kapitole se představují dnes nejpoužívanější formáty ztrátové komprese, které část informací původního signálu nenávratně zahodí. Následující kapitola doplňuje sekci hudebních formátů o formáty pracující s bezztrátovou kompresí. V principu jde o kompresi stejnou, jakou používají programy jako ZIP či RAR. Komprese je však vhodně upravena pro potřeby zvuku. Poslední kapitola je pak věnována kompresi MPEG, která patří mezi ztrátové, nicméně její vlastnosti jsou pro tuto práci stěžejní, a proto jí je věnována zvláštní sekce.

2.2.1 MIDI

Zkratka MIDI znamená Musical Instrument Digital Interface a využívá definovanou paletu hudebních nástrojů, kterým se prostřednictvím speciálního protokolu zadávají další parametry (hlasitost, výška, tempo atd.). Parametry se zadávají pro každý nástroj zvlášť a často se používá reprezentace v notách, která je přijatelná i pro člověka. Výsledný soubor neobsahuje přímo hudbu, ale pouze instrukce pro

její vytvoření. To pak záleží na konkrétním MIDI zařízení, které soubor převádí na zvukovou informaci.

Levná MIDI zařízení používají pro generování zvuku jednotlivých nástrojů generátor frekvencí. Toto řešení zní velmi uměle.

Zvukově věrnějším řešením je použití vzorků (samplů) nástrojů. Dané rozhranní má k dispozici předem nahraný (navzorkovaný) zvuk každého nástroje, takže zvuk se velmi blíží reálnému. Jedinou nevýhodou je potřeba mít někde tyto zvuky uložené.

Jedním z problémů je použití nástrojů, které nejsou obsaženy ve standardní paletě, případně zvuků, které od hudebních nástrojů nepochází vůbec. Tento problém řeší tzv. moduly, kdy v souboru jsou kromě klasických informací uložené i samply nestandardních zvuků. Samotná MIDI specifikace však moduly neobsahuje a proto vzniklo poměrně velké množství formátů, které možnosti MIDI o moduly rozšiřují. Patří sem MOD, XM, STM, MO3 a další. MO3 na rozdíl od ostatních modulů samply neukládá do běžného PCM, ale zavádí možnost využít komprimovaných formátů jako mp3 a Ogg Vorbis, což vede k další úspoře místa.

V současné době se tyto formáty příliš nepoužívají, i když některé techniky (např. použití smyček) se uplatňují nadále.

2.2.2 Ztrátová komprese

Tato kapitola se zabývá kodeky (vysvětleno níže), které používají ztrátovou kompresi. Tedy část audio dat se při převodu ztratí a po dekompresi data nejsou stejná jako originál. Pokud je však komprese zvolena vhodně, lidské ucho nemusí slyšet žádný rozdíl.

Slovo enkodér je odvozeno ze slova „encode“, a lze ho přeložit jako „zakódování“. Opačný proces provádí tzv. dekodér („decode“). Následně slovo kodek pak vzniklo sloučením těchto dvou slov. Kodek je tedy enkodér i dekodér v jednom.

Ztrátové komprese často využívají tzv. psychoakustický model, který prezentuje faktory lidského sluchu. Pomocí těchto informací se při kompresi rozhoduje, jak moc je daná zvuková složka člověkem vnímána a jestli ji lze odstranit, aniž by to lidský sluch postřehl. Pro nastavení modelu se používají poznatky z kapitoly 3.1.2 a další.

V následujícím textu se používá termín bitrate (datový tok). Bitrate je definován jako počet bitu za sekundu. V praxi se používají ještě termíny CBR (constant bitrate, stálý datový tok) a VBR (variable bitrate, proměnný datový tok). CBR tedy používá pro celý signál konstantní bitrate, což zjednodušuje synchronizaci (sekunda záznamu má konstantní velikost). Naopak při použití VBR se datový tok v průběhu signálu mění v závislosti na jeho náročnosti (počet nástrojů, dynamika atd.). Výhodou VBR je lepší kvalita než při použití CBR odpovídající průměrné hodnotě VBR. Nevýhodou pak je složitější synchronizace.

Další vlastností těchto formátů je, že zpravidla nedefinují způsob komprese. Obsahují informace o tom, jak byl signál zakódován a jak se má rozkódovat. Dále informace o struktuře dat, jak data přehrávat apod. Kvalita různých přehrávačů by tedy (při dodržení těchto informací) měla být shodná.

Průběh komprese

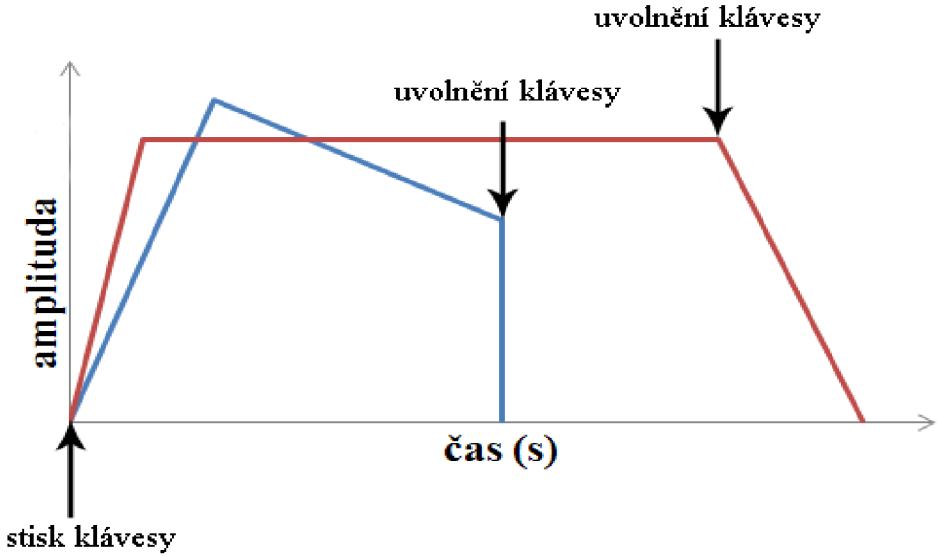
Klasické enkódování zvuku s využitím psychoakustického modelování využívá čtyř základních bloků:

- mapování
- kvantizace kódování
- framing/paketizace
- psychoaustický model

Signál se tedy nejdříve namapuje, pak dojde k jeho kvantizaci a kódování (kompresi) dle psychoakustického modelu a rozdělení na jednotlivé framy.

Mapování

Na počátku se mapuje vstupní signál. Komprese zpravidla pracují na vstupu pouze s formátem PCM. Během mapování dochází k podvzorkování na nižší frekvenci, čímž se vytvoří obálka zvuku, dle které se pak postupuje v následujících krocích. Obálka zvuku zachycuje charakteristiku vývoje amplitudy zvuku v čase.



Obrázek 4 – obálka pro zvuk varhan (červená) a klavíru (modrá)

Na obrázku 4 je patrná rozdílná charakteristika pro zvuk varhan a klavíru. Různé fáze průběhu mají své názvy, dle průběhu amplitudy (rostoucí – attack, klesající – decay, konstantní – sustain, závěr – release).

Kvantizace kódování

Fáze kvantizace a kódování bývá určujícím prvkem výsledné kvality jednotlivých druhů kompresí. Používá se vícepásmové kódování (Sub-band coding, SBC). Jak již vyplývá z názvu, celé spektrum vstupního signálu se rozdělí na pásma a každé se zpracuje zvlášť. Pro převod signálu z časové oblasti do frekvenční se použije diskrétní kosinové transformace (DCT). DCT zobrazí hlasitost jednotlivých frekvencí v signálu, takže díky psychoakustickému modelu je jednoduché odstranit frekvence, které lidské ucho není schopno vnímat (nízká hlasitost, nízká či nulová slyšitelnost frekvence, šum atd.) a zbylé frekvence pak rozdělit do jednotlivých pásem. Dochází tak samozřejmě k úspoře místa. Data, která zůstanou se pak zakódují podobně jako u bezztrátových kodeků (Huffman či Riceo kód, viz sekce *Kódování entropie*).

Mezi novější algoritmy patří Spectral Band Replication (SBR), který vylepšuje kompresi tím, že odstraňuje harmonickou redundanci ve frekvenční oblasti. Princip je založen na porovnání originálu se zakódovaným signálem a tento zakódovaný signál se rozšíří o další informace. Zejména o vyšší harmonické frekvence, které kompresí s nižší hodnotou bitrate bývají ořezány. Nejlepší výsledky podává algoritmus SBR u hlasu.

Framing

Framing (paketizace) rozděluje souvislý datový tok na framy (pakety) – viz kapitola *Bezztrátová komprese - Framing*. Navíc se přidávají informace důležité při dekompresi (různé parametry zvuku, metody kódováni apod.).

Psychoakustický model

Psychoakustický modelu analyzuje vstupní signál a řídí kvantizaci kódování. Psychoakustický model je u různých kodeků velmi podobný, neboť jeho jediným úkolem je zohlednit vlastnosti lidského sluchu.

MP3

Rozšířeným pojmem mp3 se obvykle myslí MPEG Audio Layer 3. Tento formát byl vyvinut Fraunhofferovým institutem ve spolupráci s poznatky dalších firem a jeho využití je podmíněno zaplacením licenčních poplatků. Vychází z formátu MPEG-1 Audio Layer 2, narozdíl od něj však používá modifikovanou DCT (MDCT). Tato transformace umožňuje používat delší časová okna,

která se navíc mohou překrývat. Lze použít až 576 pásem, což dovoluje větší kompresi ve spojitosti s agresivnějším psychoakustickým modelem.

Další vlastností mp3 je tzv. joint-stereo mód, kdy dochází k odstranění neslyšitelných stereo efektů. Levý (L) a pravý (P) kanál se převedou na L+R kanál a rozdílový kanál L-R. Tímto se opět zvyšuje možná komprese, nicméně ve vícekanálových systémech (Dolby Sorround, ProLogic atd.) tak dochází ke ztrátě prostorové informace.

Formát podporuje několik vzorkovacích frekvencí (32, 44,1 a 48 kHz), ale nejpoužívanější je 44,1 kHz (vychází z audio CD). Datový tok nabývá hodnot od 32 do 320 kbit/s. Navíc umožňuje použít rozdílný bitrate pro různé framy, tzv. VBR (viz úvod kapitoly 2.2).

Celým standardem MPEG se do větší hloubky zabývá kapitola 2.3 MPEG.

AAC

Tento formát byl definován v rámci video formátu MPEG-2 a MPEG-4. Vychází z mp3 a dále rozšiřuje jeho vlastnosti. Nabízí možnost využít 256 až 2048 pásem u MDCT a přidává vzorkovací frekvenci 96 kHz a zvyšuje počet kanálů na 48 a dalších 15 nízkofrekvenčních kanálů (např. pro subwoofery). Kvůli přenosovým chybám zavádí i detekci a korekci chyb. V současné době již existuje nástupce formátu AAC označovaný jako AAC plus nebo High Efficiency AAC, který aplikuje SBR.

Ogg Vorbis

Tento formát je narozdíl od mp3 otevřený a není zatížen žádnými patenty. O jeho správu se stará Xiph.Org Foundation. Jeho rozšíření však není tak velké jako právě u mp3, byť při srovnávacích testech dosahuje lepšího výsledky při stejném datovém toku. Jeho kvalita je srovnatelná s AAC a je nižší než u AAC plus.

Formát pracuje s MDCT a umožňuje využít až 8192 pásem. Podporuje i vícekanálový zvuk a kvůli možné redukci dat podporuje provázání kanálů (tzv. coupling). Komprese je čistě VBR, takže není možné nastavit přesně požadovaný bitrate.

2.2.3 Bezztrátová komprese

V poslední době se začínají rozmáhat kodeky využívající bezztrátové komprese. Po dekomprimaci jsou data zcela identická s originálem. Oproti kodekům z minulé kapitoly jsou náročnější na výpočetní výkon i na skladovací kapacitu. Toto jsou také hlavní důvody, proč k jejich rozvoji dochází až v současné době.

V podstatě se princip bezztrátové komprese, kterou tyto kodeky využívají, neliší od známých komprimačních programů jako je například Zip, Rar atd. Ovšem tyto aplikace nejsou příliš vhodné pro komprimaci hudebních dat, neboť nejsou pro tuto činnost dostatečně optimalizované a jejich výsledky tomu odpovídají.

Pro kompresi hudebních dat se tedy používají specifické postupy. Především se komprimuje každý kanál zvlášť, jelikož každý může obsahovat různá data, hodí se pro něj i různá úroveň komprese.

Schéma bezztrátové komprese je v principu vždy stejné a dá se rozdělit do následujících kroků:

- paketizace
- dekorelace
- kódování entropie

Paketizace

Používá se pro zajištění efektivní editace dat, což obnáší zejména rychlý přístup k jejich libovolné části. Zvuk je rozdělen na krátké časové úseky (tzv. pakety/rámce z anglického slova framy). Tyto úseky jsou na sobě nezávislé a lze s nimi pracovat samostatně. Každý paket má svoji hlavičku, která popisuje základní údaje o kompresi aplikované na daný paket (v typu komprese se mohou jednotlivé pakety lišit). Ideální je, aby délka paketu byla co nejkratší, což ovšem snižuje kvalitu komprese, jelikož se zvyšuje počet hlaviček a tím pádem roste i velikost dat. V praxi je vhodné najít kompromis.

Dekorelace

Tento proces se snaží najít redundantní data. Ta lze potom lépe zakódovat a doplnit jen zbytek signálu. Nejčastěji se používá metoda predikce. U predikce se signál rozdělí na vzorky a amplituda vzorku je vyjádřena přes vzorky předcházející. Po popsání signálu se zjistí chyba oproti originálu. Jedná se o neurčitý signál – entropie (míra neurčitosti).

Kódování entropie

Pro kódování entropie se používají stejné postupy (Huffmanovo kódování, Run Length Encoding, Riceovo kódování) jako u běžných komprimačních aplikací (Zip, Rar atd.).

