VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ

DEPARTMENT OF TELECOMMUNICATIONS

ROZPOZNÁNÍ A KLASIFIKACE EMOCÍ V HUDBĚ A V ŘEČI

THE EMOTIONS AT THE MUSIC AND SPEECH RECOGNITION AND CLASSIFICATION

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — Vojtěch Rezek

AUTHOR

VEDOUCÍ PRÁCE — prof. Ing. Jana Tučková, CSc.

SUPERVISOR

BRNO 2022

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní program **Audio inženýrství**
specializace Zvuková produkce a nahrávání
Ústav telekomunikací

Student: Vojtěch Rezek ***ID:*** 219343
Ročník: 3 ***Akademický rok:*** 2021/22

NÁZEV TÉMATU:

Rozpoznání a klasifikace emocí v hudbě a v řeči

POKYNY PRO VYPRACOVÁNÍ:

Vytvořte databázi zvukových nahrávek emocí obsažených v řeči a v hudbě. Ve vytvořené databázi zvolte příznaky pro klasifikaci emocí pro oba typy signálu. Nahrávky musí obsahovat emoce. Proveďte parametrizaci dat. Data rozdělte na trénovací a testovací. Ke klasifikaci emocí použijte samoorganizující se umělé neuronové sítě (včetně Kohonenových map). Praktickou část práce řešte v prostředí MATLAB. Výsledky klasifikace porovnejte poslechovými testy a analytickým vyhodnocením získaných výsledků.

DOPORUČENÁ LITERATURA:

[1] SYROVÝ, V. Hudební akustika. Akademie muzických umění, Praha 2003. ISBN 80-7331-901-2
[2] TUČKOVÁ, J. Aplikace umělých neuronových sítí při zpracování signálů. Nakladatelství ČVUT, 2009. ISBN 978-80-01-04400-1

Termín zadání: 7.2.2022 ***Termín odevzdání:*** 31.5.2022

Vedoucí práce: prof. Ing. Jana Tučková, CSc.

doc. Ing. Jiří Schimmel, Ph.D.
předseda rady studijního programu

ABSTRAKT

Práce se zabývá problematikou rozpoznávání a klasifikace emocí v řeči a v hudbě. V praktické části byla vytvořena řečová databáze a hudební databáze. Z těchto databází byly vytvořeny datasety obsahující vybrané parametry určené pro trénink (testování) umělé neuronové sítě. Úspěšnost klasifikace testovaných modelů je porovnána s úspěšností klasifikace prováděné posluchači. Při klasifikaci řeči přesnost trénovaných neuronových sítí přibližně odpovídá přesnosti dotazovaných posluchačů. V případě klasifikace hudby je přesnost umělých neuronových sítí vyšší než u posluchačů.

KLÍČOVÁ SLOVA

emoce, řeč, hudba, analýza signálů, strojové učení, umělé neuronové sítě

ABSTRACT

KEYWORDS

REZEK, Vojtěch. *Rozpoznávání a klasifikace emocí v hudbě a v řeči*. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav telekomunikací, 2022, 64 s. Bakalářská práce. Vedoucí práce: prof. Ing. Jana Tučková, CSc.

Prohlášení autora o původnosti díla

Jméno a příjmení autora:	Vojtěch Rezek
VUT ID autora:	219343
Typ práce:	Bakalářská práce
Akademický rok:	2021/22
Téma závěrečné práce:	Rozpoznávání a klasifikace emocí v hudbě a v řeči

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

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

Brno ....

podpis autora*

*Autor podepisuje pouze v tištěné verzi.

PODĚKOVÁNÍ

Rád bych poděkoval vedoucí mé bakalářské práce paní prof. Ing. Janě Tučkové, CSc. za poskytnutí seznamu vhodných odborných materiálů pro nastudování problematiky, trpělivost a podnětné návrhy k práci. Dále pak chci poděkovat panu Ing. Danielovi Kováčovi za velmi přínosnou konzultaci a zodpovězení mnohých otázek. Obrovský dík patří všem zúčastněným řečníkům, jejichž hlasy jsou obsaženy v řečové databázi. V neposlední řadě také musím poděkovat mé snoubence za vytvoření skvělých podmínek k tomu, abych se v klidu mohl věnovat intenzivně této práci.

Obsah

Seznam obrázků

Úvod

S vývojem metodiky a technologií v oblasti strojového učení se neustále posouvají hranice toho, co vše lze zkoumat při analýze řeči potažmo hudby. Již dlouho dobu jsou v praxi úspěšně implementovány systémy pro rozpoznávání slov či jejich následný převod do textu. V praxi se také již dávno uplatňuje automatická kategorizace hudebních žánrů. Nelze se proto divit, že snahou všemožných bádání je se posunout ještě hlouběji a pokoušet se o klasifikaci vlastností, které mnohdy nemusí zvládat přesně určovat ani samotný člověk. Jednou z takových vlastností jsou emoce. Jejich klasifikace je předmětem zkoumání již řadu let a možné uplatnění teoretických poznatků lze sledovat v mnoha oborech. Objektivně lze konstatovat, že v oblasti rozpoznávání řeči toho bylo objeveno více než v též disciplíně pro hudbu. To však v žádném případě neznamená, že neexistuje dostatek materiálů, které by se problematikou klasifikace emocí v hudbě zabývaly. Z hudebních děl lze taktéž získat nespočet užitečných parametrů, mezi kterými lze hledat souvislosti. Moji snahou v této práci bylo vyjít z některých poznatků a pokusit se o vytvoření vlastních datasetů obsahujících validní údaje pro účely tréninku umělých neuronových sítí, které poté tyto emoce měly rozpoznávat. Dále bude cílem zanalyzovat, jak moc se přesnost klasifikace neuronových sítí liší od klasifikace prováděné dotazovanými posluchači.

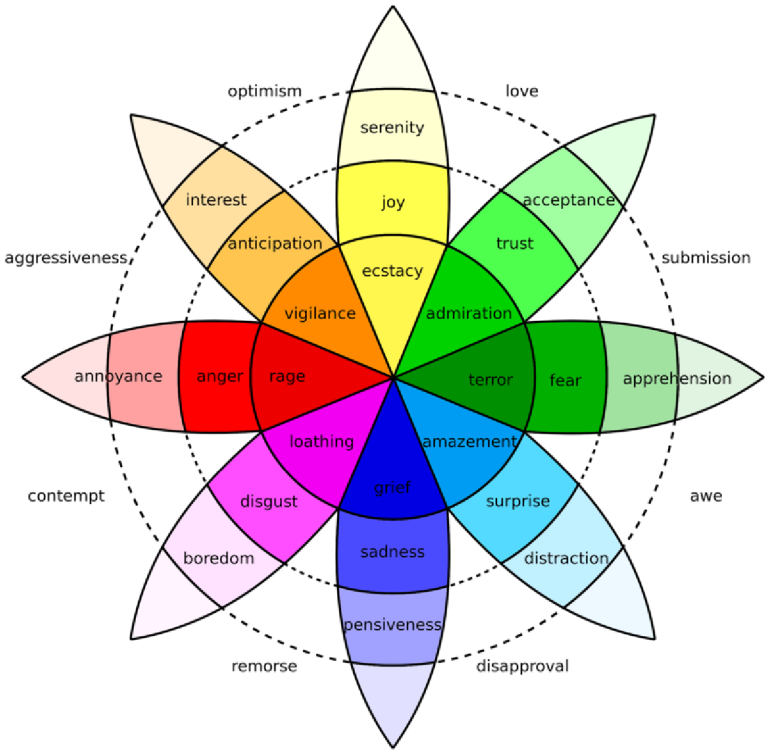
1 Emoce

Kapitola pojednává o emocích z převážně psychologického úhlu pohledu. Nadefinujeme si zde základní druhy emocí, kterými se budeme poté v následujících kapitolách věnovat i z hlediska analýzy řečového, respektive hudebního signálu. Cílem bude si ujasnit, co je pod danými emocemi přesně myšleno. Emoce předcházejí mluvenému slovu, a protože se je pomocí řeči budeme snažit analyzovat, je více než žádoucí mít jasno, co je tímto pojmem myšleno. Emoce lze charakterizovat jako komplexní jevy, pro které je typická vysoká citlivost a proměnlivost. Důsledkem vysoké citlivosti je proměnlivost emocí na základě subjektivního vyhodnocení dané situace. [1] Emoce jsou vyvolávány právě probíhajícím zážitkem či zážitkem již proběhlým (hovoříme zde o emoci vyvolanou silnou vzpomínkou). Emoce mívá obvykle svůj protiklad a lze ji prožívat v různé intenzitě. Emoce se velice těžko dají ovlivňovat. Je to z toho důvodu, že jsou evolučně starší než rozumové chápání. [2] Funkcí emocí je příprava jedince na reakci na konkrétní událost. Kromě toho emoce vyvolávají vštípení zážitku. Například strach má jedince připravit na nebezpečí a vtisknout danou situaci do paměti jako nebezpečnou. Emoce tak vedou k vymezení a hierarchizaci hodnot, vytvoření schopnosti seberegulace. [3] Projevy emocí jsou fyziologické (změna srdečního tepu, změna rychlosti dýchání atd.) a motorické (mimika, gestikulace). [4] Je nutné dodat, že řeč objektivně není tím nejspolehlivějším ukazatelem emočního stavu řečníka. Řeč v sobě nese pouze 10 procent informace o emoci.[2]

1.1 Dělení emocí

Jak je již výše uvedeno, emocí je více druhů. Velmi známým konceptem dělení emocí je Plutchikův kruh emocí. S tímto konceptem přišel v roce 1980 Robert Plutchik. Máme zde 8 základních skupin a sekvenci dějů utvářející samotnou emoci - podnětovou událost, vyvozené poznání, pocit, chování a efekt. Jak je již, řečeno Plutchik ve své práci definuje 8 základních bipolárních emoci. Tyto emoce se vyvinuly evolučně ze základních obranných mechanismů (např. strach je derivovaný z obranné reakce proti hrozícímu nebezpečí). [5] Celý tento koncept je názorně vyobrazen na kruhu. Směrem od středu nám stoupá intenzita dané emoce, radiálně se pak na kruhu rozprostírá 8 základních emocí.

V této práci se budeme snažit rozpoznávat a analyzovat 6 emocí, z nichž si jednu zadefinujeme jako emoci referenční.



Obr. 1.1: Kruh emocí [5]

1.1.1 Neutrální

Tato emoce bude mít v práci čistě referenční úlohu. Můžeme si ji představit jako střed emočního kruhu. V části práce věnující se klasifikaci emocí budou hodnoty příznaků u neutrálních promluv velmi podstatné, neboť zde bude kladen zřetel na to, jak k velkým změnám oproti neutrální náladě tu bude docházet.

1.1.2 Vztek

Z evolučního hlediska se tato emoce vyvinula z obranných mechanismů souvisejících s likvidací konkurenta či nepřítele. Vztek může být reakcí na nepříznivý chod událostí, tato emoce bývá často namířena proti jiné osobě.Kdybychom se uchýlili ke kategorizaci z hlediska výdeje energie na projevení této emoce, zřejmě bychom tuto emoci zařadili mezi emoce energeticky nejnáročnější. Tento laický předpoklad do určité míry potvrzují i závěry mé práce.

1.1.3 Smutek

Dle Plutchikovy teorie zabývající se vývojem emocí ze základních obranných mechanismů má smutek svůj původ v reakci jedince na izolaci a osamění. Smutek jako

takový samozřejmě nemusí být pouze reakcí na tyto zmíněné podněty.

1.1.4 Radost

Radost je emocí protichůdnou ke smutku. Prapůvodně to byla reakce na ukořistění potravy. Radost bývá spojována s pro jedince příjemnými prožitky.

1.1.5 Strach

Evolučně se jedná o podnět vyvolaný již existující či potenciální hrozbou. Na emočním kruhu je zobrazen strach jako opak vzteku. Narozdíl od vzteku je totiž jedinec postaven do defenzivní pozice.

1.1.6 Nuda

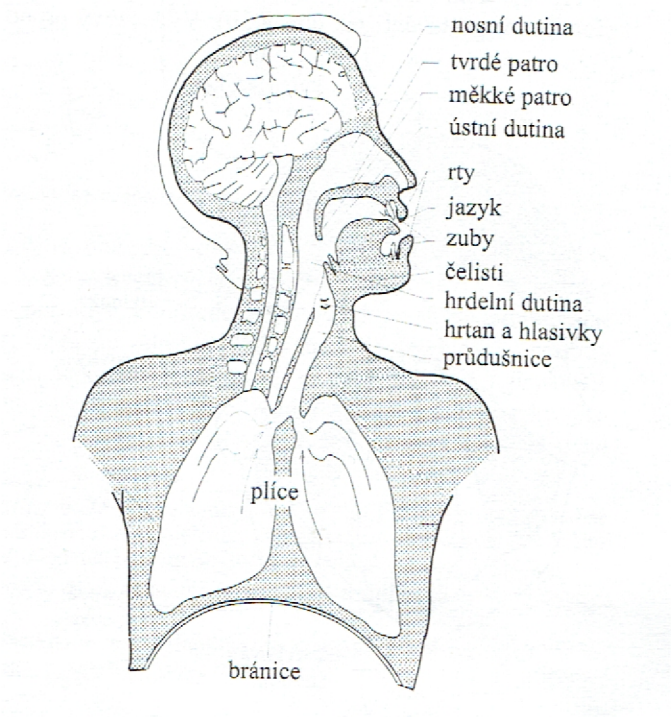
Nuda je odvozena od pocitu znechucení. Můžeme si představit našeho pravěkého předchůdce ochutnávajícího hořkou plodinu, která mu silně nechutná. Nuda je pravým opakem zájmu, jedná se o stav určité rezignace třeba na danou činnost.