Huffmanovo kódování zjednodušeně funguje tak, že se vytvoří tabulka (strom) četnosti výskytu jednotlivých symbolů ze vstupních dat a těm se přiřadí zástupný kód. Čím je četnost symbolu větší, tím kratší je kód a naopak. Výsledná data jsou pak zapsána pomocí těchto kódu a pro dekompresi je k nim přidána i data sloužící pro zpětný převod na původní symboly.

Riceovo kódování rozděluje vstupní hodnotu na dvě pomocí volitelného vstupního parametru. První hodnota je výsledek po dělení parametrem a druhá pak zbytek po dělení. Tyto dvě hodnoty jsou kódovány samostatně za sebou.

Run Length Encoding (RLE) kóduje tak, že vstupní data s posloupností stejných hodnot (např. „aaaxxxx“) převede na dvojici typu délka posloupnosti a hodnota (např. „3a4x“).

Závěr

Obecně se úroveň bezztrátové komprese hudebních dat pohybuje kolem 50%. Vždy však velmi záleží na konkrétním signálu (čím větší počet nástrojů, tím horší úroveň komprese apod.). Mezi nejznámější zástupce bezztrátových kodeků patří FLAC, Monkey Audio, Apple Lossless a další.

2.3 MPEG

Následující text se zaměřuje na popis MPEG (Motion Picture Experts Group), resp. jeho variant v podobě Layer-1, 2 i 3. Největší důraz je kladen na poslední z nich, který je nejvíce rozšířen pro skladování hudebních dat v počítačích. Oficiální specifikace ISO *11172-3* je zpoplatněna. Kapitola čerpá především z [7], odkud byly převzaty i některé obrázky (zejména grafy), dále pak [8] a [9].

2.3.1 MPEG úvod

Struktura mp3 souboru je tvořena jednotlivými framy, kde každý frame má hlavičku a vlastní audio data. Pokud obsahuje hudební soubor i informace o interpretovi, název písně atd. je typicky na začátku hudebních dat tzv. ID3Tag, který je právě nositelem těchto informací.

Hlavička má konstantní velikost 32 bitů. Vše znázorňuje tabulka níže, kde velikost buněk obsahujících data odpovídá počtu bitů, která data zaujímají v hlavičce. V tabulce jsou uvedeny ukázkové hodnoty včetně jejich významu. Některé bity nastavení se příliš nepoužívají nebo se jejich využití omezilo pouze na konkrétní aplikace. Z pohledu práce jsou důležité zejména informace týkající se přímo hudebních dat (samplovací frekvence, datový tok, layer, kanály a verze).

<table>
<tr><td colspan="4">Synchronizace</td></tr>
<tr><td colspan="2">Verze [1 = MPEG]</td><td>Layer [01 = Layer 3]</td><td>Hlídání chyb [1 = Ne]</td></tr>
<tr><td colspan="4">Datový tok [1010 = 160 kbit/s]</td></tr>
<tr><td colspan="2">Frekvence [00 = 44100 kHz]</td><td>Padding bit</td><td>Private bit</td></tr>
<tr><td colspan="2">Kanály [01 = joint stereo]</td><td colspan="2">Rozšiřující mód [pouze pro joint stereo]</td></tr>
<tr><td>Autorsky chráněno</td><td>Originál [1 = originál]</td><td colspan="2">Emphasis [nepoužívá se]</td></tr>
</table>

Tabulka 2 – struktura mp3 hlavičky

Struktura mp3 souboru je navržena tak, že jednotlivé framy na sobě nejsou závislé, proto lze přehrát jen část mp3 souboru (existují výjimky, kdy na sobě framy mohou být závislé) či možnost využít tento formát při streamování audia.

MPEG obecně podporuje tři různé vzorkovací frekvence (32, 44,1 a 48kHz) a maximálně dva kanály ve čtyřech různých módech:

- monofonní pro jeden audio kanál
- duální monofonní pro dva nezávislé kanály (funkčně shodné se stereem)
- stereo pro stereo kanály (mohou sdílet bity, ale nepoužívají joint-stereto kódování)
- joint-stereo, mód kdy se využívá toho, že není podstatná změna fáze mezi oběma kanály nebo korelace mezi kanály, případně obou

Jednotlivé specifikace kodeku se pak liší svými vlastnostmi, které směřují jejich využití. Nicméně všechny tři jsou dostatečně jednoduché, aby bylo možné vytvořit hardwarový dekodér na jediném čipu, který by byl schopen dekódovat v reálném čase. Kodek volitelně umožňuje také CRC (Cyclic Redundancy Check) detekci chyb a ukládání pomocných dat.

Layer-1 je nejjednodušší, nejvíce se hodí pro datový tok nad 128 kbit/s na kanál. Je používán například firmou Philips u Digital Format Cassette (DCC) v datovém toku o hodnotě 192 kbit/s na kanál.

Layer-2 je středně výpočetně náročný a používá se s datovým tokem kolem 128 kbit/s. Používá se pro Digital Audio Broadcasting (DAB) či pro ukládání synchronizovaných audio-video sekvencí na video CD.

Nejkomplexnější je pak Layer-3, který nabízí nejlepší audio kvalitu, zejména pro bitové toky okolo 64 kbit/s. Hodí se také pro přenos audio dat po ISDN.

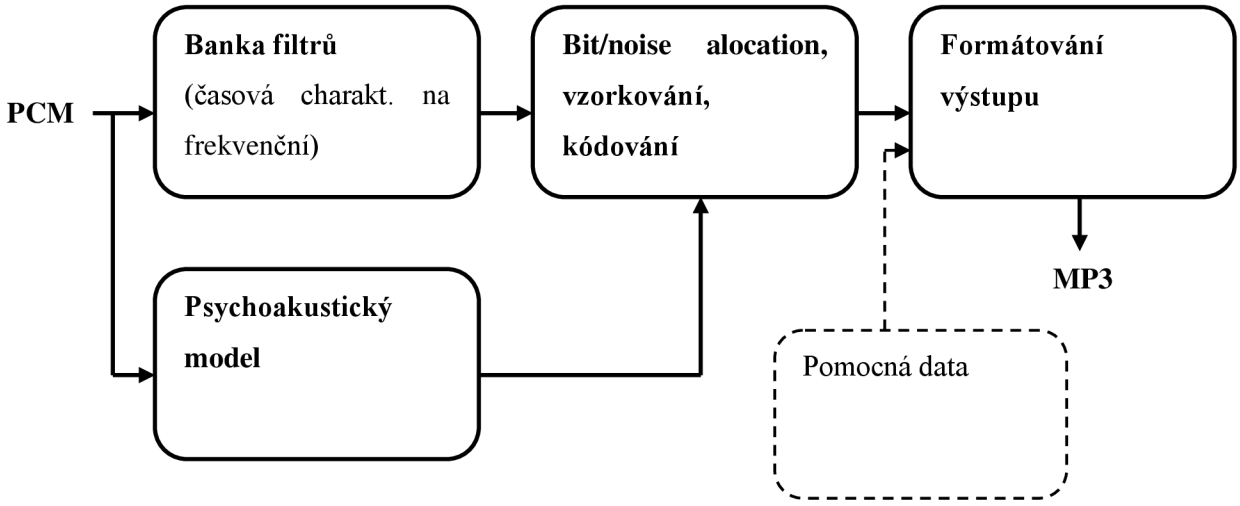
2.3.2 MPEG a ztrátovost

Všechny zde popisované formáty z rodiny MPEG jsou ztrátové. Jejich kvalita byla testována na datech těžkých pro kompresi skupinou hudebních odborníků. Přesto se ukázalo, že i komprese šest ku jedné (stereo, 16 bitů na vzorek, vzorkovací frekvenci 48 kHz a toku 256 kbit/s) již nebyla rozeznatelná od originálu, tj. odborníci nebyli schopni určit, který signál je původní nahrávka a který je komprimovaný.

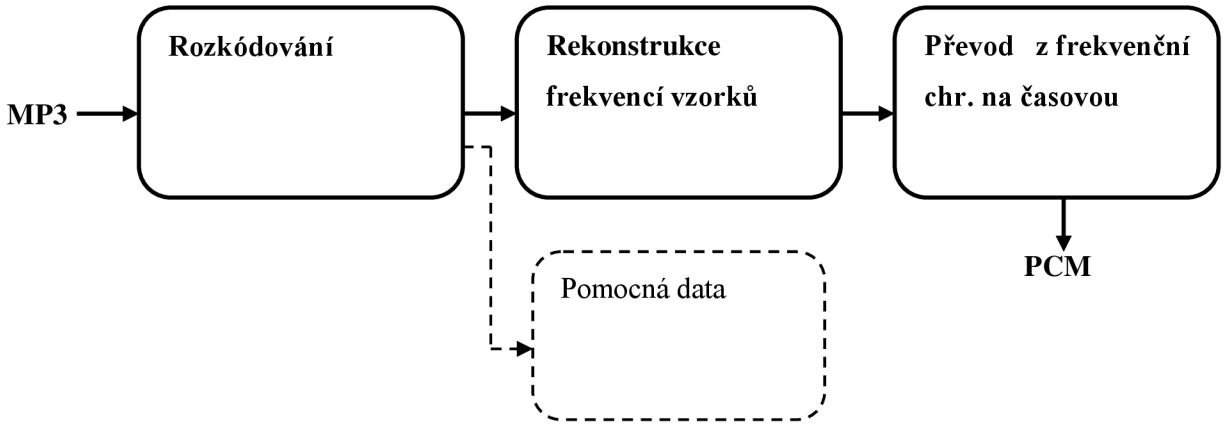
2.3.3 MPEG princip kódování

Vstupní signál prochází bankou filtrů, která jej rozděluje do několika frekvenčních hladin. Současně signál vstupuje do psychoakustického modelu, kde se určí odstup energie signálu k maskovacímu

prahu pro každou hladinu. Dalším krokem je tzv. „bit/noise allocation block“, kde se pomocí této hodnoty vypočítají kvantizační hladiny, čímž se minimalizuje slyšitelnost šumu vzniklého při vzorkování. Následně se vstupní signál navzorkuje a zakóduje. V posledním kroku se tato data společně s informacemi o nich potřebných pro dekodér zformátují do výsledného výstupu. Navíc se mohou přidat i pomocná data, která nemusí přímo souviset s audio daty. Dekodér pak funguje opačně, tedy nejdříve data rozkóduje, získá navzorkované hodnoty jednotlivých hladin a z nich pak vytvoří audio signál. Oba principy znázorňují obrázky 6 a 7 níže.



Obrázek 6 – průběh kódování

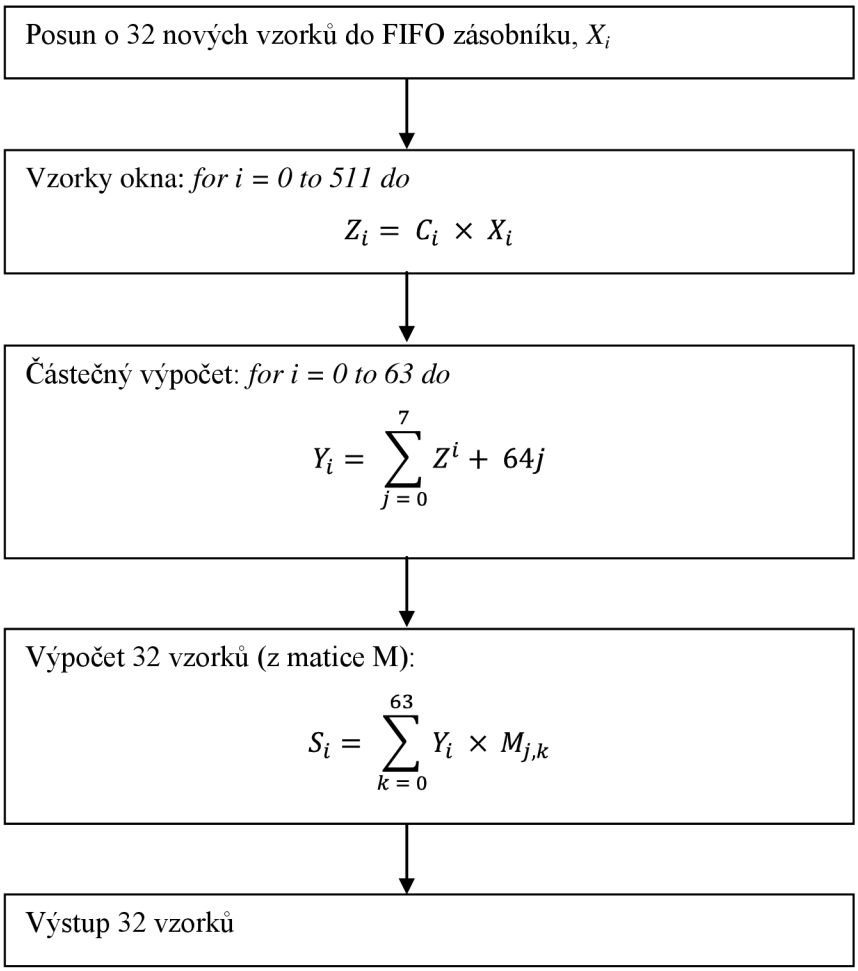


Obrázek 7 – průběh dekódování

Banka filtrů

Tento princip používají všechny verze z rodiny MPEG. Banka rozdělí vstupní signál na 32 stejně širokých frekvenčních pásem. Filtry poskytují dobré rozlišení v čase s rozumným rozlišením i ve frekvenční oblasti.

Rovnoměrné rozdělení do pásem o stejné šířce však příliš nereflektuje vlastnosti lidského sluchu. Naopak, sama šířka „kritického pásma“ je dobrým indikátorem těchto vlastností. Velká část psychoakustických efektů využívá právě škálování frekvence kritického pásma. Například jak vnímaná hlasitost signálu i jeho slyšitelnost v maskovaném signálu se liší pro signály, které jsou v rámci jednoho kritického pásma a pro signály, které pokrývají těchto kritických pásem několik. V nižších frekvencích jedna hladina obvykle pokrývá několik kritických pásem. Za těchto podmínek nelze nastavit vhodné maskování šumu na převodníku pro jednotlivá kritická pásma. Naopak, pro celou hladinu se použije maskování dle kritického pásma s nejmenší úrovní šumu.