2 Řeč

Kapitola je věnována řeči a její tvorbě. V této kapitole se podíváme na řeč z pohledu fonetiky. Cílem bude porozumět fungování lidského řečového ústrojí. Podíváme se na řeč i z pohledu její reprezentace. Velký důraz bude pochopitelně kladen na specifika českého jazyka, neboť zkoumané hlasové nahrávky jsou namluvené v češkém jazyce. Řeč je pro člověka nejvýznamnějším komunikačním prostředkem, nejobvykleji realizovaným jakožto zvukový projev. Řečí člověk vyjadřuje svoje myšlenky i (a to nás v této práci velmi zajímá) svoje emoce. Řeč inteligentního člověka je charakterizována jistou akustickou strukturou, lingvistickou strukturou a subjektivním vlivem osobnosti řečníka.[6] Nejmenší stavební jednotkou řeči je foném.

2.1 Tvorba řeči

Pro tvorbu řeči jsou stěžejní tři ústrojí tvořící hlasový trakt - dechové, hlasové a artikulační. Orgány podílející se na tvorbě řeči mají v lidském těle často dosti rozličnou primární úlohu. Hlasový trakt tak spolu utváří orgány sloužící např. k dýchání, příjmu potravy atd.



Obr. 2.1: Ilustrace hlasového traktu [6]

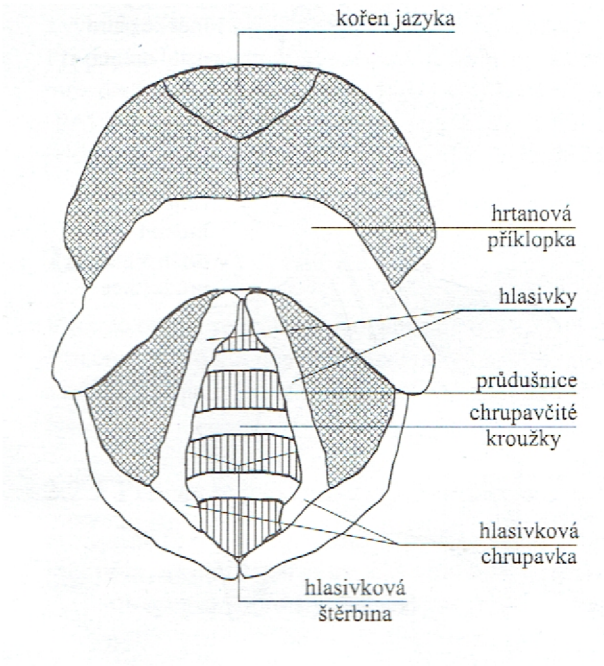
2.1.1 Dechové ústrojí

S určitou nadsázkou lze dechové ústrojí považovat za pohonnou jednotku celého hlasového traktu. Tím nejhlavnějším orgánem v tomto ústrojí jsou plíce, které mají oporu v dýchacích svalech. Kapacita plic dospělého muže je asi 4 - 5 litrů, přičemž asi 1 - 2 litry tvoří tzv. zbytkovou kapacitu plic, která musí být vždy zachována. Vitální kapacita plic je přitom asi 5 litrů.[7] Vydechnutý vzduch je tou potřebnou dávkou energie v procesu tvorby řeči. Obvyklá dávka vzduchu je 0,5 litru v pravidelných intervalech po 3 - 5 sekundách. Výdechový proud vzduchu je odváděn ven průdušnicí, a pak prochází hrtanem a nadhrtanovými dutinami, kde se modifikuje, a jako řečový signál je vyzařován rty do okolního prostoru. Síla s jakou je vzduch vydechován z plic má vliv na sílu hlasu a částečně i na jeho výšku. Pro vytvoření slyšitelné řeči je potřeba během několika sekund vytlačit více než 0,5l vzduchu. [2]

2.1.2 Hlasové ústrojí

V tomto ústrojí se z vydechnutého proudu vzduchu stává to, co nazýváme hlasem. Fundamentální stavební jednotkou tohoto ústrojí jsou hlasivky, které jsou uloženy v hrtanu, konkrétně v hrtanové uzlině za ohryzkem. Hlasivky jsou dvě ostré slizniční řasy, které vedou napříč hrtanem v místě nejužšího průchodu. [7] Typická délka hlasivek je pro muže 15mm a pro ženy 13mm (délka hlasivek určuje charakteristickou hloubku hlasu pro daného člověka). [7] Při vytváření hlasu se hlasivky nacházejí v hlasovém postavení, kdy jsou napnuty. Vydechovaný proud vzduchu prochází z plic až k hrtanu. V hrtanu se do cesty postaví hlasivky, které cestu vzduchu úplně uzavřou. Hlasivky se pod tlakem vzduchu stávají pružnými,začínají kmitat a střídavě se otvírají a uzavírají. V důsledku kmitání hlasivek se ze vzduchového proudu stává vzduchová vlna, kterou vnímáme jako zvuk. [2] Hlas tvoří základní tón a představuje nosný zvuk řeči. Frekvence kmitání hlasivek se označuje F0, je to tzv. fundamentální frekvence (obvyklý rozsah 60 - 400 Hz). Její převrácená hodnota je tzv. perioda základního tónu.[7] Hodnota frekvence základního tónu se uvádí pro muže v rozmezí 80 - 160 Hz. Pro ženy v rozmezí 150 – 300 Hz. Děti mívají tuto hodnotu mezi 200 – 600 Hz. [2] Existují i případy, kdy se hlas dostane nad nebo pod uvedené frekvence (hluboký bas, vysoký soprán). Frekvence základního tónu však není konstantní. V delších řečových úsecích se projevuje vliv intonace promluvy. Délka periody i amplituda jednotlivých pulsů základního hlasivkového tónu se však mírně liší i vrámci velmi krátkého signálu. Takové mírné kolísání délky základní periody se nazývá jitter a je závislé na duševním stavu mluvčího. Kolísání amplitudy hlasivkových pulsů se označuje jako shimmer (rozdíl mezi amplitudami jednotlivých pulsů - udáváno v dB).[7] Při normální promluvě se hodnota změny periody (jitter) pohybuje okolo

1 procenta. Hodnoty, které jsou posluchačem postřehnutelné jsou nad 2 procenty a hodnoty změny amplitudy (shimmer) 1dB. [2]



Obr. 2.2: Hlasivky [6]

2.1.3 Artikulační ústrojí

V tomto ústrojí získává řeč konkrétní srozumitelnou podobu. Ústrojí je složeno z nadhrtanových dutin (dutiny - hrdelní, ústní a nosní) a artikulačních orgánů. Mezi dutinami tvoří hranici čípek, špička měkkého patra zamezující či umožňující přístup vzduchu z dutiny. Dutiny se při procesu nehýbou, čili jejich účast je pasivní. [7] Aktivními, tedy pohyblivými účastníky procesu jsou - jazyk (podle umístění jazyka se mění tvary dutin a tím se vytváří různé zvuky), rty, měkké patro a menší měrou i zuby. [2]Velice zajímavým ústrojím je hrtan, který kromě toho, že se zapojuje do tvorby znělosti, se také pohybuje nahoru a dolů a mění délku proslovu.[2]

2.2 Fonetika českého jazyka

2.2.1 Samohlásky

Při artikulaci samohlásek je snahou udržet průchod vzduchu hlasovým traktem co nejvolnější. Kromě fundamentu se zde vyskytují i zesílené vyšší harmonické (též nazývané formanty) vznikající rezonancí v dutinách hlasového traktu. V češtině jsou

u samohlásek stěžejní první dva formanty. Formant F1 souvisí s otevřeností, F1 má nižší frekvenci u vysokých (zavřených) samohlásek. Formant F2 souvisí s předností/-zadností. Přední samohlásky mají vyšší frekvenci F2 než zadní. Zadnost je však lépe vystihována vzájemným poměrem F1 a F2 než samotným F2. Nutno dodat, že u každé samohlásky lze vypozorovat pět i více formantů, avšak na detekci samohlásky bezpečně stačí první dva. První tři formanty jsou vlastně rezonanční kmitočty největších dutin hlasového traktu. F1 - dutina hrdelní, F2 - dutina ústní, F3 - dutina nosní. [8] Jejich výška a intenzita je závislá na anatomických parametrech ústní, ale i hrdelní dutiny. Změnu výšky a intenzity pak zajišťují rty, čelist a měkké patro. [6] Základní artikulaci vokálů vytváří pohyb jazyka z neutrálního postavení (obdobného jako při klidném dýchání) dopředu a vzhůru nebo dozadu a vzhůru. Tímto posunem se mění poměr objemů dutiny ústní a hrdelní.[9] Postavení jazyka v ústech ukazuje tradiční obecné schéma soustavy vokálů jazyku, vokalický trojúhelník. Jeho východiskem je trojúhelníkový obrazec vytvořený r. 1781 Ch. F. Hellwagem. Obrazec vychází z vertikálního a horizontálního posunu jazyka z klidové polohy při tvoření jednotlivých samohlásek.

2.2.2 Souhlásky

Narozdíl od samohlásek je u souhlásek ve spektru přítomný šum. Souhlásky jsou vytvářeny vzduchovou turbulencí, která vzniká třením výdechového proudu vzduchu o překážku vytvořenou artikulačními orgány. Překážka může být úplná (závěr - typické u šumových souhlásek) nebo částečná (úžina). [6] Existue také speciální hybridní typ souhlásek tzv. polozávěrový, kdy závěr je velmi slabý a přechází hned do úžiny [9] Souhlásky se potom dále člení podle místa tvoření (retná, zubná atd.). [9]

3 Hudba

V této kapitole se ponoříme do problematiky emociálního prožitku vyvolaného poslechem hudby. Zaměříme se na některé obecné charakteristiky emocí v hudbě a také neopomeneme subjektivitu klasifikace emocí v hudbě na základě typu posluchače. Hudba je organizovaný systém zvuků produkovaný člověkem a určený pro lidské vnímání; oproti řeči, která slouží především k dorozumívání, je cílem hudby především estetické působení. Kvalitu, funkci a estetické působení hudby určuje výběr zvuků (v člověkem slyšitelném rozsahu), jejich rytmické členění a jejich uspořádání.[10] Dle historického poznání hudba jako taková vznikla dříve než písmo. Asi není velkým překvapením, že prvním hudebním nástrojem v dějinách byl lidský hlas.[11] Má to jednu zjevnou výhodu – lze lépe sledovat kontinuitu od prapočátků až po dnešek, protože hlasové ústrojí se na rozdíl od hudebních nástrojů zásadně nezměnilo.[11] Podle některých teorií se hudba vyvinula z řeči, oddělením prozodických prvků od syntaktických.[11] Teorií však existuje celá řada a ve vědecké obci není přesná shoda na tom, jaký skutečný původ hudba má.

3.1 Emoce v hudbě

O emocích v hudbě se lze bavit ze dvou hledisek. Hudební dílo se může určitou emoci snažit cíleně vyjadřovat (exprese) nebo také může hudební dílo danou emoci u posluchače vyvolávat (indukce). Zatímco první situaci je možné na základě mnohých objektivních parametrů kategorizovat, v druhém případě se vše vesměs odvíjí od subjektivního úsudku posluchače. Prvotní dochované zaznamenané úvahy o hudbě jakožto nositelce emocí pochází z antického Řecka. Řekové tehdejší doby hudbě přisuzovali velký společenský význam. Tehdy byly položeny základy hudební teorie. Nelze také opomenout kupříkladu objev významného řeckého matematika a fyzika Pythagorase, který objevil číselný poměr mezi jednotlivými tóny. Řekové posuzovali náladu a emoce v hudbě podle užívaných modů (modus je ekvivalentem toho, co v moderní hudební teorii nazýváme stupnicí). Platón byl zarytým zastáncem myšlenky, že by se měl používat pouze dórský nebo frygický modus. Vojákům hudba v dórském a frygickém modu dodá sílu, zatímco ostatní mody by je obměkčovaly. Platón byl dokonce přesvědčen, že změna modů státu by způsobila sociální revoluci.[11] Klasifikování emocí v hudebním dílu na základě stupnice, kterou využívá melodie bylo zcela běžným jevem. Typické dogma používané i v dnešní době nám káže, že mollová stupnice značí něco smutného zatímco stupnice durová značí něco veselého. Významy však nebývaly přiřazovány pouze a základě využívaných modů či stupnic. V dobách středověku bylo zakázáno používat v hudbě interval velkou kvartu též nazývanou tritón. Bylo to považováno za “ďáblův” prvek v hudbě. Dnes je tento

interval hojně využívaným prostředkem v metalové hudbě, kde skvěle funguje pro evokování temné atmosféry.V odborných kruzích však nepanuje ani shoda v tom, zda jsou emoce během poslechu skutečně prožívány. [11]

3.1.1 Emoce v podobě exprese

Výše již byla uvedena určitá dogmata přiřazující určitým hudebním postupům určité emoce. Jakým způsobem však definovat v hudbě emoce na základě změřitelných parametrů? Budu pracovat se třemi základními parametry. Tonalitou, tempem a dynamikou.

Tonalita

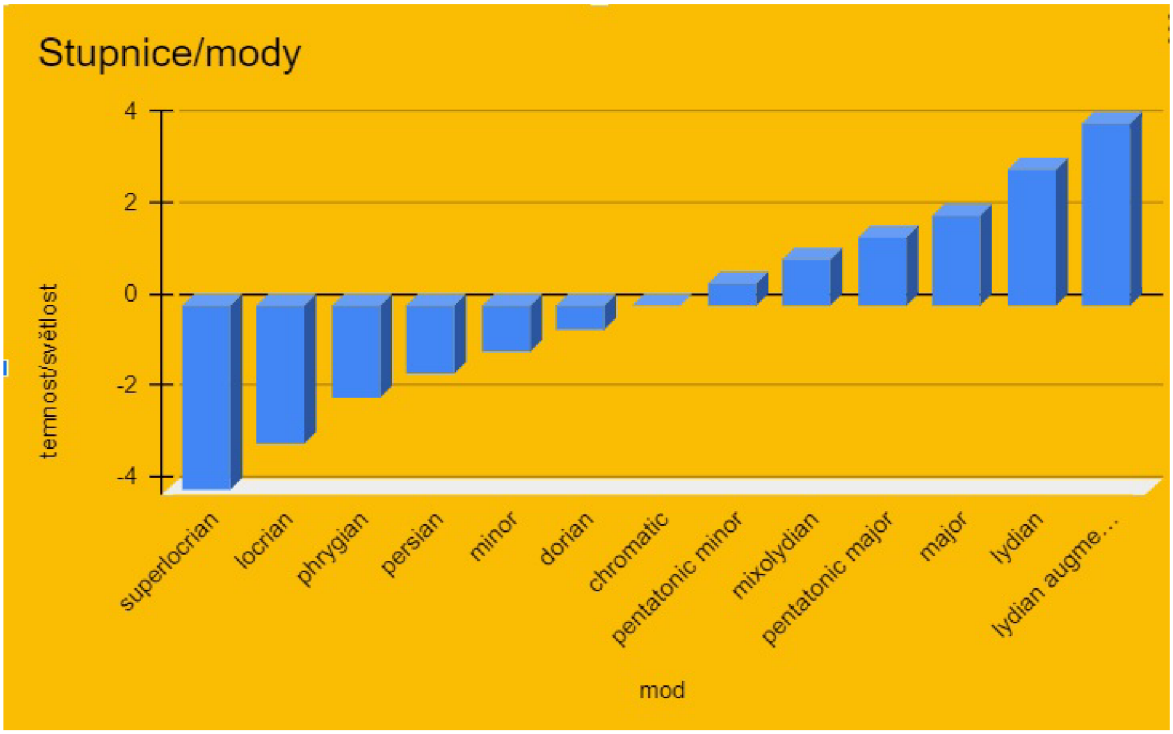
Tonalita je v hudbě charakterizována za pomocí tóniny, konkrétněji poté za pomoci stupnice či modu. Všechny stupnice či mody lze seřadit dle světlosti či tmavosti. Můžeme si představit škálu jdoucí od černé barvy přes odstíny šedé barvy až po barvu bílou, kde černá barva vyjadřuje nejhlubší smutek a bílá barva chcarakterizuje tu nejryzejší radost. Vyšel jsem z konceptu, který na svých přednáškách prezentoval skladatel, dirigent a slavný populizátor hudby Leonard Bernstein. Původní posloupnost 9 modů (superlokrický, lokrický, frygický, mollový, dórský, mixolydický, mollový, durový, lydický a lydický rozšířený) jsem doplnil o perskou stupnici, chromatickou stupnici (mající především referenční charakter k neutralitě) a o durovou a mollovou pentatoniku (hojně využívána v bluesové a rockové hudbě). Jednotlivým stupnicím/-modům jsem přiřadil koeficienty dle tmavosti/světlosti. Záporné hodnoty hodnoty koeficientů vyjadřují příslušnost k tmavším stupnicím/modům, zatímco koeficienty kladné příslušnost k stupnicím/modům světlejším. Uprostřed na hodnotě 0 se nachází stupnice chromatická.

Tempo

Významným parametrem k posouzení emoce v expresi hudebního díla je tempo. Pro určité emoce může být tempo určující, jiné emoce mohou být na tempu do určité míry zcela nezávislé. Obecně však lze říci, že vztek a štěstí si většina z nás asociuje s rychlejšími tempy, zatímco smutek či nudu s pomalejšími.

Dynamika

Mnohem klíčovější než dynamika samotná jsou její změny. Lze usuzovat, že takový vztek si většinou spojíme s dynamikou forte až fortissimo. Dynamiku v hudební nahrávce dokážou poměrně spolehlivě reflektovat veličiny spojené s energií.



Obr. 3.1: Temnost/světlost stupnic

3.1.2 Emoce v podobě indukce

Zde se bavíme o čistě subjektivních prožitcích jednotlivých posluchačů. Rozlišujeme více typů posluchačů.

Typy posluchačů

O přesnější kategorizování posluchačů hudby se pokusil muzikolog Th. W. Adorno. Jeho typologie není pouze psychologická, ale zahrnuje i prvky sociologické. [12] 1)hudební expert - profesionální muzikant, co při poslechu přemýšlí nad všemi souvislostmi 2)dobrý posluchsč - člověk neznalý hudební teorie, avšak hudbu vnímající téměř stejně sofistikovaně jako hudební expert. 3)vzdělaný posluchač - v teorii vzdělaný jedinec, co však poslouchanou hudbu vnímá více pocitově 4)emocionální posluchač - hudbu vnímá silně citově 5)posluchač hudby pro zábavu - nejběžnější typ 6)lhostejný jedinec - hudba mu je zcela ukradenná

4 Analýza a zpracování řečového signálu

V této kapitole se budeme opět věnovat řeči, tentokráte z více technického úhlu pohledu. Bude řeč o tom, co vše lze za parametry ze zaznamenaného řečového signálu vyčíst a jakým způsobem lze dále řečový signál zpracovávat. V kapitole Řeč jsem již lehce zavadil o některé termíny, které v této kapitole budou hrát významnou roli. Věnovali jsme se již fonetice, konkrétně fonetice českého jazyka. Tato kapitola se bude z velké části zabývat možnými zobrazeními řečových signálů a tím, co vše z nich lze vyčíst, což je disciplína pro pochopení problematiky této práce zcela zásadní.

4.1 Disciplíny potřebné pro komplexní analýzu řeči

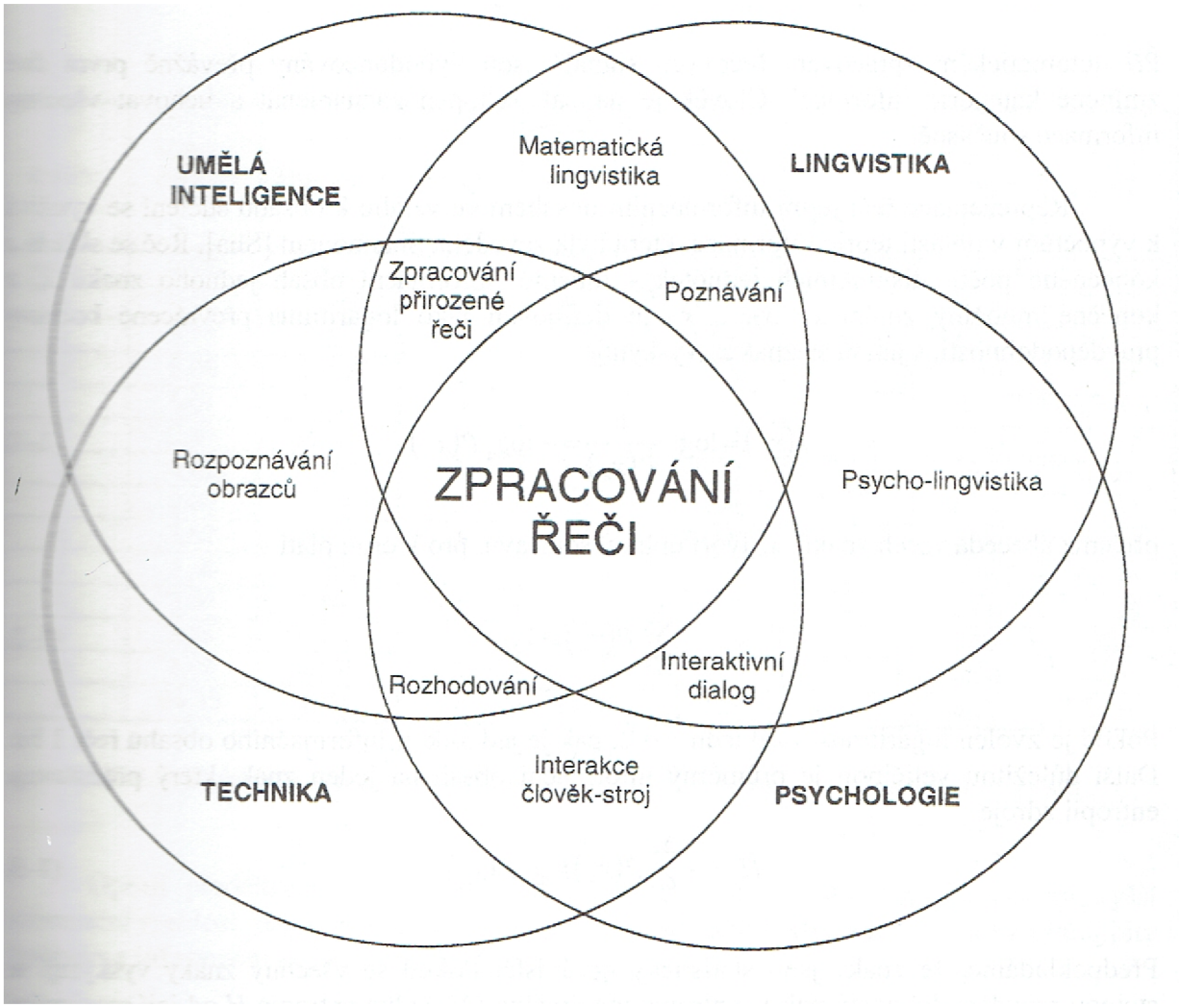
Jak je již zmíněno, problematika vyžaduje znalosti z mnoha oborů, tyto obory se při analýze řeči velmi často prolínají.[8] Z toho důvodu vlastně není příliš překvapivé, že mezi používanými materiály pro zpracování této práce, se vyskytují díla a práce napříč širokým spektrem oborů přes pedagogiku, psychologii až po elektrotechniku.

4.1.1 Výčet disciplín

Fonetika - Fonetikou jsme se zaobírali již ve druhé kapitole. Součástí disciplíny jsou znalosti o tvorbě řeči, definice inventáře dostupných hlásek a popis jejich artikulační a akustické vlastnosti.[8] **Fonologie** - Tato disciplína se zabývá hláskami z pohledu jejich výskytu a možnosti utváření kombinací. [8] **Prozodie** - Pro analýzu emocí v mluveném projevu zcela stěžejní disciplína. Na samotnou disciplínu se ještě blíže podíváme. Je zaměřena na zvukovou stránku jazyka, určování melodie a přízvuku v jednotlivých slovech. [8] **Lexikologie** - Nauka zaměřená na jazykovou stránku problematiky konkrétně se zabývající jednotlivými slovy. [8] **Syntax** - Nauka o větné skladbě [8] **Sémantika** - Určování významu slov pro aplikaci při vhodnému výběru slov ze seznamu. [8]

4.2 Znázornění řečových signálů

Řečový signál lze zobrazovat mnoha způsoby v závislosti na tom, jaké konkrétní parametry řečového signálu chceme zkoumat. My se podíváme na ty pro nás nejpodstatnější.



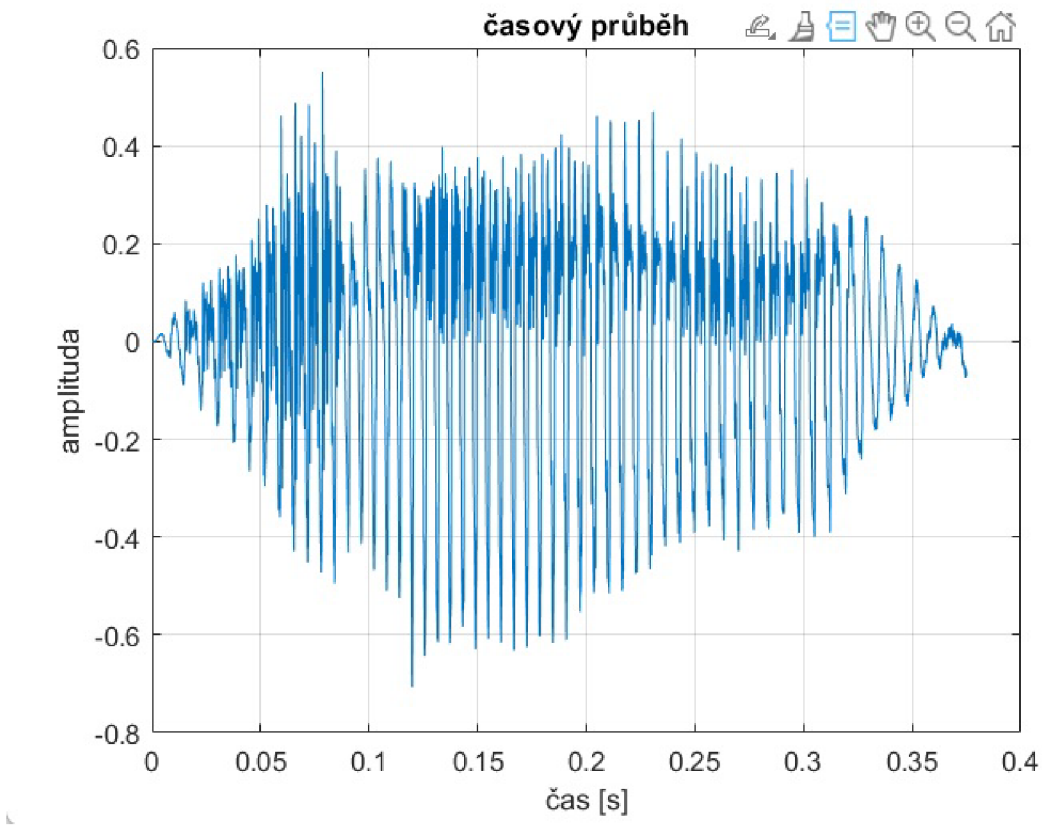
Obr. 4.1: Disciplíny potřebné v analýze řeči [8]

4.2.1 Časový průběh

Zřejmě jako první se většině lidí ze všech možných zobrazení vybaví časový průběh. Z průběhu lze vyčíst tvar signálu, velikost amplitud, pořípadě délka periody. Když budeme například detailně zkoumat změny délek jednotlivých period, budeme moct odhadnout, jak moc velký jitter je v řeči přítomen. Dále když budeme detailně zkoumat proměnlivost amplitud, budeme moct odhadnout míru shimmeru. S pomocí mnohých matematických operací se pak z průběhu dá vyčíst mnohem více.