Obrázek 8 – banka filtrů (proces)

Převod bankou filtrů není bezztrátová transformace. I kdyby se neuvažovalo vlastní vzorkování, tak není možné přesně obnovit původní signál. Nicméně vzniklá chyba je díky dobrému návrhu malá a neslyšitelná.

Sousedící filtry v bance se navzájem překrývají. Proto signál na jedné frekvenci může být zpracován dvěma sousedícími filtry.

Vlastní analýza signálu a výpočet využívá principu (obrázek 8) navrženého Rothweilerem [6]. Výstup filtrů lze popsat následovně:

$$s_t[i] = \sum_{k=0}^{63} \sum_{j=0}^{7} M[i][k] \times (C[k + 64j] \times x[k + 64j]) \qquad (1)$$

kde *i* je index hladiny a může nabývat hodnot 0 až 31; $s_t[i]$ je výstup filtru pro hladinu *i* v čase *t*, přičemž *t* je násobek 32 intervalů hudebních vzorků; *C[n]* je jeden z 512 koeficientů analyzovaného okna (definováno ve standardu); *x[n]* je vstupní vzorek načtený ze zásobníku, který celkem obsahuje 512 vzorků a

$$M[i][k] = \cos\left[\frac{(2 \times i + 1) \times (k - 16) \times \pi}{64}\right]$$

jsou koeficienty matice analýzy.

Rovnice (1) je částečně optimalizována, aby se snížil počet nutných výpočtů. Protože funkce v závorce je nezávislá na *i* a *M[i][k]* je nezávislé na *j*, 32 výstupních filtrů potřebuje jen 512 + 32 × 64 = 2 560 operací násobení a 64 × 7 + 32 × 63 = 2 464 operací sčítání. Tedy přibližně 80 operací násobení i sčítání na jeden výstup. Další optimalizace je možná za použití FDCT či FFT.

Aby banka filtrů vracela 1 výstupní vzorek na 32 vstupních, musí každý filtr každou skupinu 32 výstupních vzorků podvzorkovat, aby dostal právě jeden výstupní vzorek.

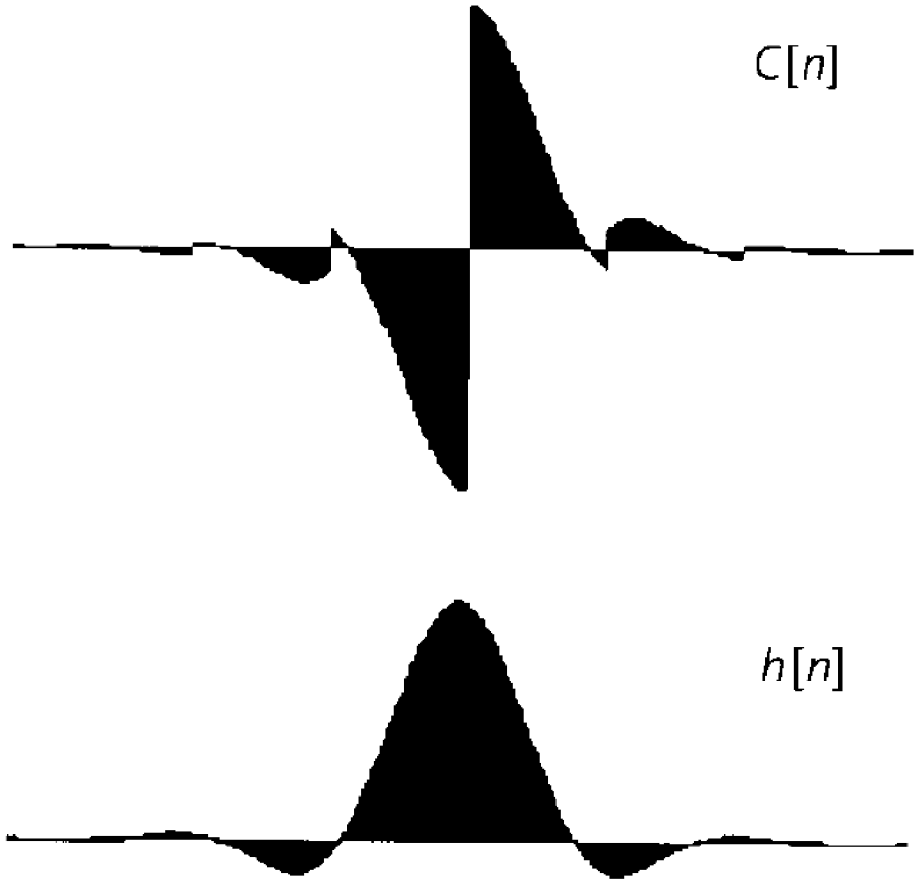
Rovnici (1) lze zapsat také jako konvoluční filtr:

$$s_t[i] = \sum_{n=0}^{511} x[t - n] \times H_i[n] \qquad (2)$$

kde x_t je audio vzorek v čase *t* a

$$H_i[n] = h[n] \times \cos\left[\frac{(2 \times i + 1) \times (n - 16) \times \pi}{64}\right]$$

kde *h[n]* = *-C[n]* pokud celá část výrazu (n/64) je lichá, jinak *h[n]* = *C[n]*, pro *n* od 0 do 511 (viz obrázek 9). Tato forma je sice více vhodná pro klasickou analýzu, není ale příliš efektivním řešením z hlediska náročnosti výpočtu. Přímá implementace by vyžadovala 32 × 512 = 16 384 operací násobení a 16 × 511 = 16 352 operací sčítání pro výpočet 32 výstupů.



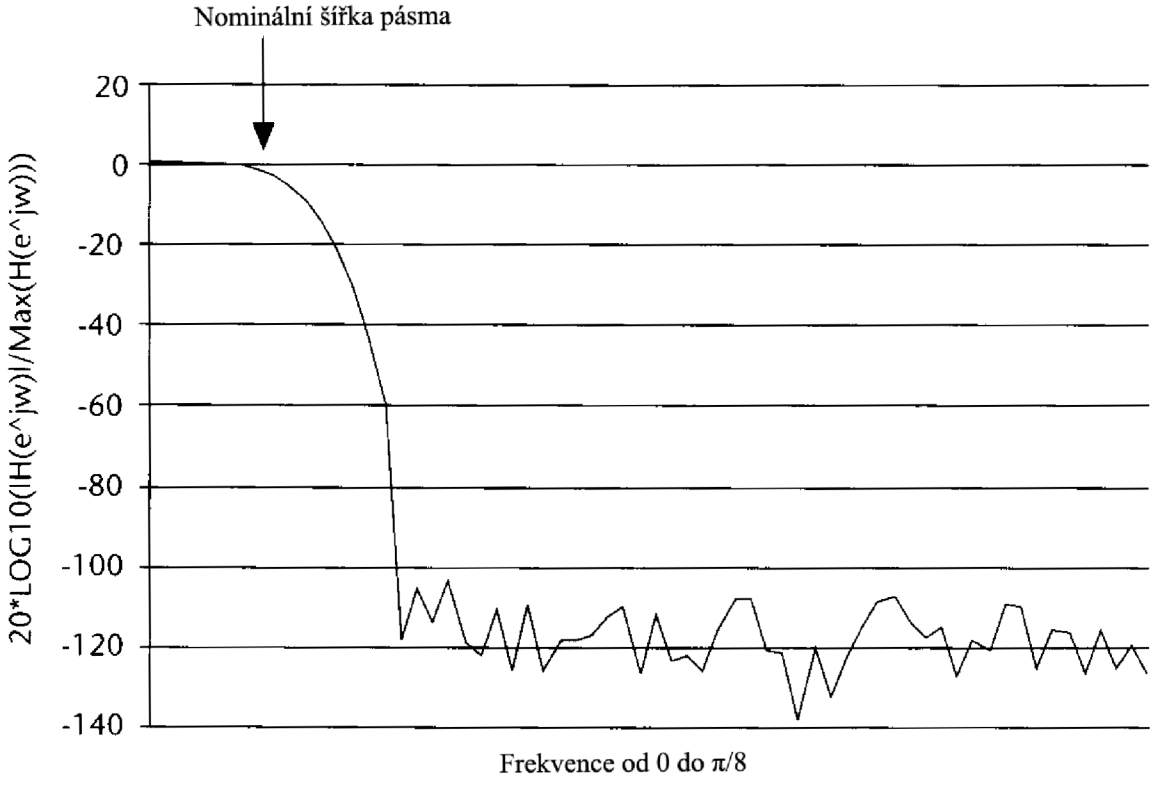
Obrázek 9 – srovnání grafů C[n] a h[n]

Koeficienty *h[n]* odpovídají odezvě filtru s dolní propustí banky filtrů. Hodnota *C[n]* má každou lichou skupinu 64 koeficientů *h[n]* negovanou pro vyvážení *M[i][k]*. Cosinus v *M[i][k]* pracuje v rozsahu $k = 0$ až *63* a tedy pokrývá liché periody, zatímco cosinus ve výrazu $H_i[n]$ pokrývá rozsah *n* od 0 do 511, tedy osm půlperiod.

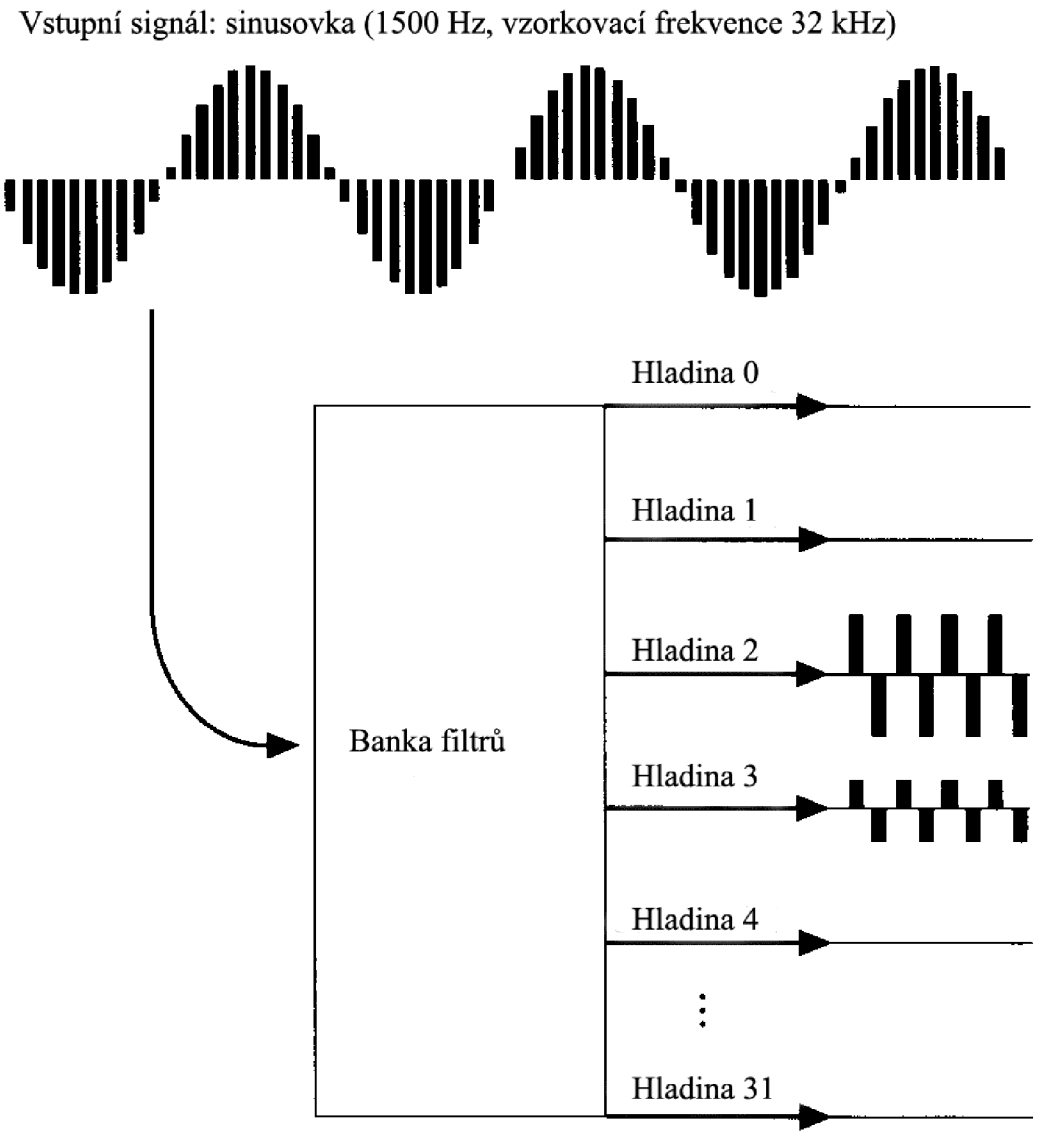
Výpočet $H_i[n]$ pak ukazuje, že každý z těchto výpočtů je modulací prototypu filtru s dolní propustí, kde cosinus posouvá odezvu tohoto filtru do patřičné frekvenční hladiny (proto se těmto filtrům říká vícefázové („polyphase“)). Tyto filtry mají střední frekvenci na lichých násobcích $\pi/(64T)$, kde T je vzorkovací perioda a každý filtr má nominální šířku pásma $\pi/(32T)$.

Odezva základního filtru nemá optimální frekvenční charakteristiku, proto při podvzorkování vzniká nezanedbatelné množství aliasu. Design filtru za předpokladu použití patřičných fázových posunů umožňuje dekodéru naprosté odstranění toho jevu.

Další vlastností použitých filtrů je přesah mezi sousední filtry. To může být nežádoucí pro efektivní kompresi, protože energie signálu, která je na rozhraní dvou filtrů se projeví na výstupu obou dvou (viz obrázek 11).