4.2.2 Spojité kmitočtové spektrum

Obecně se jedná o vztah intenzity na frekvenci. Tento průběh získáme Fourierovou transformací časového průběhu. V tomhle zobrazení lze vyčíst časové změny



Obr. 4.2: Časový průběh signálu

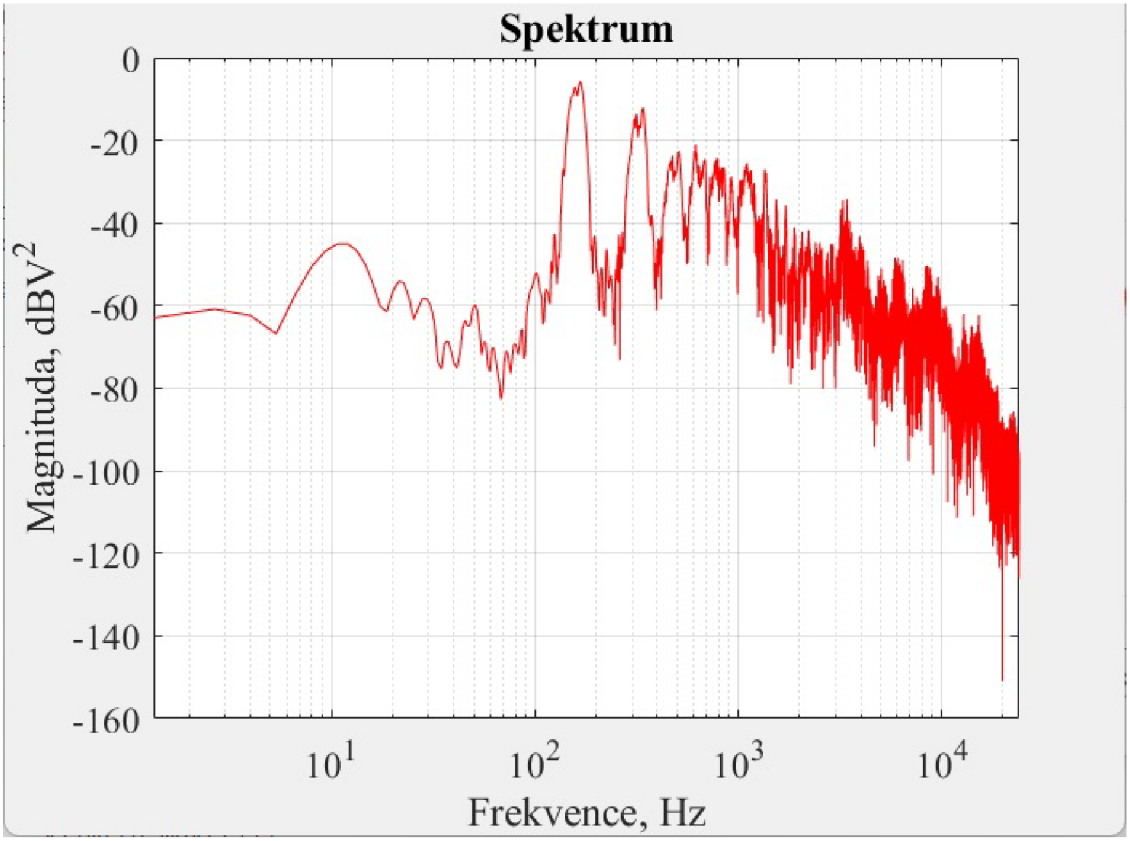
jednotlivých formantů.

4.2.3 Spektrogram

Skvělým pomocníkem pro zkoumání akustických vlastností řečového signálu je spektrogram. Pracuje se třemi parametry - čas, frekvence a intenzita. Může být realizován ve 3D provedení, ale pro zjednodušení bývá nejčastěji využívána 2D verze, kde je problém většího množství parametrů vyřešen analogií s barvami, kdy určitá barva vždy znázorňuje danou hodnotu intenzity.Spektrogramy mají výrazné využití ve zkoumání barvy zvuku.

4.3 Příznaky řečového signálu

Po převedení signálu z analogové podoby do podoby číslicové lze z navzorkovaného signálu vyčíst mnoho důležitých parametrů. Z pohledu matematického lze považovat digitalizovanou řeč za stochastický (náhodný) proces. Bohužel, tento signál obsahuje



Obr. 4.3: Spojité kmitočtové spektrum

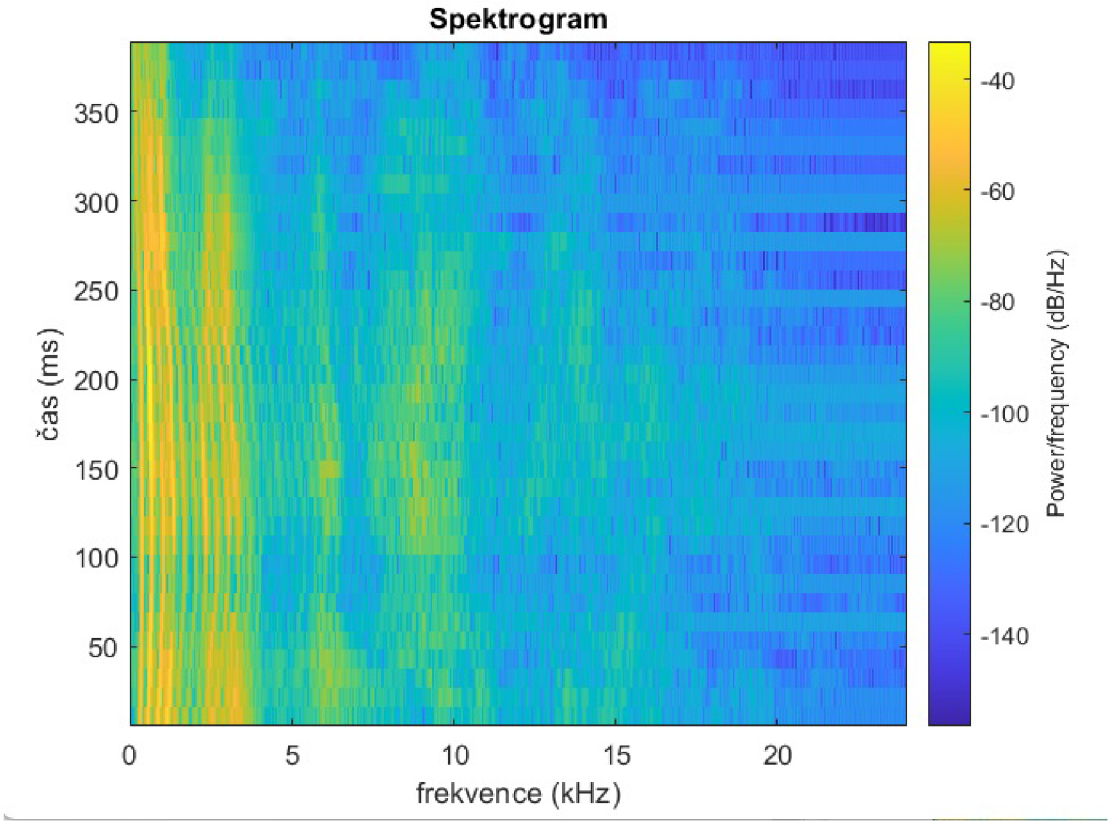
mnohé pro nás irelevantní informace, které z celého souboru informací je nutné vyloučit. [8]

4.3.1 Energie

Energie je definovaná jako suma čtverce hodnot signálu. Tento parametr je velmi užitečný při rozlišování jednotlivých hlásek.[8]

4.3.2 Počet průchodů nulou

U navzorkovaných signálů nastane průchod nulou, pokud mají dva sousední vzorky různé znaménko. Z počtu těchto průchodů lze odhadnout rozložení energie v kmitočtovém spektru. [8] Dle tohoto parametru lze například analyzovat v jaké části nahrávky se mluví a v jaké nikoliv, což pomáhá úpravě délky souboru, aby poté neobsahoval nepotřebná data.



Obr. 4.4: Spektogram

4.3.3 Pásmová filtrace

Metoda při níž se za pomocí filtrace realizované bankou filtrů zjišťují potřebné veličiny z vybrané kmitočtové oblasti. Lidské ucho zpracovává řeč podobným postupem jako analýza bankou filtrů. I vzhledem k této podobnosti se často používá banka filtrů s logaritmickými pásmy. [8] Filtrace je stěžejní například pro výpočet mel-kepstrálních koeficientů, protože každý koeficient reprezentuje jednu z frekvenčních oblastí.

4.3.4 LPC

Lineární predikce, známá pod zkratkou LPC (linear predictive coding), je jednou z nejpoužívanějších metod analýzy řečového signálu. Její princip spočívá v předpovídání n-tého vzorku řečového signálu pomocí lineární kombinace vzorků předcházejících. Vztah pro hodnotu signálu je následující:

$$\sum_{m=1}^{M} a_m s(\mathrm{n} - \mathrm{m})$$

Za pomoci součtu M předchozích prvků téhož signálu a vynásobením příslušným koeficientem (je nutné správné nastavení koeficientů) získáme následující hodnotu signálu.[8]

4.3.5 Autokorelační koeficienty

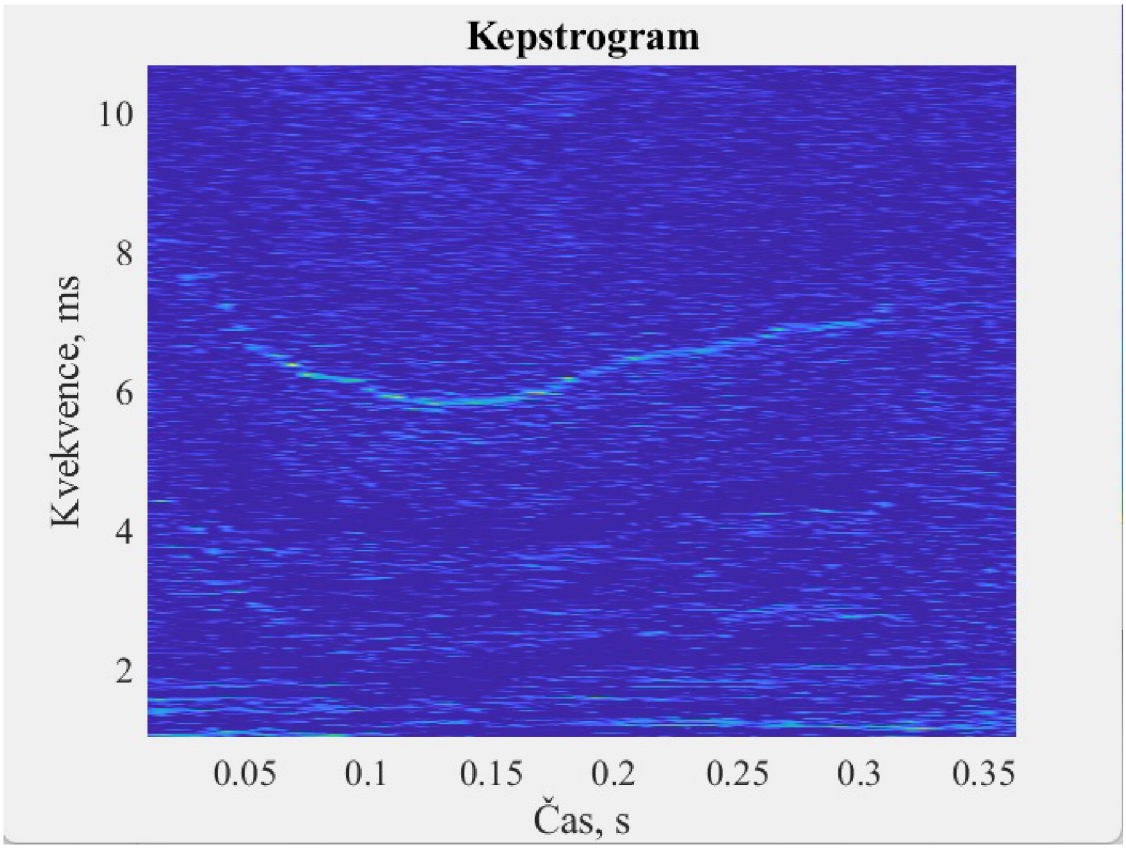
Nejdůležitější metoda analýzy řečového signálu je autokorelační funkce známá pod zkratkou AKF.

$$R(k) = \sum_{n=-\infty}^{+\infty} s(n)s(n+k)$$

Je zde posuzována shoda signálu se stejným signálem posunutým o k vzorků. [8]

4.3.6 MFCC

Z anglického mel-frequency cepstral coefficients. Práce s celým spektrem je vysoce neúčelná, spektrum totiž obsahuje mnoho pro nás zcela nedůležitých informací. Proto bylo nutné vymyslet vhodnější zobrazení signálu, kde by se již pracovalo pouze s užitečnými a s ničím nekorelujícími parametry. Z čistě matematického hlediska je kepstrum výsledek následujících operací - přenesení časového průběhu přes Fourierovu transformaci do frekvenční domény - výpočet logaritmu ze spektrálních amplitud - transformace do kvefrenční domény. Kvefrence je specifiální časová veličina popisující výkonovou spektrální hustotu. Metoda při níž jsou odděleny parametry buzení a hlasového ústrojí [13] Každou harmonickou složku v signálu lze chápat jako součin těchto dvou složek. Když tento vztah zlogaritmujeme, tak namísto součinu zíkáme součet. To je poté mnohem praktičtější pro jejich následné oddělení. [13] Klasická kepstrální analýza nebere v potaz subjektivitu vnímaní lidským uchem. Kýženým výsledkem mel-kepstrální analýzy jsou koeficienty (obvykle 13) popisující spektrum, kdy každý z koeficientů popisuje určitou frekvenční oblast. Výsledkem použití funkce MFCC v Matlabu je matice o 13 sloupcích a různém počtu řádků odvíjejících se od délky zpracovaného signálu. Výpočty koeficientů jsou totiž prováděny vždy pro krátký časový segment. To, jak se koeficienty v čase mění, popisuje speciální barevný graf. Z matice lze poté udělat vektor zprůměrováním těchto hodnot.