Obrázek 10 – odezva filtru



Obrázek 11 – vznik nenulových hodnot ve dvou hladinách

2.3.4 Psychoakustický model

Psychoakustický model využívá vlastností lidského sluchového ústrojí popsané dříve. Jak již bylo uvedeno, MPEG audio funguje tak, že rozdělí audio signál do frekvenčních hladin, které se blíží kritickým pásmům. Poté se každá hladina vzorkuje dle slyšitelnosti vzorkovacího šumu v daném pásmu. Pro nejefektivnější kompresi by měla být každá hladina vzorkována právě tak, aby vzorkovací šum byl neslyšitelný.

Psychoakustický model analyzuje audio signál a počítá úroveň maskování šumu jako funkci frekvence. Maskovací schopnosti signálu závisí na frekvenci signálu a hlasitosti. Enkodér používá tuto informaci pro rozhodnutí, jak nejlépe reprezentovat vstupní signál pomocí omezeného počtu bitů.

MPEG standard obsahuje dvě ukázkové implementace psychoakustického modelu (bude použito označení model-A a model-B). Model-A je méně komplexní než model-B a obsahuje více

kompromisů pro zjednodušení výpočtů. Oba modely jsou schopny pracovat se všemi třemi vrstvami („layer"), nicméně pouze model-B má potřené specifické úpravy, aby plně odpovídal pro Layer-3.

V návrhu modelů je ponechána vcelku značná volnost. Vyžadovaná přesnost modelu závisí na cíleném kompresním poměru a zamýšlené aplikaci. Pro nízké úrovně komprese, kde je dostatek bitů pro zakódování signálu, je možné psychoakustický model vynechat úplně. V takovém případě proces alokace bitů iterativně přiděluje bity hladinám s nejnižším odstupem signál šum. Pro archivaci hudby by však měl být psychoakustický model mnohem přísnější.

Časové zarovnání hudebních dat

Vyhodnocení psychoakustiky se provádí jednou za frame. Audio data zaslaná psychoakustickému modelu musí být totožná s daty, která se právě mají zakódovat. Model tedy musí počítat se zpožděním dat při průchodu bankou filtrů a s offsetem, aby relevantní data byla vycentrována v okně psychoakustické analýzy.

Převod zvuku do frekvenční oblasti

Psychoakustický model by měl používat vlastní a nezávislý převod do frekvenční oblasti než jaký používá banka filtrů. Potřebuje totiž lepší frekvenční rozlišení pro přesný výpočet prahů pro maskování. Oba zmíněné modely používají FT a standardní Hannovo okno (aplikované před FT kvůli redukci rušivých efektů na hraně filtru).

Model-A využívá okno z 512 vzorků pro Layer-1 a 1024 pro Layer-2 a 3. Protože Layer-1 má jen 384 vzorků (podrobnosti jsou uvedeny v kapitole MPEG Layer-1) na frame, není problém je tímto oknem pokrýt a výpočetní náročnost je nižší. Frame u Layer-2 i 3 má 1152 vzorků, takže okno s 1024 vzorky nepokryje frame celý. Ideální by bylo, kdyby okno pokrylo všechny vzorky, které se mají kódovat, 1024 vzorků je jakýsi kompromis. Vzorky, které nepokryje okno analýzy, obecně nemají příliš velký vliv na výsledek psychoakustiky.

Model-B používá okno s 1024 vzorky pro všechny vrstvy. Pro Layer-1 pak funguje obdobně jako u předchozího modelu, pro Layer-2 a 3 pak provede pro každý frame dvě kalkulace. V první kalkulaci se spočítá první polovina z 1152 vzorků a v druhé kalkulaci pak polovina druhá. Model-B poté zkombinuje výsledky obou kalkulací, přičemž v oblasti překrytí vybere vždy hodnotu vyššího odstupu signálu od šumu pro každé pásmo. Důsledkem toho je, že se vybere nižší ze dvou dostupných prahů pro maskování.

Rozdělení spektrálních hodnot na tóny a šum

Oba modely se snaží detekovat a oddělit tónové a šumové složky signálu, protože schopnost maskování se pro tyto dva typy liší.

Model-A identifikuje tóny na základě lokálních maxim výkonového spektra. Po zpracování všech tónů sečte zbývající spektrální hodnoty do jediné šumové složky na kritické pásmo. Frekvenční

index každé netónové složky je hodnota nejbližší ke geometrickému průměru ohraničení daného kritického pásma.

Model-B složky přímo neodděluje, ale počítá tónový index jako funkci frekvence. Index pak říká, zda je složka spíše tón nebo spíše šum. Dle hodnoty indexu se tedy i rozhodne, jaké maskování se použije. Index sám o osobě je založen na měření předvídatelnosti. Model-B používá data ze dvou předchozích analyzovaných oken, aby pomocí lineární extrapolace předpověděl hodnoty složek v aktuálním okně. Složky tónů jsou snáze předpověditelné, a proto jejich index bude vyšší. Protože tento proces vychází z více dat než u prvního modelu, měl by podávat i lepší výsledky.

Aplikace funkce rozložení

Schopnost maskování daného signálu se rozkládá přes sousední kritická pásma. První model určuje prahy pro maskování šumu aplikací empiricky definovanou funkcí. Druhý model používá funkci rozložení.

Nastavení dolní meze pro prahové hodnoty

Oba modely obsahují empiricky definovaný absolutní maskovací práh, práh ticha. Tento práh je dolní mezí slyšitelnosti zvuku.

Hledání maskovacího prahu pro každou hladinu

Oba modely počítají maskovací prahy s větší přesností než banka filtrů. Pro potenciálně velké množství prahů pro jednotlivé frekvence v pásmu musí určit jednu hodnotu pro celé pásmo.

Model-A vybírá minimum v každém pásmu. Tento postup funguje dobře pro hladiny s nižší frekvencí, kde je hladina úzce svázána s kritickým pásmem. U vyšších frekvencí může být nepřesný, protože zde kritické pásmo pokrývá interval několika hladin. Tato nepřesnost má v průběhu rostoucí tendenci, protože první model sjednocuje všechny šumové složky z jednoho kritického pásma do jediné hodnoty na jediné frekvenci. Ve výsledku tak model-A převede šumovou složky do formy složky tónové. Hladina ležící v širokém kritickém pásmu, ale daleko od sjednocené šumové složky pak nezíská přesné ohodnocení pro maskování šumu. Tento přístup je kompromisem pro snížení výpočetní náročnosti.

Model-B používá stejný princip pouze u hladin, které jsou dostatečně široké vzhledem ke kritickému pásmu v dané frekvenční oblasti. U ostatních hladin pak počítá průměr ze všech prahů. Model-B má stejnou přesnost jak pro hladiny s vyššími frekvencemi tak pro hladiny s nižšími frekvencemi, protože u něj nedochází ke sjednocování šumové složky.

Výpočet odstupu signálu od maskování („signal-to-mask ratio“ zkráceně SMR)

Psychoakustický model počítá poměr signálu k maskování jako poměr energie signálu v dané hladině k minimálnímu prahu maskování v této hladině. Model tuto hodnotu následně předává enkodéru do kroku „bit/noise allocation“ (viz obrázek 6).

Ukázka analýzy psychoakustického modelu

Obrázek 12a demonstruje graf spektra signálu, jedná se o sinusovku na frekvenci 11 250 Hz s nízkofrekvenčním šumem. Model-B ze specifikace MPEG je poněkud názornější, a proto bude psychoakustická analýza demonstrována nejprve na něm.

Obrázek 12b zobrazuje výsledek transformování zvuku do percepční oblasti a aplikování funkce rozložení. Obrázek 12c zobrazuje vypočítaný tónový index daného signálu. Je vidět zřetelný posun extrému sinusovky a rozšíření rozložení nízkofrekvenčního šumu. Transformace zvětšuje nízkofrekvenční oblast a naopak komprimuje oblast vysokofrekvenční. Protože funkce rozložení je použita v percepční oblasti signálu, její tvar přibližně odpovídá lineární funkci rozdělení („partititon“). Obrázek 12d zobrazuje graf funkce rozložení.

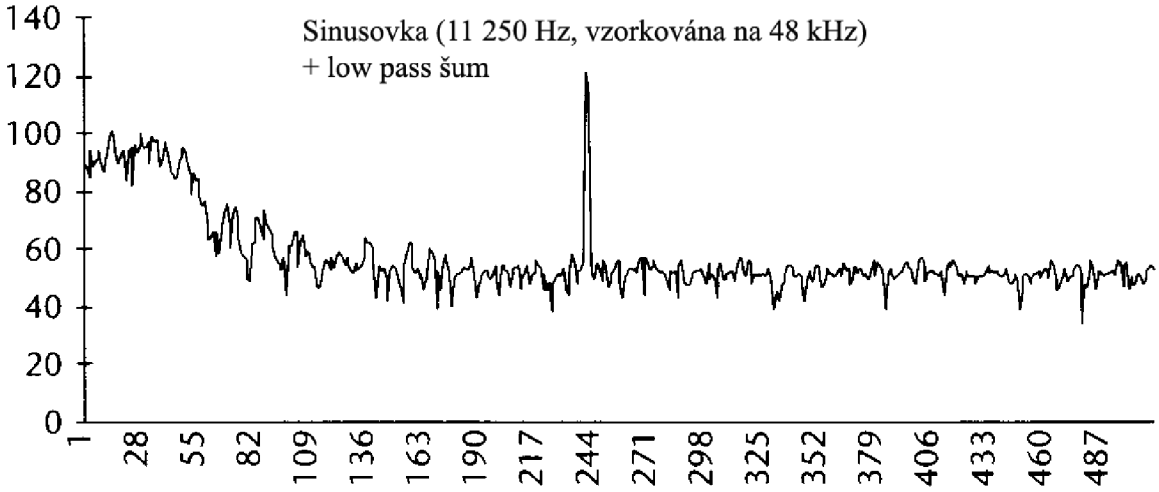
Obrázek 12e ukazuje graf prahu maskování spočítaného modelem na základě rozložení energie a tónového indexu. Obrázek zahrnuje jak graf maskovacího prahu před aplikací nižší meze prahu (viz *Nastavení nižší meze pro prahové hodnoty*) i po aplikaci pro ilustrování rozdílu. Je zřejmé, že aplikace nižší meze prahu znatelně zvyšuje maskovací práh v oblasti vyšších frekvencí, což souvisí s tím, že lidský sluch je na vyšší frekvence méně citlivý. Také je dobře patrné, jak sinový signál zvedá maskovací prahy sousedním frekvencím.

Maskovací práh je počítán ve frekvenční oblasti s rovnoměrným rozložením místo percepčním jako příprava na poslední krok, tj. výpočet odstupu signálu od maskování (SMR) pro každou hladinu. Obrázek 12f zachycuje tyto výsledky a obrázek 12g je frekvenční graf daného audio signálu z jednotlivých hodnot SMR. V tomto případě byla komprese silná (768 na 64 kBit/s), tudíž kodér nemůže zamaskovat veškerý vzorkovací šum.

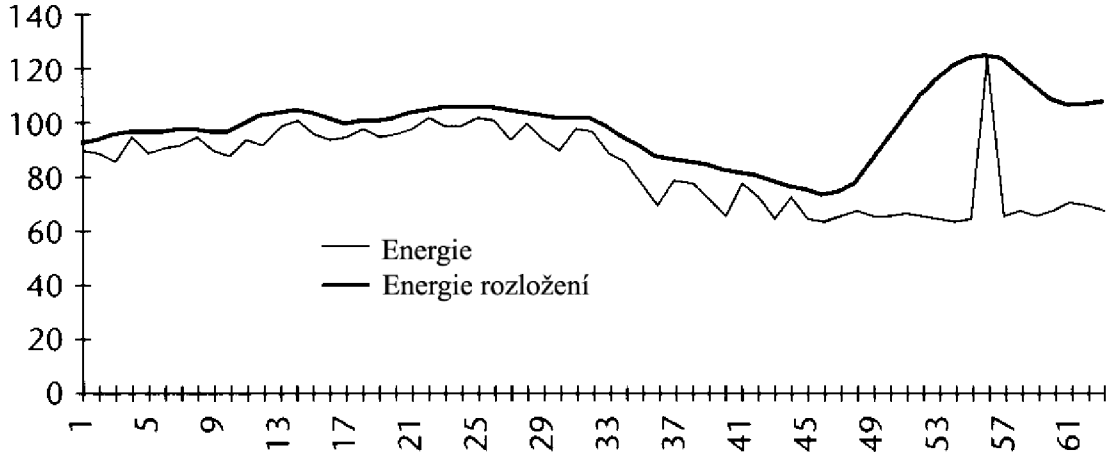
Model-A je demonstrován na stejném vstupním audio signálu. Obrázek 12h znázorňuje detekci lokálních spektrálních maxim jakožto tónových a netónových složek. Obrázek 12i zobrazuje zbylé složky po decimaci. Tento proces odstraňuje jak složky, které jsou menší než práh ticha, tak slabší tónové složky přibližně z poloviny šířky kritického pásma (0,5 Bark) silnějších tónových složek.

Model-A použije zbylé tónové a šumové složky pro určení globálního maskovacího prahu ve podvzorkované frekvenční oblasti. Tato podvzorkovaná oblast pak přibližně odpovídá percepční oblasti.

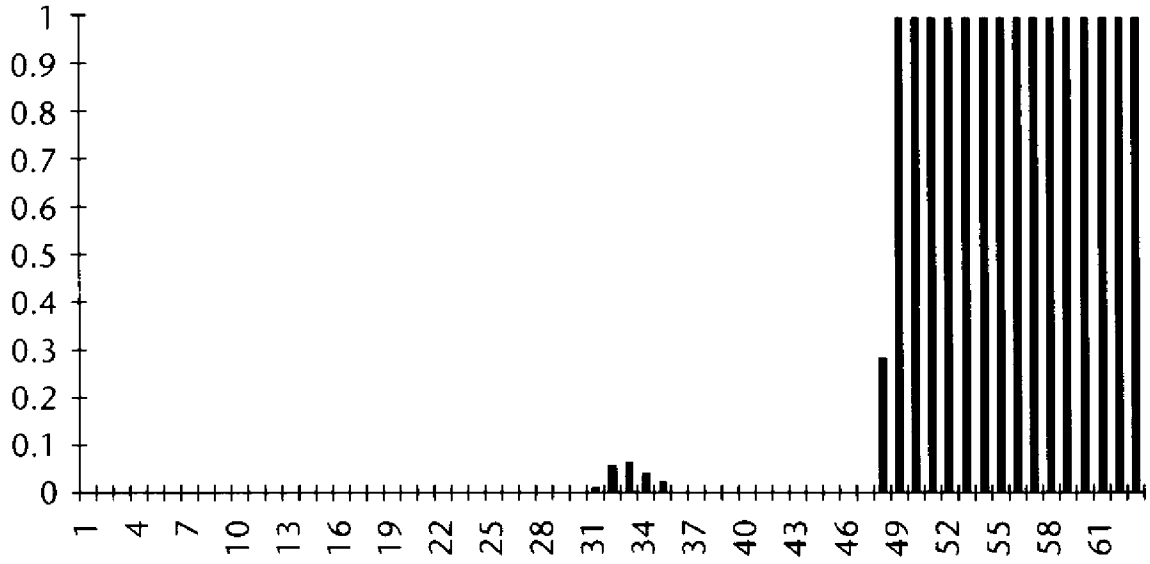
Obrázek 12j zobrazuje globální maskovací práh pro vstupní signál. Model-A následně vybírá nejmenší z globálních prahů pro každou hladinu signálu pro spočítání hodnot SMR (obrázek 12k). Obrázek 12l pak zobrazuje výsledný graf frekvence vstupního signálu z hodnot SMR.