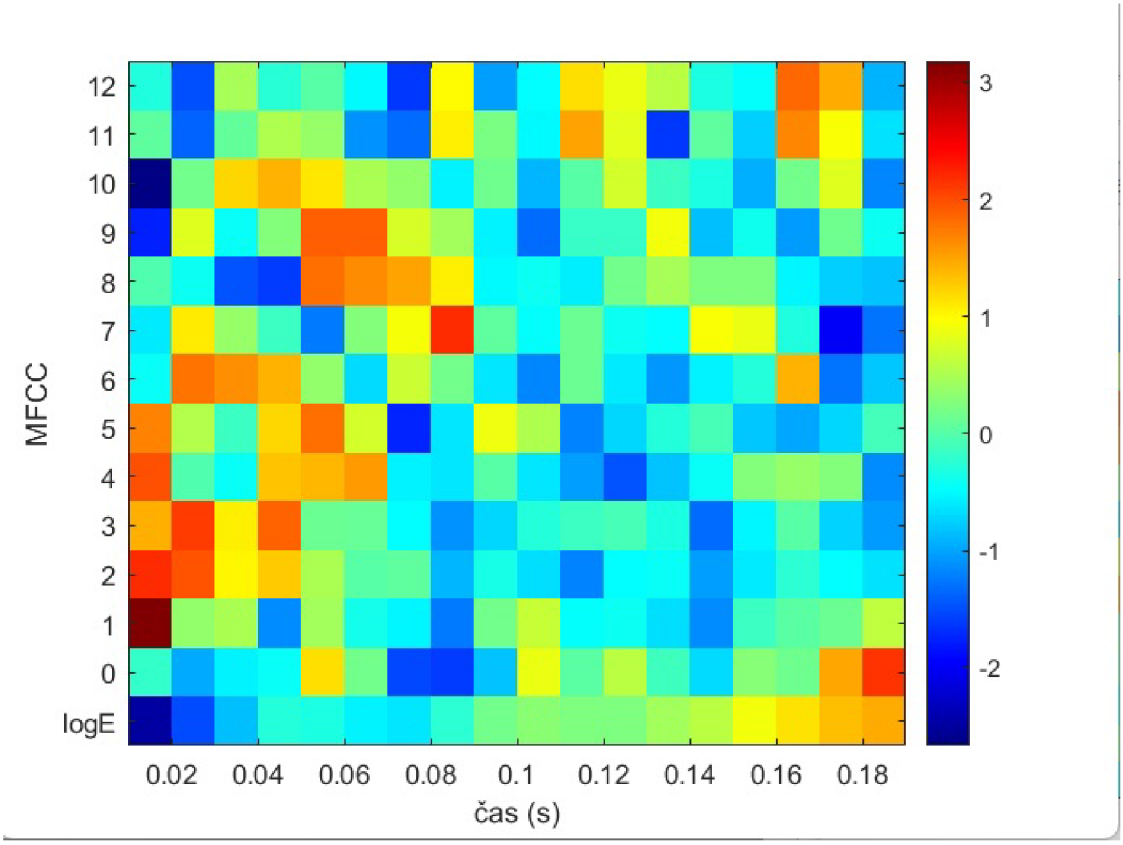
Obr. 4.5: Kepstrogram

4.4 Prozodické informace v řečovém signálu

Zatímco většina disciplín z oblasti analýzy řečového signálu si láme hlavu nad obsahem sdělení, prozodie se prioritně zabývá způsobem, jakým je daná věc řečena. To je při klasifikaci emocí v řeči mnohem důležitější než se zabývat slovy, která byla vyřčena.

4.4.1 Frekvence základního tónu

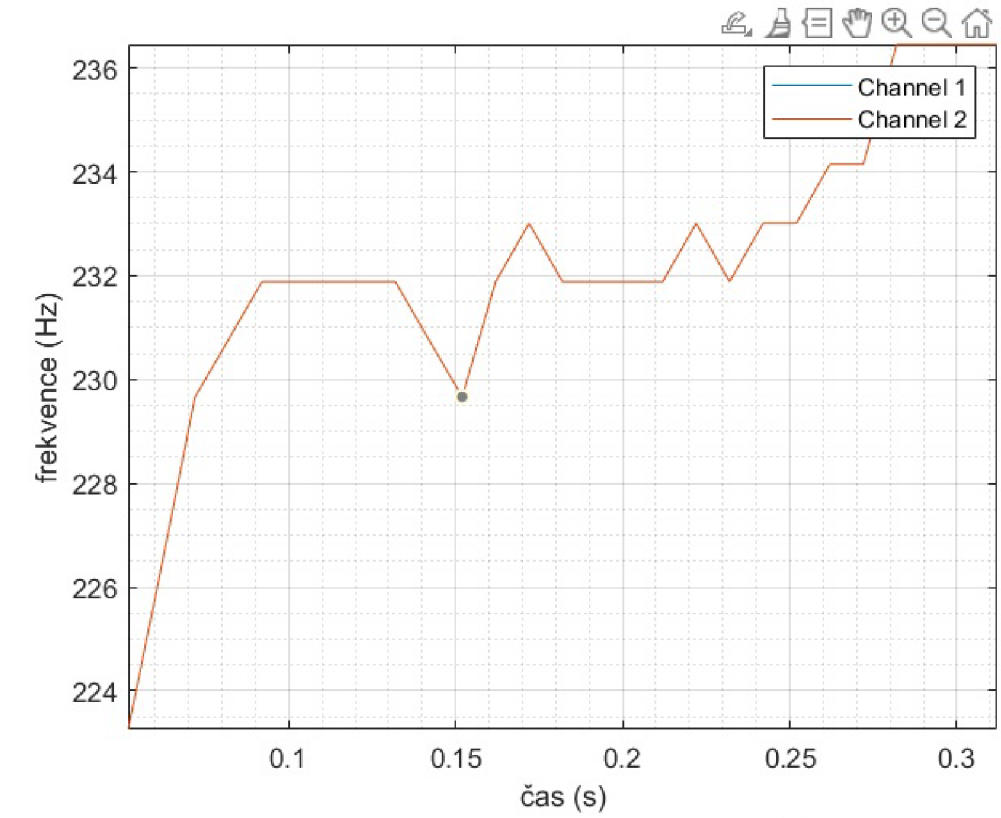
Fundamentální kmitočet, na kterém mluvčímu kmitají hlasivky.[2]Jak je již zmíněno v předcházejících stránkách, tato frekvence bývá proměnlivá. Míra změn této frekvence určuje intonaci. Výrazně to lze sledovat na víceslabičných slovech či delších větných celcích. Určité možnosti, jak si s touto proměnlivostí poradit, představím v praktické části práce.



Obr. 4.6: MFCC

4.4.2 Energie

Můžeme defacto vnímat jako ukazatel hlasitosti. Obvykle počítáno jako RMS z vektoru reprezentujícího číslicově zpracovaný signál. V práci bude kladen důraz i na zohlednění faktu, že vliv na celkovou energii budou mít i podmínky, za jakých byly pořizovány nahrávky, které se u mnohých nahrávaných osob lišily. Z toho důvodu si pomůžeme hodnotami pro neutrální promluvy a bude nás zajímat, jak moc se v dané emoci energie liší oproti emoci neutrální.



Obr. 4.7: Fundament v čase

5 Analýza a zpracování hudebního signálu

Tato kapitola bude svou obsahovou stránkou dosti podobná kapitole předchozí s tím rozdílem, že místo řečového signálu bude řeč o signálu hudebním. Vědecký pokrok nám totiž dává do rukou účinný nástroj v podobě číslicového zpracování signálu, jenž stejně tak, jak jsme to udělali u řeči, můžeme využít i u hudby. [14] I zde je využíváno metod, které jsme si představili v kapitole o analýze a zpracování řeči.

5.1 Zvukové parametry využívané při analýze hudební nahrávky

Představíme si nejpodstatnější parametry, jež jsou předmětem zkoumání v analýze hudebního signálu. Některé parametry nám jsou již známé z kapitoly předchozí, s některými se seznámíme v této kapitole nově.

5.1.1 Změny spektra

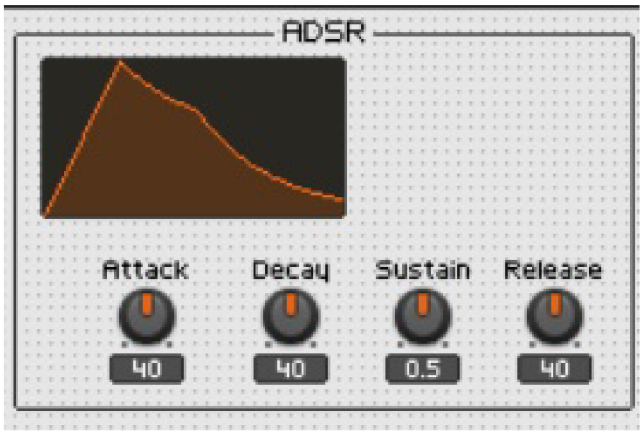
Typickým projevem změny spektra může být třeba změna barvy tónu hudebního nástroje při jeho nasazení. [14] Veličin, které tyto změny spektra umí popsat, je více. Využít lze MFCC, LPC, či autokorelaci. Obrovské využití veličin popisujících změny ve spektru má například klasifikace hudebních žánrů.

5.1.2 Tvar energetické obálky

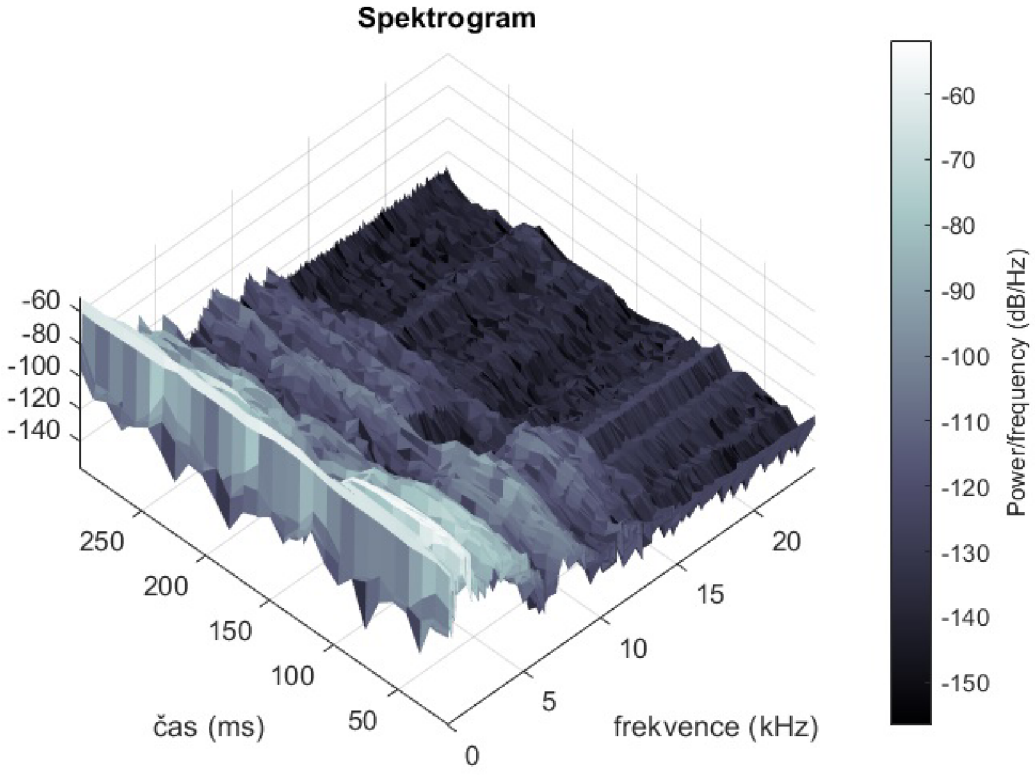
Pojem známý pod zkratkou ADSR (attack, decay, sustain, release). Parametr popisuje průběh energetického nárůstu a následného poklesu při nasazení tónu. Celý tento proces má 4 fáze - attack - prudký nárůst energie na začátku, decay - mírný pokles po prvotním náporu, sustain - ustálení a udržování stejné energie, release - pokles k nule. Každý nástroj má svůj zcela unikátní a charakteristický tvar obálky. Obvykle se ze zkoumaného signálu extrahují časové hodnoty pro jednotlivé fáze. Ve většině případů se analýze podrobují hodnoty pro fázi attack.

5.1.3 Spektrogram

O spektogramu již byla řeč v minulé kapitole, ale pro zopakování - zobrazení intenzity jednotlivých frekvencí obsažených ve spektru umožňují lépe určit pozice tónů i v polyfonicky složitějším zvuku.[14] Spektrogram graficky znázorňuje změny intenzit jednotlivých frekvenčních frekvencí v čase.