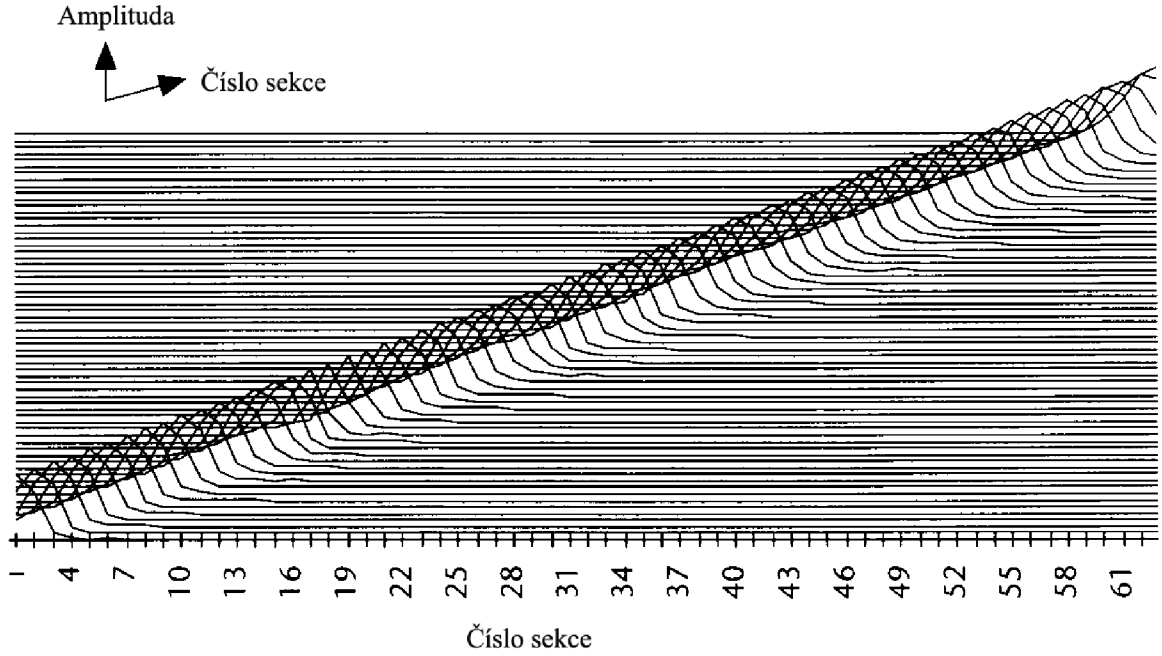
Obrázek 12a



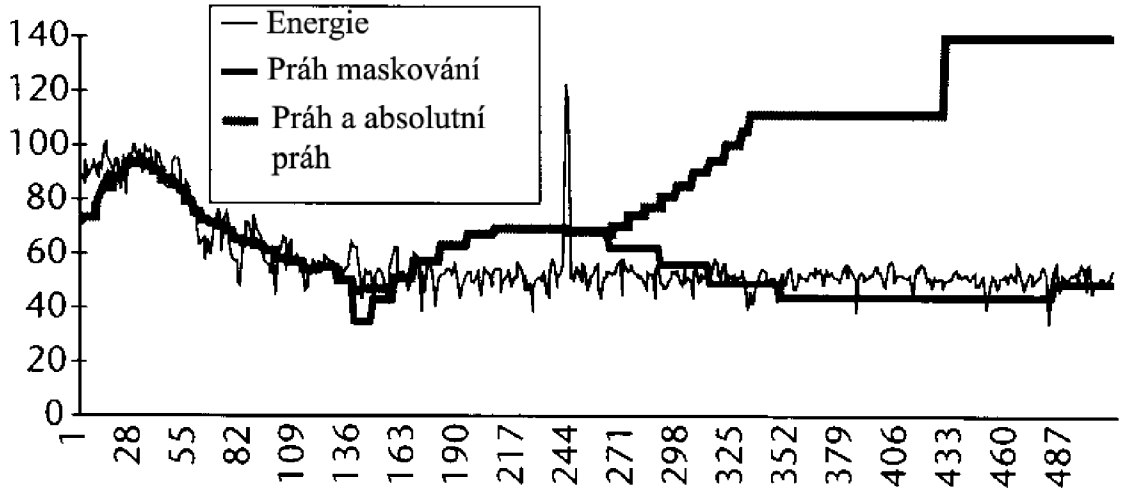
Obrázek 12b



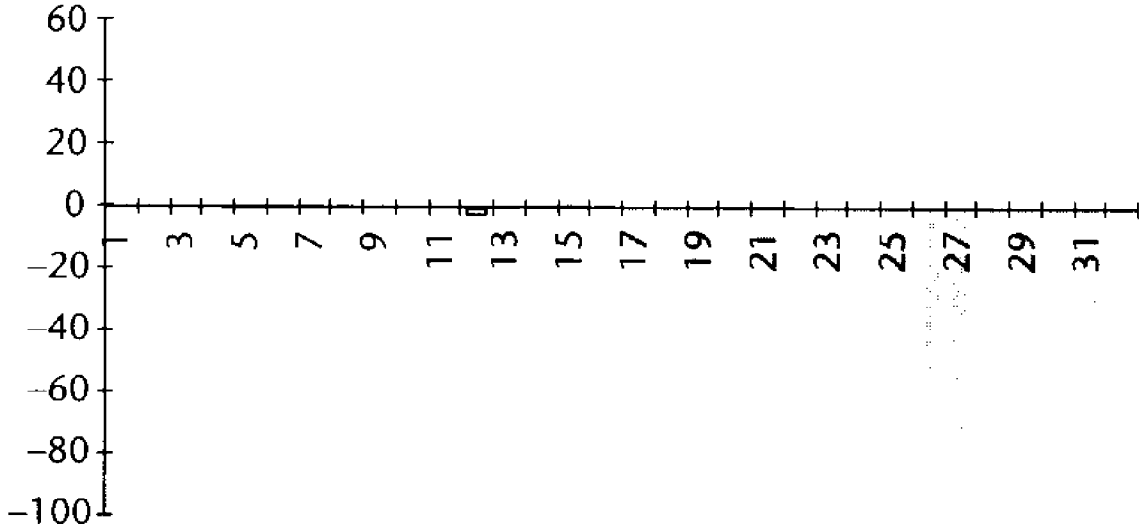
Obrázek 12c



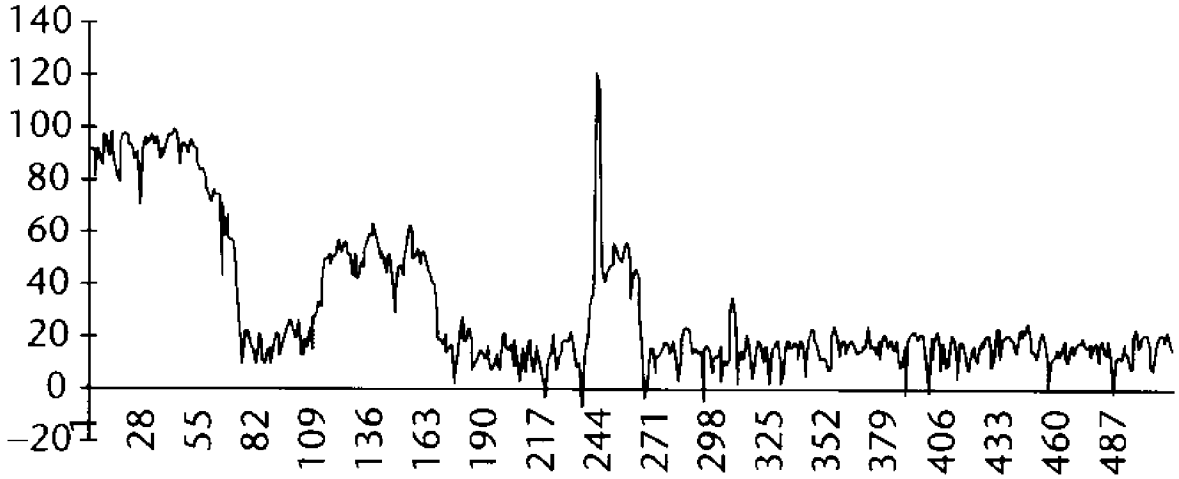
Obrázek 12d



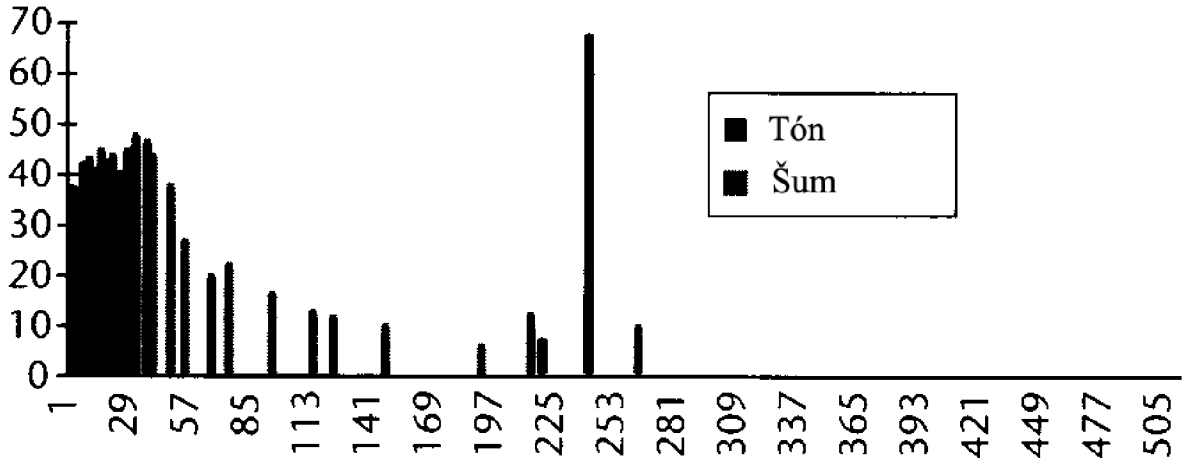
Obrázek 12e



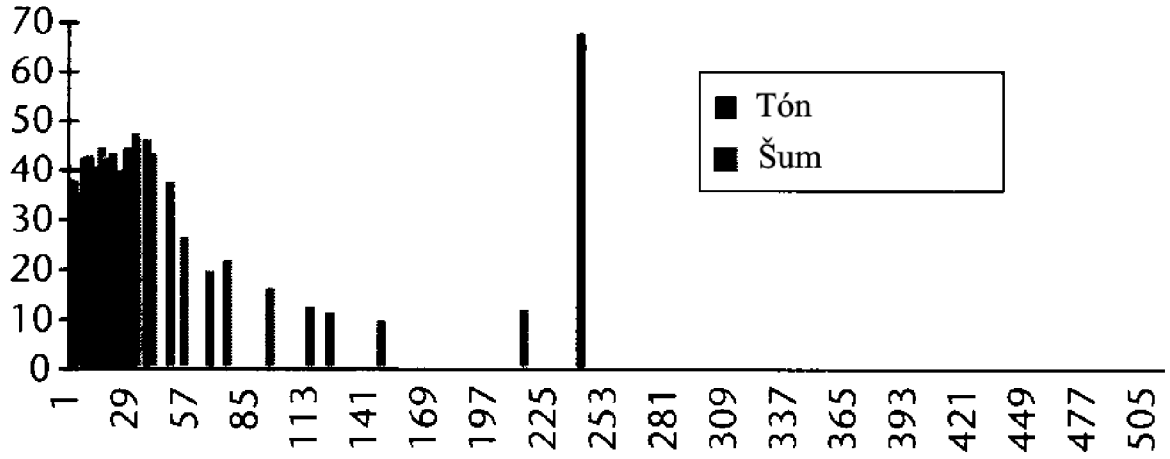
Obrázek 12f



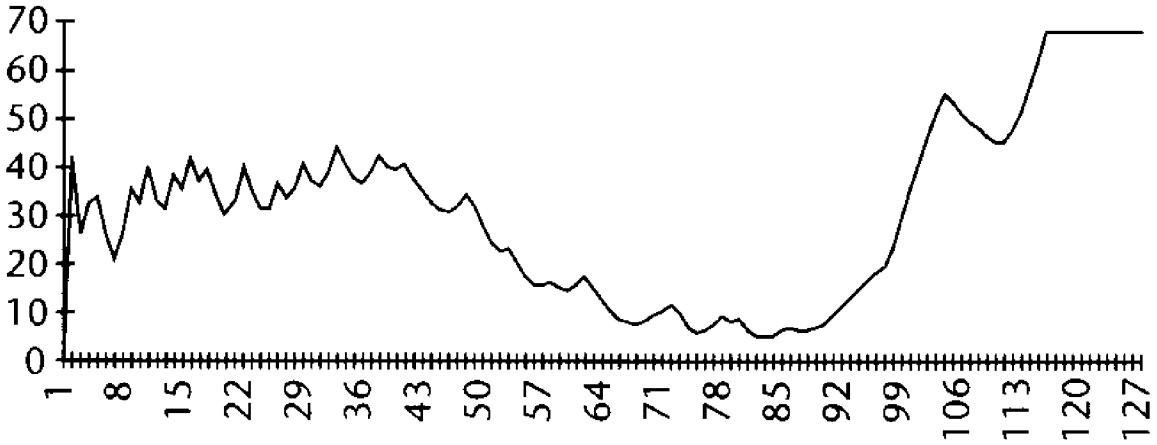
Obrázek 12g



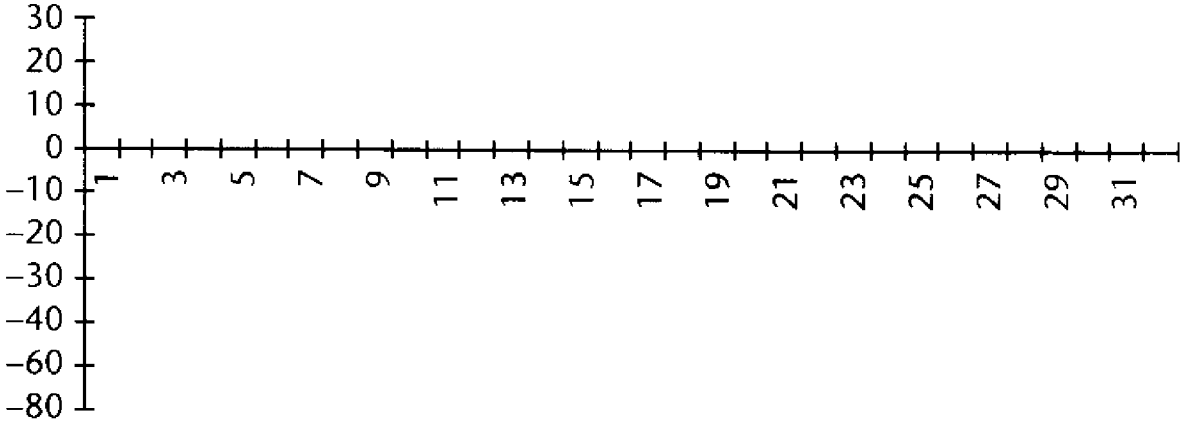
Obrázek 12h



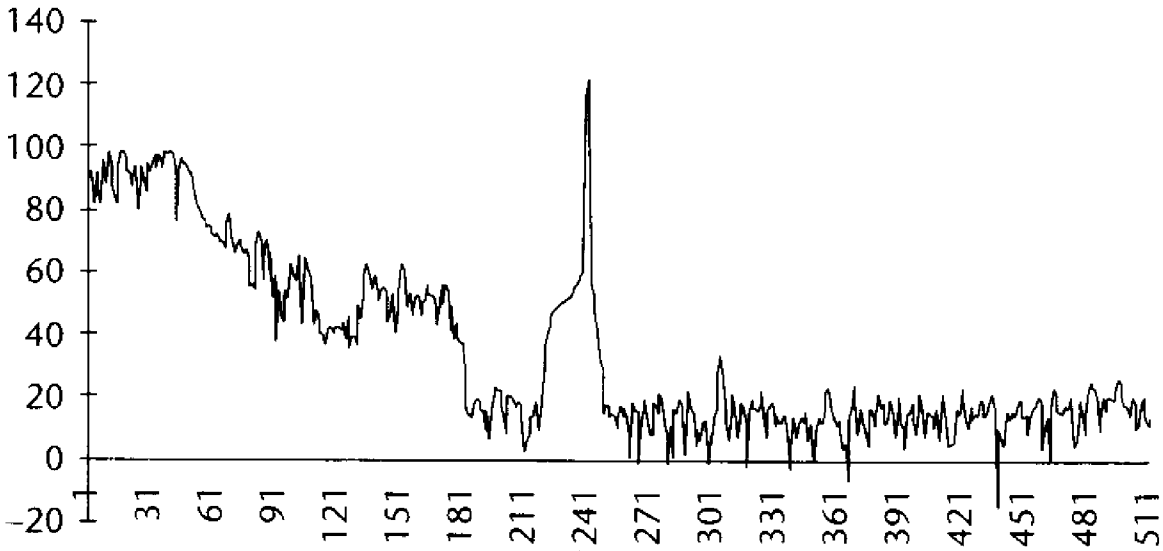
Obrázek 12i



Obrázek 12j



Obrázek 12k


140
120
100
80
60
40
20
0
-20
1
31
61
91
121
151
181
211
241
271
301
331
361
391
421
451
481
511

Obrázek 12l

3 Příznaky pro rozpoznávání (nejen) hudebních stylů

Existují různé přístupy k dané problematice, liší se pojetím jak vlastních hudebních žánrů tak příznaky používanými k rozpoznávání i algoritmy. Tato kapitola čerpá informace především z [13].

Příznak je vlastnost posuzovaného zvukového signálu. Pro přesnější určení hudebního stylu se využívá celá množina příznaků neboli vektor příznaků. Základním problémem je určit, které znaky nám umožňují určit žánrovou příslušnost. Další problém pak navazuje a týká se extrakce těchto znaků z daného signálu. Příznaky lze rozdělit do několika kategorií dle vlastnosti, kterou se snaží popisovat:

- barva tónu
- rytmus, tempo apod.
- výška tónu

3.1 Příznaky barvy tónu

Tyto znaky patří při rozpoznávání hudebního žánru mezi ty důležitější. Obvykle se získávají z každého framu vstupního signálu.

3.1.1 Koeficienty Furierovy transformace (FFC)

Signál je typicky reprezentován grafem s amplitudou signálu závislou na čase. Avšak většina důležitých informací pro rozpoznávání hudebního žánru je skryta ve frekvenčním spektru signálu. Pro převod informací se používá Furierova transformace. Souvislost s diskrétní kosinovou transformací (DCT) používanou u ztrátové komprese je více než zřejmá.

3.1.2 Cepstrální a mel cepstrální koeficienty (MFCC)

Tyto koeficienty dobře reprezentují tvar frekvenčního spektra (používají se i při detekci nástroje či v rozpoznávání mluveného slova). Cepstrální koeficienty se získávají z Furierovy transformace či koeficientů LPC (viz následující kapitola). Vlastní cepstrum se používá v rozpoznávání řeči pro oddělení buzení (v lidském těle probíhá v hrtanu) a modifikace (hltan, měkké patro, zuby, jazyk atd.). V rozpoznávání řeči je praktické tyto informace oddělit. Mel-cepstrální koeficienty se pak snaží více přiblížit principu lidského slyšení aplikací psychoakustických korelací (viz sekce 2.1.2).

3.1.3 Lineární predikce (LPC)

Lineární predikce je algoritmus pro zajištění reprezentace spektrální obálky digitalizovaného signálu. Používá se zejména v práci s řečí jako účinný nástroj analýzy i rekonstrukce hlasového signálu. LPC při analýze vychází z předpokladu, že signál je generován zařízením (např. hlasivky) a dále modifikován (např. rezonance způsobené hlasovým traktem člověka). Prostřednictvím filtrů se snaží odstranit tyto modifikace a analýzu provádí až s následně získaným signálem.

3.1.4 Spectral Centroid

Spectral centroid se snaží určit, v jakém frekvenčním rozsahu se nachází hlavní část signálu. Pocitově se v hudbě projevuje podobně jako „brightness“ (jas).

3.1.5 Spectral Flux

Tento příznak se soustředí na rychlost změny energetického spektra signálu. Výpočet probíhá přes porovnávání hodnot dvou různých rámců (jako Euklidovská vzdálenost mezi normalizovanými spektry dvou rámců). Slouží k hodnocení barvy signálu.

3.1.6 Zero Crossing Rate (ZCR)

Název příznaku lze přeložit jako četnost protínání nuly a tento příznak opravdu vyhodnocuje četnost přechodů signálu ze záporné oblasti do kladné. Tato vlastnost se používá k vyhodnocení základní frekvence, která je důležitá pro určení výšky tónu. Díky ní se určuje melodie signálu a celkové rozložení harmonie. Nicméně u hudby, kde na výsledném signálu se podílí více zdrojů (hudební nástroje, hlasy, zvukové efekty) není jasné, který zdroj se má považovat za zdroj základní frekvence a i vlastní určení frekvence se tak stává velmi obtížné. Navíc pro rozpoznání hudebního žánru není melodie skladby příliš vypovídající.

3.1.7 Spectral Roll-Off

Tento příznak se hledá jako hodnota, pod kterou se nachází požadovaná část (např. 80%) spektra.

3.1.8 Low order statistics

Pro získání příznaku se frekvenční spektrum signálu rozdělí na různá pásma. Podle toho jak velká část signálu se nachází v kterém pásmu, se určuje podobnost signálu, což může být vhodné i pro rozpoznávání hudebních žánrů. Jednotlivá pásma mohou například zjednodušeně odpovídat různým hudebním nástrojům a dle použitých hudebních nástrojů lze určit odpovídající hudební žánr.

3.1.9 Delta-coefficients

Příznak hodnotí dynamiku signálu, kde opět různým hudebním stylům často odpovídá různá dynamika.

3.2 Příznaky tempa

Dle některých studií nejsou příznaky získané z barvy tónu dostatečné pro určení hudebního stylu. Proto přidávají příznaky získané z tempa daného signálu. Rozpoznání taktu, ač pro člověka je jednoduché (např. tzv. „klepání nohou do rytmu“), je pro počítače velmi složité. Většina systému není schopna zaregistrovat případnou změnu taktu, přenesení doby přízvučné na dobu nepřízvučnou (tzv. synkopa) a další rytmické změny. Naštěstí pro rozpoznání hudebního žánru není potřeba určovat přesně takt, ale spíše tempo skladby. I když se jedná o subjektivní hodnotu, každému signálu lze přiřadit nějaké tempo.

Jednou z možností je využití „rytmického histogramu“ získaného autokorelační funkcí signálu. Histogram reprezentuje zastoupení (váhu) různých rytmických period v signálu. Případně se zpracují poměry mezi periodami.

Další možností je porovnávání průměrné amplitudy signálu a jejích špičkových hodnot. Ty mohou často určovat důraz na rytmiku daného signálu.

3.3 Příznaky výšky tónu

Rozpoznávače založené na výšce tónu nejsou příliš používané. Podobně jako u tempa lze použít histogram reprezentující jednotlivé frekvence. Nabízí se varianta slučovat oktávy nebo naopak např. pro zjištění výškového rozsahu signálu oktávy neslučovat. Pro určení žánru pak slouží předem vytvořené vzory. Vzor rockové hudby určitě nebude tak rozmanitý jako u hudby jazzové, kde je pravděpodobnost zahrání libovolné noty v běžném hudebním rozsahu poměrně vysoká.

Tyto rozpoznávače obvykle pracují s MIDI daty, kde je výška tónu implicitně dána. Vzhledem k velké variabilitě hudby je poměrně obtížné pouze na základě výšky tónu určit hudební žánr.

4 Rozpoznávací systém

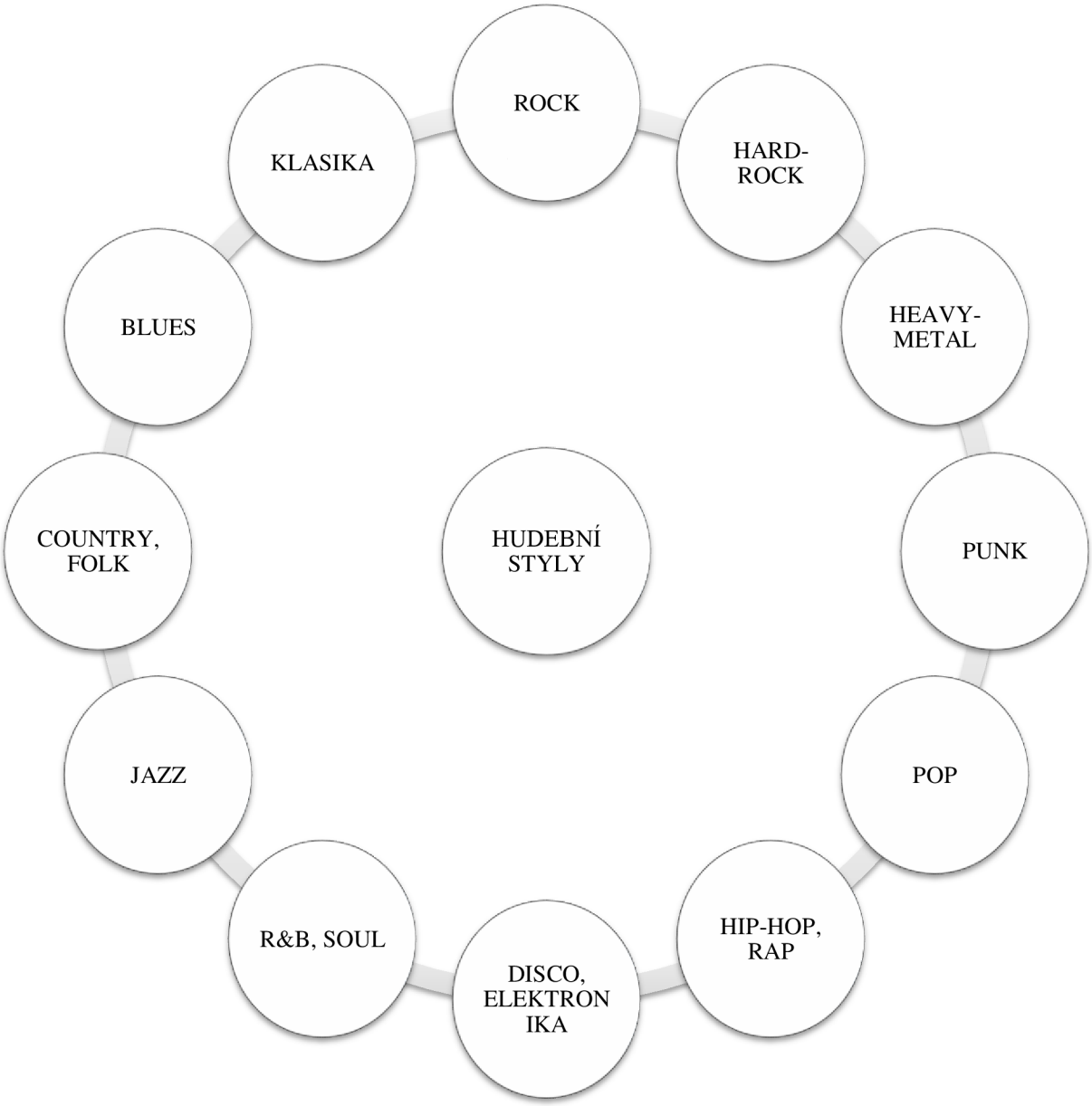
Rozpoznávací systém v této práci se snaží o rozlišení základních hudebních žánrů. Nicméně jeho hlavním smyslem je demonstrace rozdílnosti výsledků mezi klasicky získanými příznaky (analýza wav formátu) a příznaky získanými přímo z komprimovaných hudebních souborů (mp3).