Obr. 5.1: Tvar obálky



Obr. 5.2: 3D Spektrogram

6 Umělé neuronové sítě a jejch aplikace

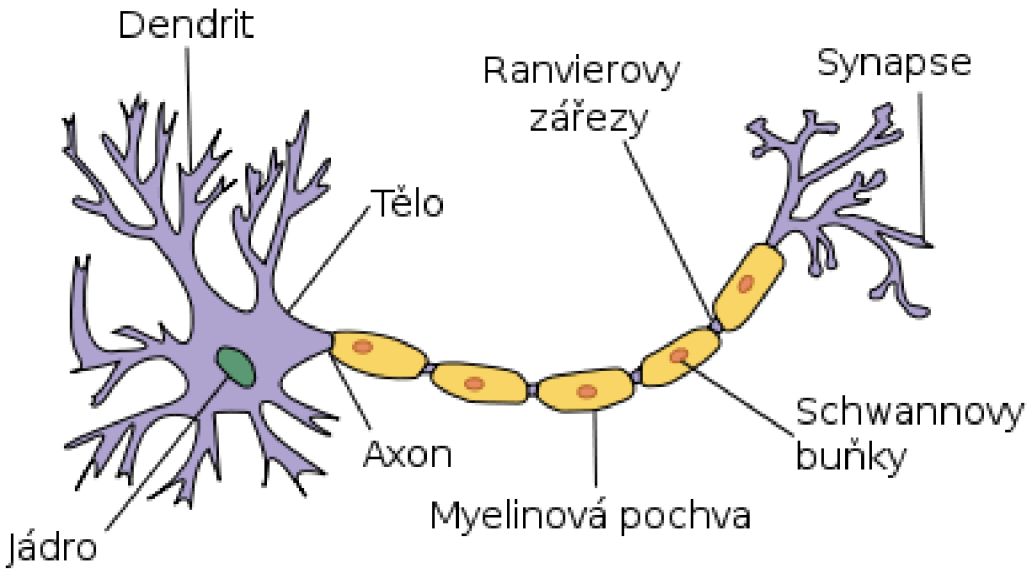
Cílem kapitoly bude vysvětlit základní koncepty umělých neuronových sítí a jejich následné využití v praxi i s ohledem na využití v bakalářské práci, která bude navazovat na tuto semestrální práci. Umělé neuronové sítě - UNS (v angličtině artifical neural networks - ANN) jsou matematickým modelem biologických neuronových sítí. Samotný pojem umělých neuronových sítí nelze jen tak sloučit pod jednotnou obecnou definici, nicméně si je můžeme představit jako soustavu mnoha jednoduchých procesorů, z nichž každý má několik vstupů a jeden výstup. [15]

6.1 Biologické neuronové sítě

Pro lepší pochopení umělých neuronových sítí je ideální pochopit to, v čem byla nalezena inspirace pro jejich vznik. Jsou zde samozřejmě myšleny biologické neuronové sítě, které jsou základem všech částí biologického informačního systému, jehož základní jedotkou je neuron.[15] Neuron je s trochou nadsázky takový kurýr rozvážející zásilky plné informací. Existence informačního systému je nezbytnou nutností pro existenci a přežití všech systémů. Nervový systém a mozek člověka tvoří jednu z nejsložitějších soustav vůbec.[15] Samotný neuron je složen z těla s buněčným jádrem nazývaným soma. [15] Na tělo jsou napojeny výběžky dvojího typu - buďto krátké a dostředivé dendrity (jejich matematický ekvivalent jsou váhy) nebo dlouhé a odstředivé axony.[16] Dendrity tvoří vstup do neuronu, axony zase tvoří výstup.[15] Axon je pouze jeden. Zde už si můžeme všimnout té spojitosti s umělou neuronovou sítí, kde jsme se bavili o tom, že pomyslné jednotky tvořící síť mají mnoho vstupů a pouze jeden výstup. Axon se stýka s tzv. Ranvierovými zářezy. Axony jsou ukončeny synapsemi. Průchodnost synapsí je proměnná. Průchodnost synapsí se mění například při učení. [15]

6.2 Základní pojmy

Je nutné se seznámit s některými pojmy souvisejicími s umělými neuronovými sítěmi. Zde se nevyhneme ani některým matematickým pojmům. K posouzení časových a prostorových vztahů mezi vzorky signálu a k vyhodnocení kvalitativních i kvantitativních vazeb je nezbytné využít matematických vztahů.Budou nás zajímat statistické vlastnosti jednotlivých vzorků[17] Nemálo dějů v přírodě lze popsat signálem časově a prostorově proměnným. Časové a prostorové vazby sousedních hodnot signálů lze popsat pomocí vektorů či matic. [17] Grafické zobrazení může vypadat následovně.



Obr. 6.1: Anatomie neuronu [16]

Při trénování Umělé neuronové sítě hraje významnou roli metrika,neboť budeme řešit rozdíly mezi jednotlivými vzorky reprezentovanými vektory. Míru tohoto rozdílu nazýváme vzdálenost. [17] Nejčastěji se zde uplatňují tři následující matematické definice - Euklideova vzdálenost, Minkowskiho vzdálenost a Hammingova vzdálenost.

6.2.1 Euklideova vzdálenost

Mějme dva n-rozměrné vektory x a y. Euklideova vzdálenost je definována jako odmocnina součtů čtverců rozdílů prvků dané dimenze. Bývá používaná nejčastěji.

6.2.2 Minkowskiho vzdálenost

Jedná se o zobecněnou obdobu Euklideovy vzdálenosti. Využívána je například v Kohonenově kubické neuronové síti.

6.2.3 Hammingova vzdálenost

Využívána pro určování vzdálenosti u binárních vektorů (tudíž takových vektrů, jejichž souřadnice mohou nabývat pouze dvou možných hodnot). Nicméně její využití není omezené pouze na tyto případy. Lze obecně použít všude, kde jsou možné prvky diskrétně vymezeny (např. písmena abecedy, seznam studentů atd.).

6.3 Učení neuronových sítí

Stejně jako člověk se stává schopnějším i na základě dovedností a zkušeností, tak i umělá neuronová síť musí absolvovat proces při němž se poté dostane do stádia, kdy nám začne dávat přesné a validní výsledky. Způsobů jak učit (chcete-li trénovat) umělou neuronovou síť je vícero. Musíme si odpovědět na následující tři otázky - asociativní či neasociativní? , s učitelem nebo bez učitele? , jednorázově či opakovaně?

6.3.1 Samoorganizující se neuronové sítě

Za život se k nám dostáva nepřeberné množství informací, které si náš mozek třídí na ty užitečné a ty neužitečné. Samoorganizující se neuronové sítě se snaží o to samé. Z vícedimenzionálních vstupních dat jsou vybrána ta data, která se jeví jako nejdůležitější přenositel informace. Funkce samoorganizující se mapy je založena na soutěžním učení, kdy všechny neurony přijímají stejná data a pouze jeden nejpodobnější neuron se stává vítězem. Hodnoty kolem vítězného neuronu mívají podobné hodnoty a tvoří clustery (shluky). [2] Samoorganizující se mapy jsou užitečné zejména pro analýzu velkého množství vstupních parametrů. [15]

6.3.2 Kohonenovy mapy

Finský vědec Tuevo Kohonen přišel s konceptem samoorganizujících se neuronových sítí učících se bez učitele, což znamená, že neuronové síti nejsou předhazována data s očekávatelnými parametry. [18] Poloha neuronu nejpodobnějšího vstupnímu vzoru se vždy změní tak, aby byl neuron co nejblíže. Neurony si můžeme v této situaci představit jako koule vzájemně propojené pružnou sítí. Změna polohy jednoho neuronu se tak automaticky projeví i na polohách neuronů ostatních, a protože na vstup po krocích přichází stále víc a víc vzorů, začínají se hýbat sami z vlastní vůle i další neurony.

7 Praktická část bakalářské práce

Úkolem praktické části bakalářské práce bylo vytvořit databázi nahrávek řeči a hudby, určit vhodné parametry pro klasifikaci a na závěr porovnat výsledky tréninku a klasifikace neuronových sítí s výsledky poslechových testů.

7.1 Vytvoření databáze řeči

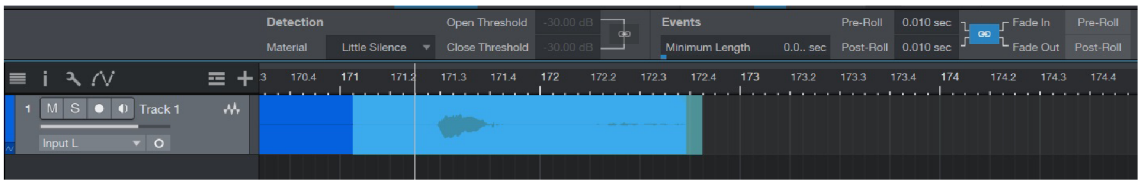
Nahrávání do řečové databáze se zúčastnilo 10 lidí různého věku a pohlaví. Každý řečník pronesl šest slov v šesti různých emocích. Celkově bylo vytvořeno 360 zvukových souborů zaznamenávajících emoční řeč. Pro nahrávání 4 z 10 jedinců byl použit kondenzátorový mikrofon Scarlett Cm25 Mkiii. Zbylé nahrávky byly pořizovány na dynamický mikrofon Shure SM58. Zvukovou kartou použitou pro nahrávání byl Focusrite Scarlett třetí generace mající jeden XLR vstup s možností zapnutí fantomového napájení a jeden vstup pro nástrojový jack s nastavitelnou impedancí pro přímé nahrávání hudebních nástrojů do linky. Použitým DAW softwarem byl PreSonus Studio One, v kterém byly nahrávky také stříhány.

7.1.1 Nahrávání řečníků

Databáze nahrávek řeči je tvořena neherci a dané emoce obsažené v nahrávkách se nahrávaní jedinci snažili co nejvěrohodněji napodobit.Při nahrávání řečníků jsem plnil roli režiséra (s výjimkou 36 vzorků, kde jsem nahrán já sám). Nahrávalo se po krátkých sekvencích, kdy byl řečník obeznámen s tím, jaké slovo má vyslovit, a já vždy řečníkovi řekl, jakou emoci má předvést. Každý z účastníků nahrávání si poté svoje krátké segmenty řeči poslechl a sám musel vyhodnotit, zda je s touto podobou nahrávek spokojen. V případě nespokojenosti nahrávaného řečníka se nahrávání nepovedených částí opakovalo. Bohužel ne všechny nahrávky bylo možné pořídit v akusticky suchém prostředí. Ve většině případů jsem musel s nahrávací výbavou cestovat a vytvářet nahrávací stanici na různých místech. Režie a nahrávácí prostor se nacházely v jediné místnosti. Obvyklé schéma bylo následující: dlouhý stůl, na jedné straně byla režie, na druhé straně stolu byl řečník stojící u mikrofonu.

7.1.2 Střih nahrávek

Nahrávky byly stříhány v DAW. Při střihu jsem se snažil citlivě odmazat tiché pasáže, ale zároveň jsem nechal vždy krátký prostor na začátku a konci vyřčeného slova. Dobrou pomocí při tomto procesu byl nástroj používány na triggering a oddělení tichých pasáží s nastavitelným pre-rollem a post-rollem. Každé slovo bylo poté vyexportováno jako samostatný wav soubor.



Obr. 7.1: Úpravy v DAW Studio One

7.2 Vytvoření hudební databáze

Databáze byla vytvořena z mých autorských děl, či děl vzniklých v rámci výuky předmětu Vybavení studia. Z hudebních děl jsem vystřihl krátké sekvence, které jsou po stránce hudebního tématu a struktury jednotné. Všechny hudební ukázky byly nahrávány při vzorkovací frekvenci 48 kHz. Taktéž bylo využito DAW Studio One. Databáze je tvořena 120 ukázkami. Některé z ukázek v databázi byly vzaty z mých již dávno vytvořených děl, jiné jsem zase pro účely této práce speciálně zkomponoval. Každé emoci je tudíž přiřazeno 20 ukázek, z toho 8 bylo zkomponováno či vybráno zcela intuitivně na základě mého subjektivního úsudku. Zbylých 12 ukázek pro danou emoci je transponovaná v DAW programovaná demo ukázka, která byla vytvořena přesně podle parametrů odpovídajícím teoretickému předpokladu o emocích v hudbě, který jsem již představil v teoretické části. Specifickým případem v hudební databázi jsou ukázky pro neutrální emoci. Ta je totiž definována jako referenční statická kategorie. Ukázky pro tuto kategorii emocí jsou proto pouze různě transponované zahrané dlouhé tóny, což má charakterizovat to, že tu nedochází k žádným změnám, které by mohly popsat jednu ze zkoumaných emocí.

7.2.1 Intuitivně určená část databáze

Tato část databáze tvoří 40 procent z jejího celkového rozsahu. Přibližně půlka ukázek spadajících do této části databáze je vzata z již hotových děl. Jedná se o písně mé kapely, písně složeny sólově mnou, nahrávky nahrané ve výuce předmětu Vybavení studia či můj umělecký projekt. Zbytek je doplněn o písně, které jsem pro účely práce zkomponoval. U některých písní jsem si pomohl dostupnými hotovými loopy ve Studiu One.

7.2.2 Část databáze určená dle teoretických předpokladů

Zde jsem vyšel z konceptu koeficientů temnosti a světlosti spojeného s tempem skladby. Pro každou emoci jsem naprogramoval kratičkou klavírní skladbu dle požadovaných parametrů,čímž jsem chtěl definovat jakýsi etalon pro danou emoci. Každá

ze skladeb má svých dvanáct různě transponovaných verzí. To má za cíl ukázat, že změna výšky celého celku je z pohledu určování veličin zcela nepodstatnou změnou.

Vztek

Skladba pro tuto emoci je v rychlejším tempu (180 BPM) a v originální netransponované verzi je v tónině C moll.

Nuda

Ukázková krátká skladba k této emoci je v pomalejším tempu (80 BPM) a originální výškově neposunutá verza skladby je v tónině C dur.

Strach

Tato emoce byla z hlediska definice vůbec ta nejobtížnější. Tempo skladby jsem nastavil na 120 BPM. Pro melodie v dílu obsažené jsem zvolil frygickou stupnici, která by pocit strachu u posluchače měla navozovat.

Štěstí

V tomto případě jsem vytvořil durovou verzi tréninkové skladby pro vztek. Tudíž co se týče nastavených parametrů, tak je zde rychlejší tempo (180 BPM) a tónina C dur.

Neutralita

Zde je pouze po dobu šesti vteřin zahraná jedna, případě dvě, noty. V případě jedné zahrané noty využívám pouze intervalu oktávy a kvinty. Striktně se vyhýbám tercii, která by mohla utvořit jasně určitelný akord, kde by v závislosti na tom, zda je zahrána malá či velká tercie, šlo rozlišit mollový, respektive durový, akord.

Smutek

Zde je vytvořena mollová a zpomalená verze skladby definující štěstí. Řečí parametrů je zde nastavené tempo 80 BPM a tónina C moll.

7.3 Parametrizace řeči

Pro řešení této problematiky lze naleznout nemálo inspirace v mnoha odborných publikacích či závěrečných pracích. O možných konceptech klasifikace emocí v řeči bylo doopravdy mnoho napsáno. Při určování nejvhodnějších parametrů jsem se

nejprve snažil vyjít z čistě laických předpokladů a poté jsem se snažil najít způsob, jak tyto niance vhodně kvantifikovat a přenést do světa vědy. Poté jsem zkoušel vybrat vhodné příznaky z oblasti analýzy řeči, s kterými jsem se seznámil při teoretické přípravě k této práci. Pro výpočet parametrů jsem poté vytvořil živý skript v MATLABu. Výsledkem byl dataset obsahující všechny mnou vybrané parametry k jednotlivým nahrávkám řeči.

7.3.1 Výkonový poměr

Při pořizování nahrávek jsem si všimnul toho, že častou metodou pro navozování určité emoce je změna hlasitosti projevu. Např. při vzteku mnozí řečníci přecházeli až do řevu, zatímco třeba u nudy měli tehdenci svůj projev výrazně tlumit. Hodnoty neutrálních promluv zde sloužily jako reference. Mimo jiné důvodem pro tuto metodiku byl i fakt, že každý z řečníku často volil jinou vzdálenost od mikrofonu, a tudíž bylo nutné tento vliv na celkovou hlasitost nějakým elegantním způsobem eliminovat. Z každého vzorku jsem přes připravený skript vypočítal efektivní hodnotu RMS, tu jsem poté vydělil časem dané nahrávky. Nakonec jsem takto vypočítaný výkon vydělil hodnotou výkonu získanou z neutrálního projevu k odpovídajícímu slovu od odpovídajícího řečníka. V praxi to znamenalo, že hodnoty výkonových poměrů pro všechny neutrální promluvy jsou rovny 1.

```
%cteni souboru, prehrani, hledani fundamentu a rms
[x, fs] = audioread("Neutral_Ano_1.wav")
sound(x, fs)
pitch(x,fs)
xlabel("čas (s)")
ylabel("frekvence (Hz)")
voicef0 = pitch(x,fs);
meanvoice = mean(voicef0);
minvoice = min(voicef0);
maxvoice = max(voicef0);
energy=rms(x);
energyLR = mean(energy);
deviation = std(voicef0);
```

```
x = x(:,1);
```

```
%vytvoreni casove osy a vypocet rms na sekundu
N = length(x);
t = (0:N-1)/fs;
cas = N/fs;
energyPerSecond = energyLR/cas;
```

Obr. 7.2: Ukázka kódu v MATLAB

7.3.2 Rozdíl fundamentu

Tento parametr se snaží reflektovat změny intonance řečníka pro různé emoce. Opět tu neutrální promluva slouží jako reference. Především z důvodu, že je nutné vzít v potaz fakt, že každý řečník má jinak vysoký hlas a kdybychom se upnuli na absolutní hodnoty fundamentů, těžko by se z toho daly vyčíst relevantní informace. Už samotné určení fundamentu pro vybraný vzorek si vyžadovalo zvolení vhodného přístupu. Fundament se v čase mění. Určitě by se daly zvolit metody jako počítání celkového průměru apod., ale to by také mohlo výsledky výrazně zkreslit. Proto jsem se rozhodl vždy podrobit zkoumání průběhy fundamentu v čase. Předpokladem bylo, že každá slabika má svůj vlastní fundament. V grafu jsem vždy hledal neměnné úseky v podobě vodorovných čar a vždy jsem v dané oblasti vyčetl hodnotu kmitočtu. Většina grafů byla v tomto směru dobře čitelná. Počet neměnných oblastí odpovídal počtu slabik. V případě dvojslabičných slov byl výsledný fundament určen jako průměr fundamentů každé slabiky. Nakonec byl vypočítán rozdíl mezi fundamentem dané emoce a emoce neutrální pro odpovídající řečníky a pro odpovídající slova.

7.3.3 Průměr MFCC

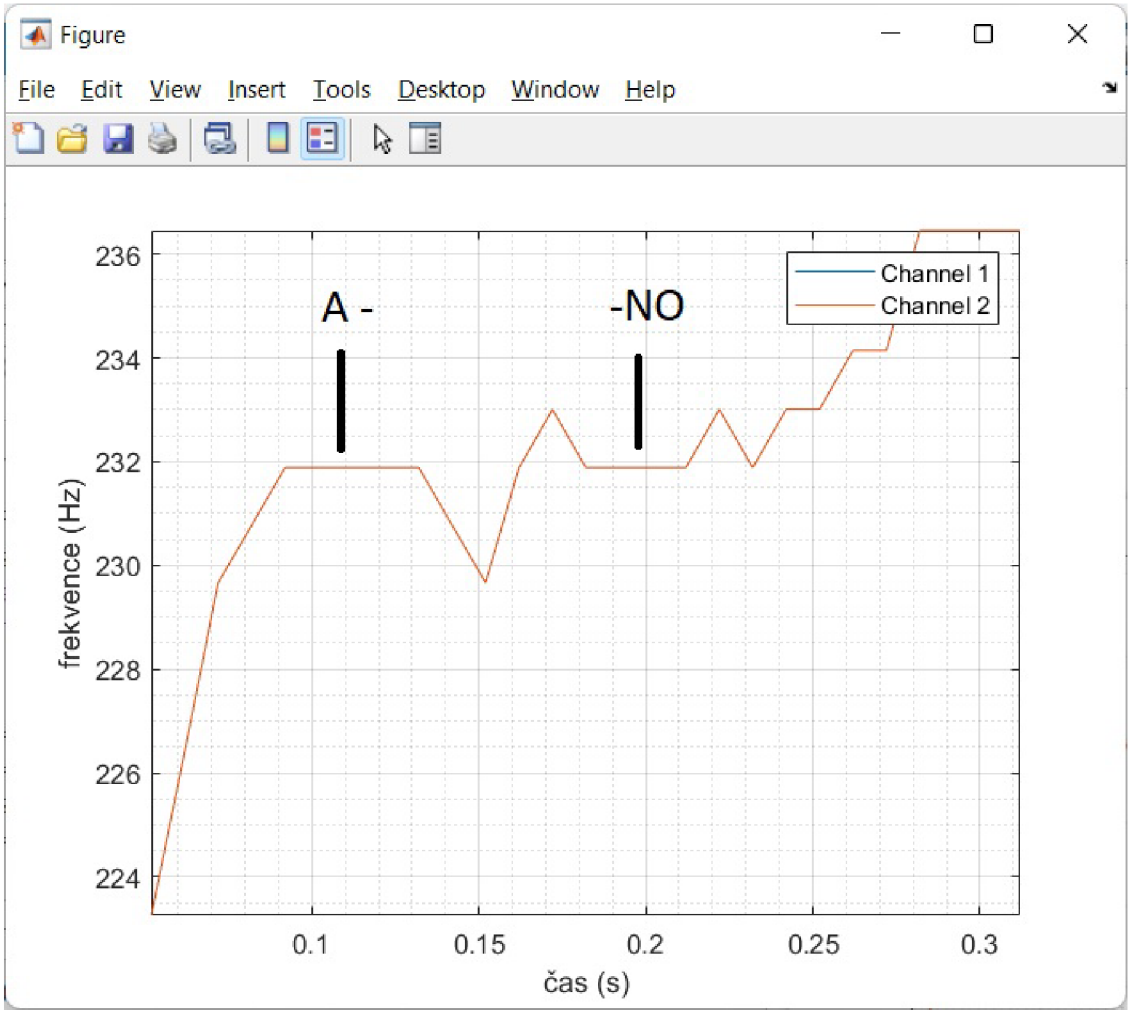
Z každého vzorku byla v MATLABU vyextrahována matice, která v sobě měla obsaženo 13 koeficientů měnících se v čase. Nejprve jsem výpočtem průměru z matice udělal vektor a následně stejnou operací skalár. Tento parametr či jeho různé variace mají široké využití v rozpoznávání řeči. Nicméně jsem si všimnul korelací určitých hodnot pro dané emoce a napadlo mě tedy tento parametr také použít.

7.4 Parametrizace hudby

Určení parametrů pro hudební ukázky si žádalo mnohem větší dávku kreativity a přemýšlení než v předchozím případě při určování parametrů pro řeč. Vycházel jsem především z parametrů užívaných v hudební teorii. Pro určité nalezení inspirace jsem si do MATLABu nainstaloval MIRtoolbox, který vyvinul Dr.Olivier Lartillot z Ženevské Univerzity. Tento toolbox slouží k získávání mnohých parametrů z hudebních nahrávek. Bohužel se mi ve většině parametrech, které tento toolbox dokázal vypočítat, nepodařilo najít korelace pro jednotlivé emoce. Nicméně věřím tomu, že při hlubším bádání by se za použití tohoto toolboxu dalo prozkoumat mnohé.

7.4.1 Tempo

U většiny nahrávek použitých v databázi jsem již tempo znal a v případě nahrávek, u kterých jsem tempo nevěděl, jsem si na pomoc vzal MIRtoolbox a jeho funkci



Obr. 7.3: Určování fundamentu

mirtempo. V případě nahrávek, u kterých tempo nebylo konstatní, byl vypočítán průměr.

7.4.2 Koeficient modu

MIRtoolbox dokáže určit sice tóninu, není však schopen detailnějšího určení použitých modů. Od úmyslu vyřešit tento problém vlastním kódem v MATLABu jsem rychle upustil, protože realizace algoritmu na určení stupnic a modů z hrané melodie by vydala za samostatnou diplomovou práci. Musel jsem tedy veškeré ukázky manuálně zanalyzovat. U sporných nahrávek jsem si vzal na pomoc jiné osoby znalé v hudební teorii.

7.4.3 Průměr MFCC

I zde jsem použil tento parametr. Jeho výpočet umožňoval i mnou používaný toolbox. Využití MFCC je úspěšně aplikováno v klasifikaci hudebních žánrů. Ukázky v databázi jsou žánrově velmi rozdílné, nicméně by se dalo říci, že u některých kategorií se v ukázkách objevují často podobné barvy. Například nahrávky pro vztek jsou zhruba z poloviny tvořeny skladbami, co využívají elektrických kytar. Byť použití tohoto parametru byl na začátku pouze takový slepý pokus, tak se mi za jeho užití podařilo dosáhnout dobrých výsledků.

7.5 Trénink a testování neuronových sítí

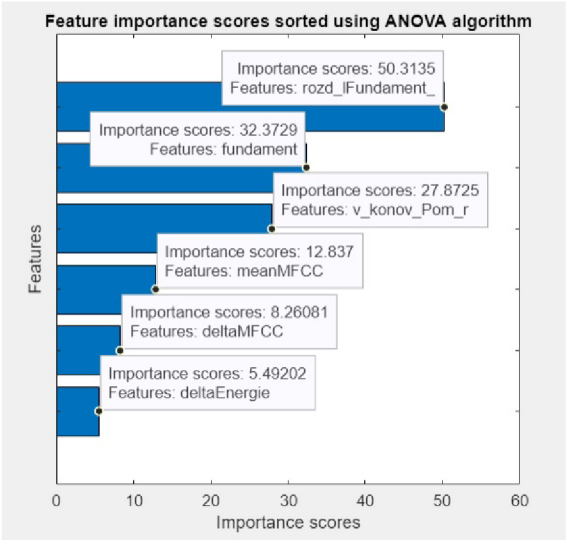
Pro účely trénování a testování neuronových sítí jsem využil aplikaci v MATLABu Classification Learner. Tato aplikace nabízí několik metod strojového učení včetně využití různých typů neuronových sítí. Z možných nabízených metod validace se mi nejlépe osvědčila 5-násobná křížová validace, kterou jsem nakonec použil pro trénink a testování všech modelů. Datasety jsem vždy rozdělil v poměru 3:2 na tréninkový a testovací.

7.5.1 Trénink a testování databáze řeči

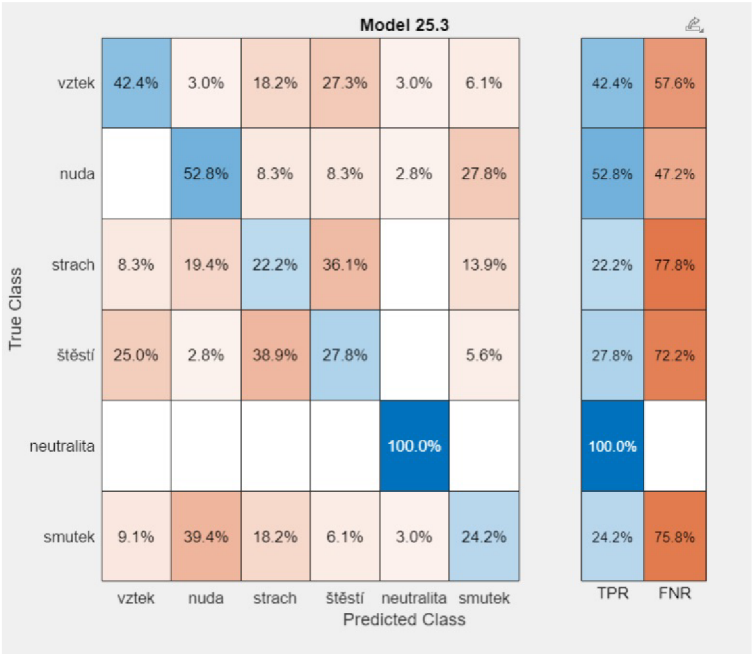
Společně s dalšími čtyřmi posluchači jsem vytvořil pořadí věrohodnosti řečníků. Část datasetu určená pro trénink obsahovala šest nejvěrohodnějších řečníků. Zbylí čtyři řečníci tvořili část datasetu určenou pro testování. Testoval jsem s různým počtem parametrů. Nakonec se nejvíce osvědčil set následujících parametrů. Zmiňované tři parametry byly doplněny o hodnoty fundamentů, rozdíl průměru MFCC od neutrální promluvy a rozdíl RMS od neutrální promluvy. Algoritmus pro vyhodnocování důležitosti vlastností ANOVA vybrané příznaky seřadil následovně: Trénink a testy byly provedeny na několika modelech. Detailně tu rozeberu tři typy neuronových sítí, u kterých se mi podařilo dosáhnout nejlepších výsledků. U všech zmíněných modelů jsou hodnoty úspěšnosti validace a testování téměř totožné

Jednovrstvá síť o sto neuronech

Třetího nejlepšího výsledku jsem dosáhl s jednovrstvou neuronovou sítí o 100 neuronech využívající aktivační funkci ReLU. Byť se jedná o třetí nejlepší výsledek, tak přesnost nedosahuje ani padesáti procent. Dle předpokladů se neuronová síť nejlépe vypořádala s neutrální emocí.