4.1 Vstupní data

Do systému vstupují hudební data ve dvou různých formátech – wav a mp3. Kromě procesu získávání příznaků je ale postup stejný. Hudba byla pro účely rozpoznávače rozdělena do dvanácti žánrů (obr. 5), které se nekladou za cíl pokrýt veškerou hudební tvorbu, ale pouze zmapovat vybrané hudební žánry a na nich demonstrovat výsledky rozpoznávání.

Každý z hudebních žánrů je pro natrénování zastoupen přibližně jednou hodinou, počet nahrávek se tedy může lišit. Testovací množina dat je tvořena počtem pěti skladeb pro každý žánr. Seznam vybraných skladeb je obsažen na přiloženém CD v sekci příloh.

Formát wav je klasický 16 bitový mono záznam, při vzorkovací frekvenci 44,1 kHz. Datový tok u mp3 je nejednotný, což by nemělo mít zásadní vliv na kvalitu rozpoznávání. Minimální hodnota je 128 kbit/s, většina záznamů je však pořízena v lepší kvalitě (192 a více kbit/s). Vzorkovací frekvence je 44,1 kHz a jedná se o mono záznam.



Obrázek 5 - hudební styly

4.2 Rozpoznávač

Pro rozpoznávání je použit HTK toolkit vyvinutý univerzitou v Cambridge. Aplikuje se jak na vlastní rozpoznávání a trénování dat, tak i pro extrakci příznaků z formátu wav. Příznaky formátu mp3 jsou získány pomoci upraveného dekodéru (kapitola 4.2.2 MP3 dekodér).

4.2.1 Mp3 dekodér

Pro získání příznaků zakódovaných v mp3 kompresi je potřeba nástroj pro dekompresi. Běžné dekodéry převádí mp3 na čistá hudební data (wav), aby je dokázal přehrát počítač resp. zvuková karta.

Z kapitoly pojednávající o specifikaci MPEG je zřejmé, že příznaky jsou obsaženy v audio datech, která následují za hlavičkou každého framu. Pro jejich extrakci byl použit opensource dekodér MAD (MPEG Audio Decoder, viz [10]). Jeho specifikace plně odpovídá ISO standardu, zajímavé je, že využívá pouze operace s celými čísly, čímž je vhodný i do různých embedded systémů apod.

Pro použití nabízí dvě API, tzv. high-level API a low-level API. Při využití low-level API je možné přesně kontrolovat jednotlivé kroky dekodéru. Pomocí tohoto low-level API (vychází se z demonstrace Bertranda Petita [11]) se získají koeficienty MDCT. Koeficienty jsou získány jako číslo s fixní desetinnou čárkou. Rozsah se může pohybovat od -8.0 do necelých +8 (přesněji do +7.99999996274709701538085937 5), přičemž nejmenší zaznamenatelná hodnota je 0.00000000372529029846191406 25 (tedy přibližně $3.725e^{-9}$). Více o formátu pojednává soubor *fixed.h* (součást knihoven *libmad*).

Pomocí low-level API (viz přiložené CD) byl dekodér upraven tak, že v něm nedochází k celkové dekompresi mp3 souboru, ale pouze k zisku MDCT koeficientů a jejich zapsání do textového souboru. K dispozici je teoreticky pro každý frame 1152 koeficientů, což je pro natrénování pomocí HTK příliš velký počet. Výběr koeficientů může do značné míry ovlivnit výsledky rozpoznávání (více viz kapitola 4.3).

Koeficienty jsou dále z textového souboru převedeny do binárního formátu shodného s formátem používaným v HTK. Další kroky zpracování jsou již totožné jako u příznaků získaných analýzou hudby ve wav formátu.

4.2.2 Struktura HMM

Modely mají celkem 3 stavy, první a poslední jsou speciální nevysílací, vysílací stav je tedy pouze jeden. Funkce hustoty rozdělení pravděpodobnosti v jednotlivých stavech jsou modelovány pomocí 10 resp. 20 Gaussových funkcí s diagonální kovarianční maticí. Pro MFCC_0 (viz kapitola 4.3) je pak jedna funkce hustoty rozdělení pravděpodobnosti popsána vektorem 23 středních hodnot (mean) a vektorem 23 rozptylů (variance). Pro MFCC_E_A_D (viz kapitola 4.3) pak vektorem 69 středních hodnot a vektorem 69 rozptylů. V modelech pro mp3 vycházejících z MDCT koeficientů bude funkce hustoty rozdělení pravděpodobností reprezentována vektorem o 32 středních hodnotách a vektorem o 32 rozptylech.

4.3 Použité příznaky a trénování modelů

Pro soubory wav byla použita dvě různá nastavení pro HCopy v HTK. První konfigurace vstupní signál převáděla na 22 MFCC koeficientů a nultý koeficient (dohromady 23 koeficientů). Druhá konfigurace opět vracela MFCC koeficienty, navíc však i log-energie, delta a double delta koeficienty (celkově 69 koeficientů). V HCopy má odpovídají nastavení notaci MFCC_0 resp. MFCC_E_A_D (pro přehlednost bude tato notace používána i v následujícím textu). Šířka okna pro výpočet jednoho rámce je standardních 25 ms a posuv okna 10ms. Jedná se o experiment a není jasné, která konfigurace bude podávat lepší výsledky pro rozpoznávání žánrů, neboť nelze příliš srovnávat s rozpoznáváním mluvené řeči.

Pro mp3 soubory jsou použity koeficienty MDCT. Počet těchto koeficientů je až 1152 na frame. Při získávání příznaků pomocí HTK se používalo 23 resp. 69 koeficientů. Počet MDCT koeficientů je tedy nutné zredukovat. Pro testování byly nakonec využity dva způsoby volby koeficientů. Pro první způsob (dále značen Metoda-30%) se rozpočítá frekvenční rozsah pro jeden koeficient a následně se vybere 32 koeficientů z dolních 30% rozsahu pro každý frame. Metoda se soustředí na koeficienty odpovídající nižším frekvencím, které popisují obálku spektra daného framu. Druhý způsob (Metoda-36) se snaží o zastoupení v celém frekvenčním rozsahu a vybírá každý 36. koeficient (vrací tedy opět 32 koeficientů pro frame).

Bylo by samozřejmě možné vyzkoušet i další typy výběru a je možné, že by dosáhly i lepších výsledků. Celkově se však jedná o experiment, který má zjistit, zda je vůbec možné koeficienty MDCT k rozpoznávání vyžít. Také je potřeba uvážit časovou náročnost výpočtu, který zejména při trénování a analýze dat není zanedbatelný.

4.4 Dosažené výsledky

4.4.1 Klasické příznaky

Testování probíhalo v několika sériích s různým nastavením. Nastavení se lišila v počtu Gaussových funkcí a v získávaných příznacích (viz předešlé kapitoly 4.3 a 4.2).

Nejprve bylo rozpoznávání otestováno na datech, na kterých se modely trénovaly. Tento proces je v práci označen jako „validace“. Obecně se má validace provádět se speciálním validačním setem. Nedostatek hudebních dat však neumožnil vytvořit vlastní validační set, proto byl použit trénovací set. Vypovídající hodnota této validace může být zkreslena, nicméně má primárně posloužit pro srovnání mezi metodami, kterými se práce zabývá.

Úspěšnost při validaci se pro všechny typy měření pohybovala nad hranicí 82%, což lze považovat za uspokojivé (viz tabulka 3). Příznaky MFCC_0 si vedly lépe než MFCC_E_D_A, podobně jako vyšší počet Gaussových funkcí vedl k lepším výsledkům.

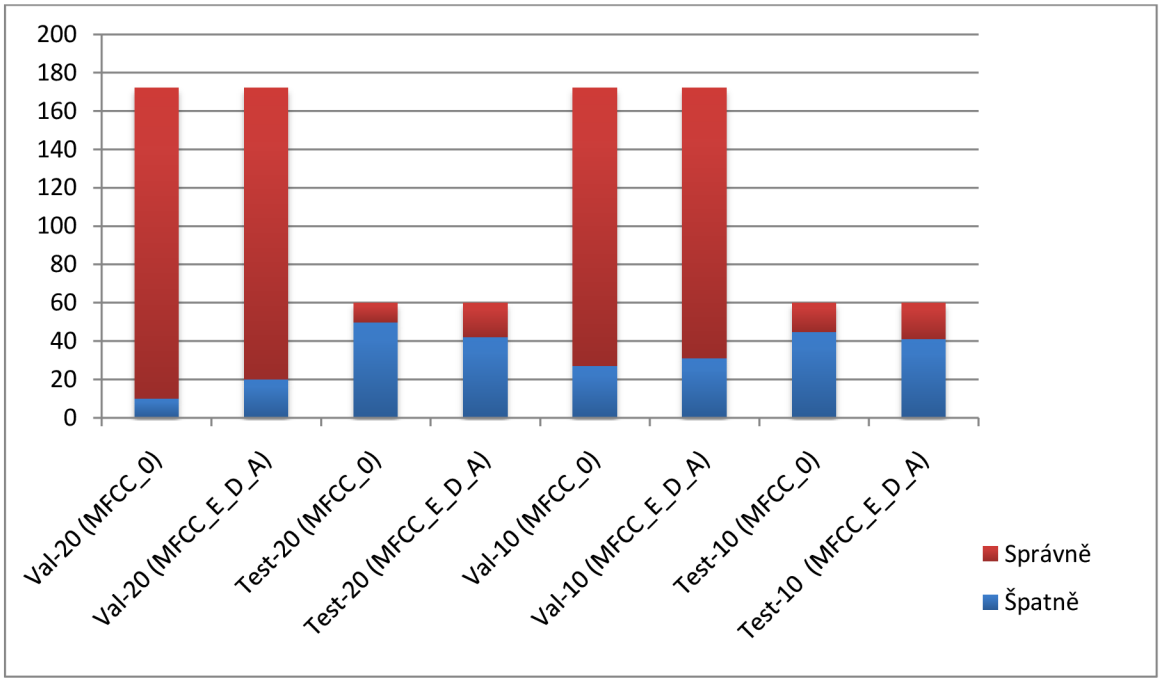
Validace	úspěšnost (%)
MFCC_0, 10 gauss.	84,3
MFCC_E_D_A, 10 gauss.	82,0
MFCC_0, 20 gauss.	94,2
MFCC_E_D_A, 20 gauss.	88,4

Tabulka 3 – úspěšnost validace

Při použití testovacích dat byly výsledky daleko horší a nelze je považovat za uspokojivé (tabulka 4). Lepší úspěšnost vykázal rozpoznávač při použití příznaků MFCC_E_D_A, které si naopak u validace vedly hůře. Zajímavý je fakt, že test s menším počtem Gaussových funkcí dopadl lépe. Pravděpodobný důvod bude rozebrán dále.

Test	úspěšnost (%)
MFCC_0, 10 gauss.	25,0
MFCC_E_D_A, 10 gauss.	31,7
MFCC_0, 20 gauss.	16,7
MFCC_E_D_A, 20 gauss.	30,0

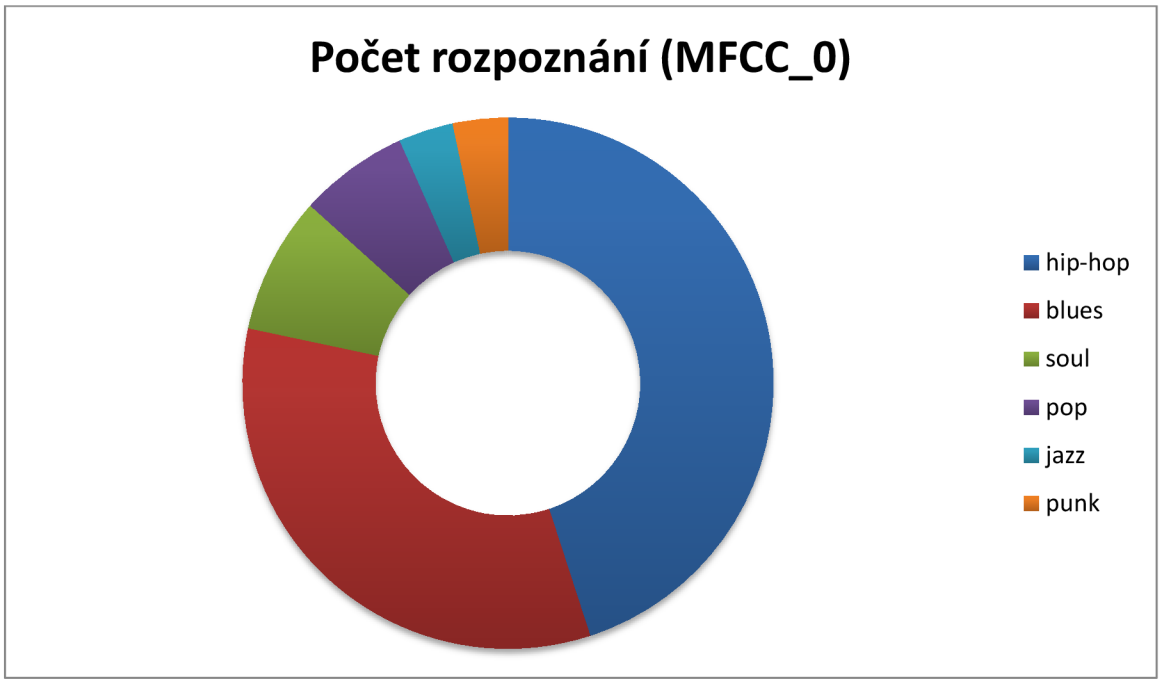
Tabulka 4 – úspěšnost testu



Graf 13 – souhrn výsledků

Rozdíl mezi úspěšností testování a validací je zřejmý. Všechny výsledky jsou zobrazeny v *grafu 13*, který zobrazuje počet úspěšného/neúspěšného rozpoznání žánru skladeb.

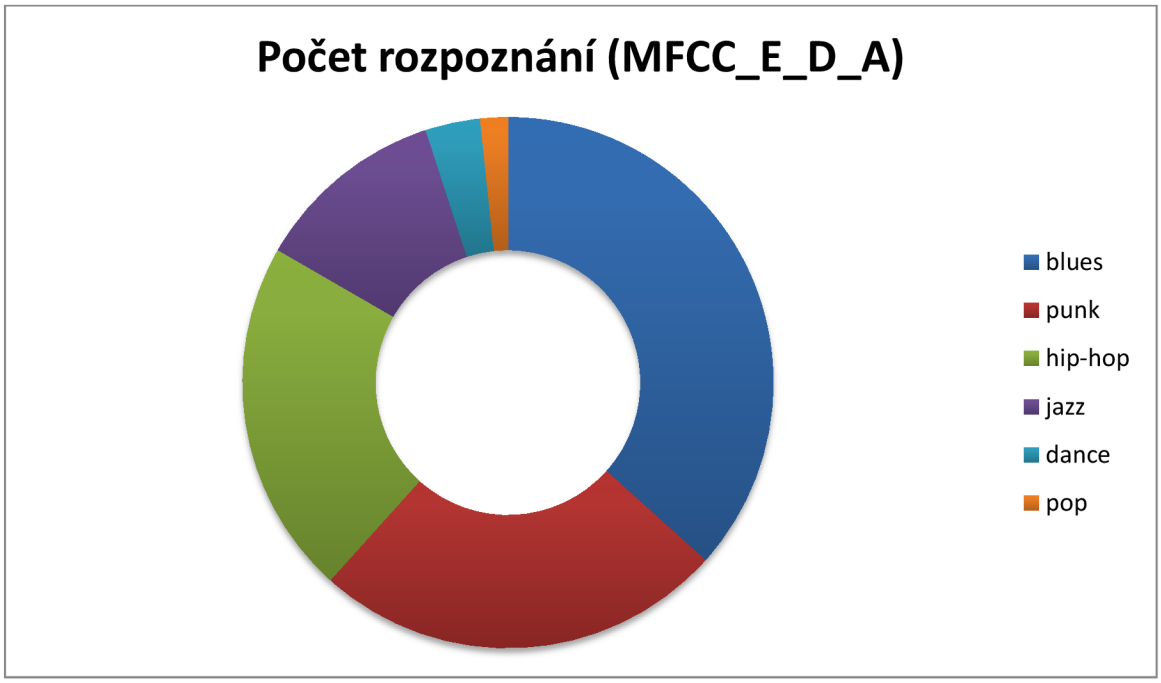
Protože větší počet gaussových křivek podal při testování horší výsledek, je vhodné výstup pro toto nastavení důkladněji analyzovat. Následující grafy se snaží zachytit, který hudební žánr byl vyhodnocován nejčastěji nezávisle na jeho správnosti. V prvním případě (*Graf - 14*) byla většina (78,3%) testovacích písní zařazena do dvou žánrů, blues (33,3%) a hip-hop (45%).



Graf 14 – dominance žánrů

Ve druhém případě (*Graf - 15*) byly dominantní (83,3%) tři žánry, blues (36,6%), punk (25%) a hip-hop (21,7%). Protože zastoupení jednotlivých hudebních žánrů bylo mezi testovacími daty rovnoměrné, znamená to, že modely pro blues a hip-hop (resp. blues, punk a hip-hop) nejsou ideálně natrénovány, neboť přesahují i do ostatních stylů. Podobně je možné vysvětlit, proč data natrénovaná pomocí 10 Gaussových funkcí podala lepší výsledky. Dominantní modely pravděpodobně nebyly natolik dominantní, a proto bylo rozložení mezi žánry rovnoměrnější a výsledky přesnější.

Také je možné, že modely s větším počtem Gaussových funkcí byly přetrénované (tzn. variance jednotlivých Gaussových funkcí je malá). Přetrénované modely obvykle velmi dobře rozpoznávají trénovací data, ale velmi špatně data testovací.



Graf 15 – dominance žánru 2

4.4.2 Příznaky získané z mp3

Testování probíhalo obdobně jako u klasických příznaků. Opět byla provedena „validace“, kde výsledky již nebyly tak dobré (tabulka 5). Zajímavý je pokles úspěšnosti u rovnoměrně vybraných koeficientů při trénování s 20 Gaussovými funkcemi. Je možné, že výpovědní hodnota koeficientů není dostatečná a v kombinaci s důkladnějším trénováním jsou pak jednotlivé modely příliš podobné.

Validace	úspěšnost (%)
Metoda-30%, 10 gauss.	61,0
Metoda-36, 10 gauss.	54,1
Metoda-30%, 20 gauss.	65,1
Metoda-36, 20 gauss.	48,3

Tabulka 5 – úspěšnost validace

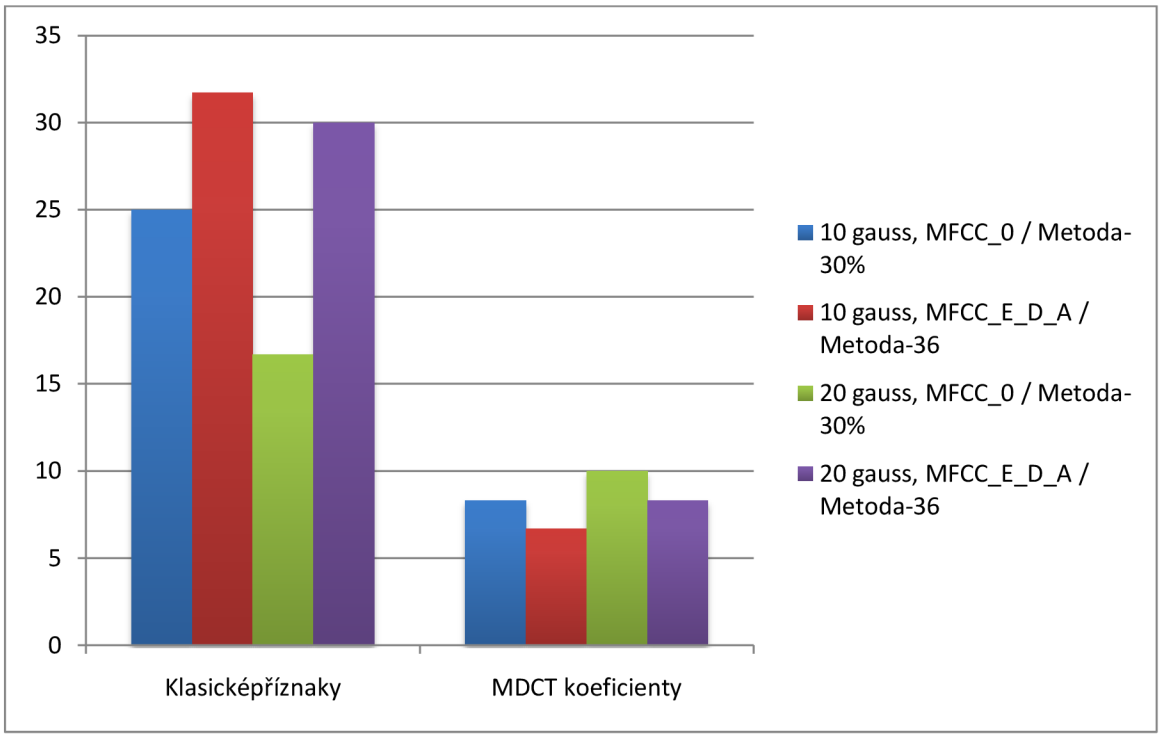
Vzhledem k nízké úspěšnosti validace se daly očekávat i nízké hodnoty při rozpoznávání testovacího setu dat. Dosažené výsledky jsou opravdu nepoužitelné v praxi (tabulka 6). A výsledky ani nelze dále příliš analyzovat, protože v nich nelze vypozorovat žádné dominantní žánry jako v předchozí kapitole ani jiné závislosti. Navíc úspěšnost je malá a vzhledem k počtu skladeb v testovacím setu se pohybuje na hranici chyby měření.

Test	úspěšnost (%)
Metoda-30%, 10 gauss.	8,3
Metoda-36, 10 gauss.	6,7
Metoda-30%, 20 gauss.	10,0
Metoda-36, 20 gauss.	8,3

Tabulka 6 – úspěšnost testu

4.4.3 Shrnutí a srovnání výsledků

Porovnání dvou metod tedy vychází jednoznačně lépe pro klasicky extrahované příznaky (graf 16). Nicméně ani jeden z výsledků nelze považovat za dostačující a lze z nich usoudit, že hudební žánr je poměrně těžko specifikovatelný z hlediska příznaků v signálu. Dále je pravděpodobné, že za použití komplexnějšího druhu příznaků při klasické analýze wav signálu by bylo možné dosáhnout lepších výsledků. Na druhou stranu určení hudebního žánru zůstává stále značně subjektivní. Zajímavé by mohlo být srovnání s výsledky ankety, kde by skladby z testovacího setu rozdělovali lidé do stejných žánrů jako zde navržený rozpoznávač.



Graf 16 – srovnání úspěšnosti

5 Závěr

V samotné oblasti rozpoznávačů probíhá v současné době intenzivní vývoj a v budoucnu budou pravděpodobně tvořit naprosto běžnou součást nejrůznějších přístrojů a zařízení. Seznámení s principy rozpoznávačů hudebních stylů je velmi zajímavé, neboť vykazují jisté podobnosti s rozpoznávači mluveného slova a řeči a přitom jsou velmi specifické. I zde je patrný velký rozdíl mezi vlastnostmi člověka a stroje. To, co je pro člověka triviální, může být pro stroj velmi složité a naopak.

Hlavním cílem práce bylo získat MDCT koeficienty přímo z formátu mp3 a otestovat jejich možné využití na rozpoznávání hudebních žánrů. Oproti klasickému extrahování příznaků tak lze ušetřit čas převodu mp3 formátu do formátu wav i vlastní čas extrakce příznaků z takto získaného wav signálu. Získat MDCT přímo z mp3 dat se podařilo, demonstrace jejich využití při rozpoznávání hudebních žánrů však vykázala nízkou úspěšnost.

Hodnocení výsledků práce je také obtížné, neboť se nenabízí žádné srovnání. Úspěšnost rozpoznávání byla nízká pro oběma způsoby získané příznaky. Úspěšné rozpoznání hudebního stylu není otázkou několika málo příznaků, ale obecně celé škály příznaků a navíc je do značné míry subjektivní. Nabízí se srovnání s výsledky, kde by rozpoznávání žánrů prováděl určitý počet posluchačů. Tématika tedy není striktně omezena na obor informačních technologií, ale jako většina komplexnějších otázek spadá do více oborů.

Literatura

[1] Berenzweig, A., Ellis, D., Lawrence, S. *Using voice segments to improve artist classification of music*. 2002.

[2] Chai, W., Vercoe, B. *Folk music classification using hidden markov models*. 2001.

[3] Tzanetakis, G., Cook, P. *Musical genre classification of audio signals.* IEEE Transactions on Speech and Audio Processing 10, s. 293-302

[4] J. Černocký. *Temporal processing for feature extraction in speech recognition.* VUT, Brno, 2003.

[5] Young, S., Jansen, J., Odell, J., Ollason, D., Woodland, P. *The HTK book*, Entropics Cambridge Research Lab., 2002, Cambridge, UK

[6] Rothweiler, J. H. *Polyphase quadrature filters – A new subband coding technique.* IEEE ASSP, 1983, s. 1280-1283

[7] Pan, D. *A tutorial on MPEG/audio compression*. IEEE Multimedia, 1996

[8] Mp3. [online], http://en.wikipedia.org/wiki/MP3. [rev. 2009-05-03], [cit. 2009-05-03]

[9] Psychoacoustic. [online], http://en.wikipedia.org/wiki/Psychoacoustic. [rev. 2009-05-03], [cit. 2009-05-03]

[10] Mpeg audio decoder. [online], http://www.underbit.com/products/mad. [rev. 2009-05-03], [cit. 2009-05-03]

[11] MPEG audio decoder low-level demonstration. [online], http://www.bsd-dk.dk/~elrond/audio/madlld. [rev. 2009-05-03], [cit. 2009-05-03]

[12] HTK book. [online], http://users.ece.gatech.edu/antonio/htkbook/node2_tf.html. [rev. 2009-05-15], [cit. 2009-05-15].

[13] Aucouturier, J., Pachet, F. *Representing Musical Genre: A State of the Art.* Journal of New Music Research, 2003