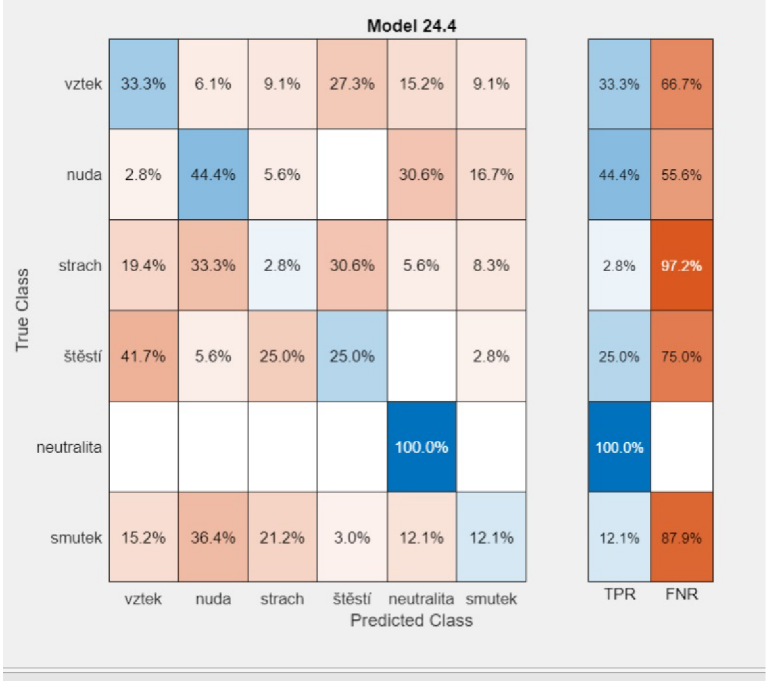
Obr. 7.4: Priorita parametrů



Obr. 7.5: model wide neural network

Kosinová Kohonenova neuronová síť

Tento konkrétní typ Kohonenovy neuronové sítě využívá kosinovou distanční metriku. Má 10 sousedů. Největší problém tomuto modelu činily emoce strach a smutek. Opět nebylo dosáhnuto ani poloviční úspěšnosti.



Obr. 7.6: Kosinova neuronová síť

Jemná Kohonenova neuronová síť

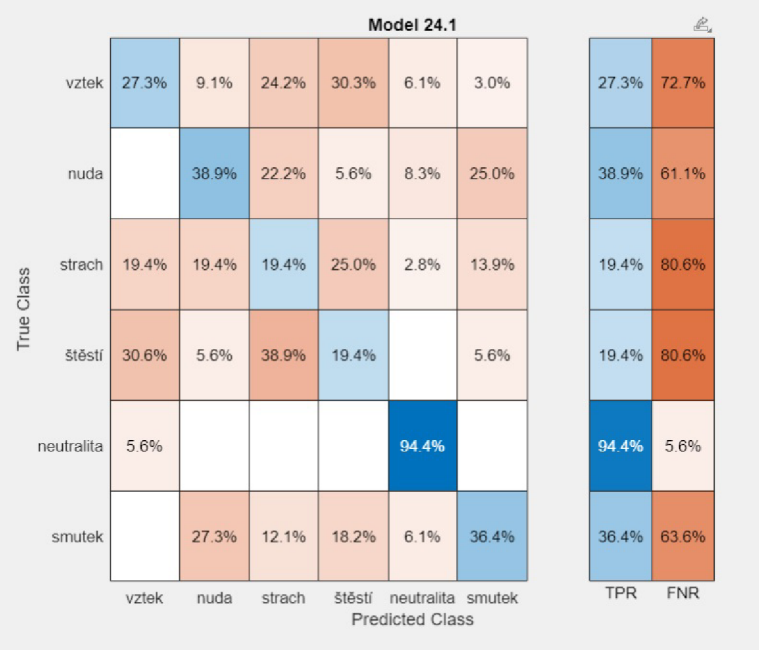
Narozdíl od předchozího modelu tato síť využívá Euklidovu vzdálenostní metriku a pouze jednoho souseda. Přes nejlepší výsledek u tohoto modelu úspěšnost klasifikace neutrální emoce nedosáhla sta procent.

Metody strojového učení nevyužívající neuronové sítě

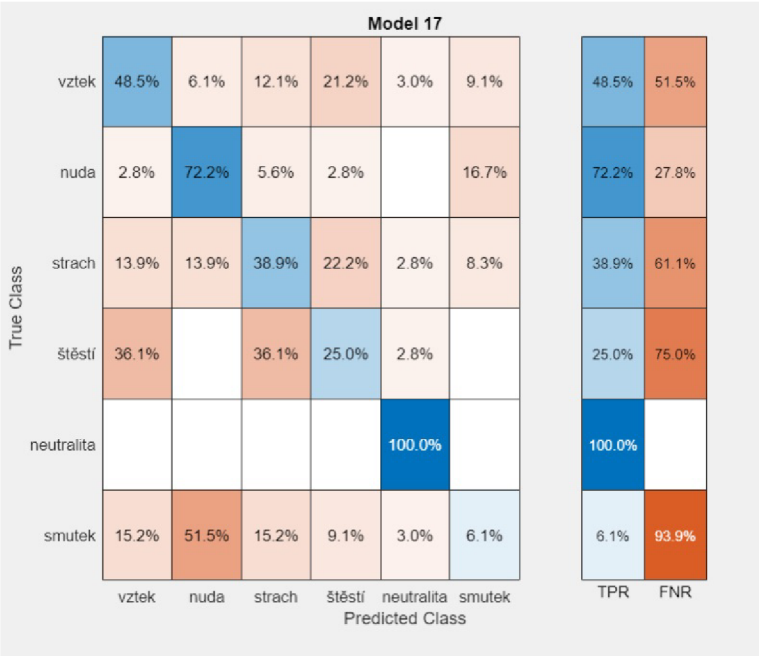
Pro srovnání jsem vyzkoušel i jiné modely než umělé neuronové sítě. Nejlepšího výsledku jsem dosáhl u rozhodovacího stromu. Podařilo se mi dosáhnout hodnot vyšších než padesát procent.

7.5.2 Trénink a testování hudební databáze

Dataset jsem rozdělil následujícím způsobem. Vždy jsem od každé emoce vzal 8 ukázek vytvořených dle teoretických předpokladů a 4 ukázky z intuitivní části databáze. Zbytek nahrávek byl potom obsahem tréninkové části datasetu. Využitými parametry byly koeficienty modu, tempo a průměr MFCC. Kruskalův-Wallisův test vyhodnotil důležitost parametrů následovně:



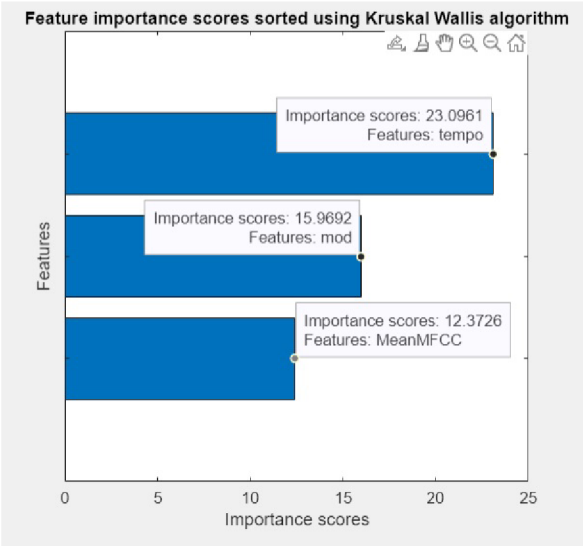
Obr. 7.7: Jemná Kohonenova neuronová síť



Obr. 7.8: Rozhodovací strom

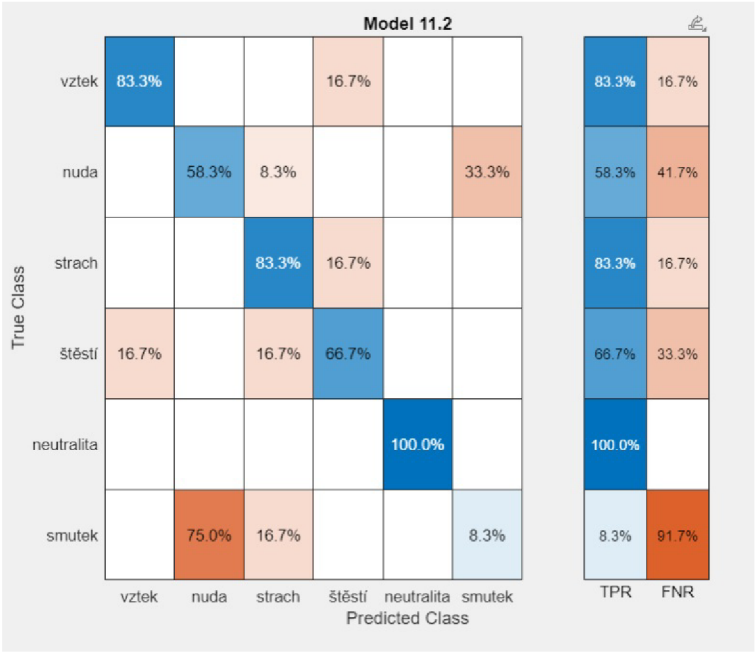
Kohonenovy sítě

Kohonenovy sítě si nevedly příliš dobře. U validace bylo ještě dosáhnuto solidních výsledků, avšak testování dopadlo tristně a nepodařilo se dosáhnout ani poloviční

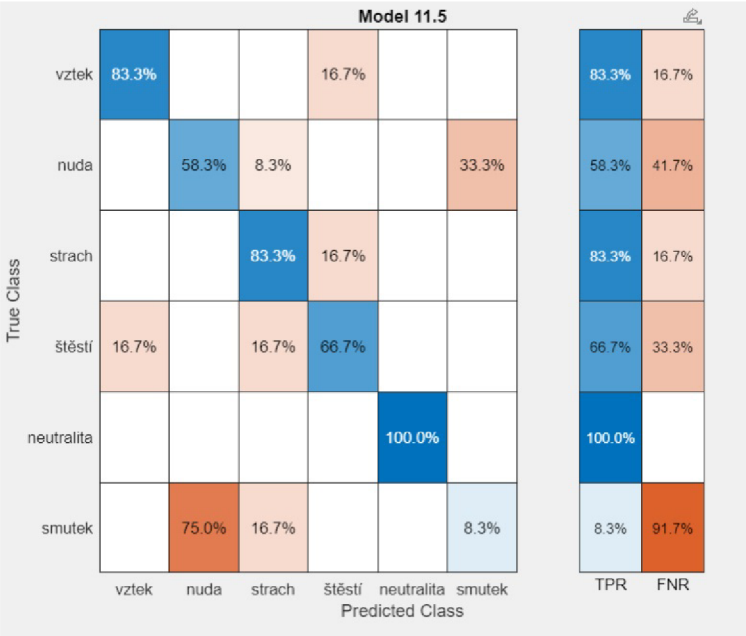


Obr. 7.9: Kruskalův-Wallisův test

úspěšnosti. Střední a kubická Kohonenova síť dosáhla shodné úspěšnosti validace, která činila 66.7 procent, a úspěšnosti klasifikace činící 47.9 procent.



Obr. 7.10: Střední KNS



Obr. 7.11: Kubická KNS

Třívrstvá neuronová síť

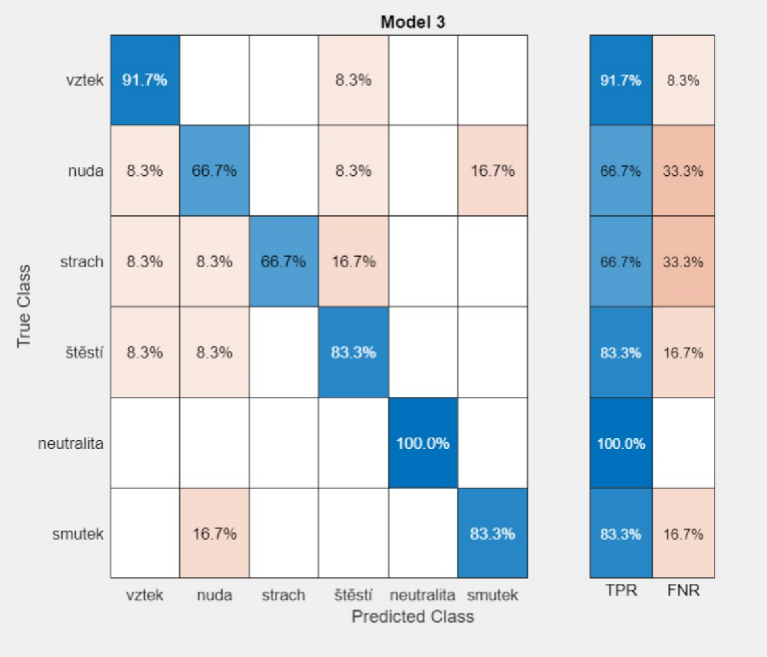
Otestoval jsem mnoho variant počtu neuronů v jednotlivých vrstvách. Nejlepších výsledků jsem dosáhnul s 9 neurony ve vstupní vrstvě, 20 neurony ve skryté vrstvě a 4 neurony ve výstupní vrstvě. Využitou aktivační funkcí byla ReLU. Validační úspěšnost byla 81.9 procent a úspěšnost klasifikace činila 79.2 procent.

7.6 Poslechový test

Poslechový test jsem provedl se čtyřmi posluchači, kterým bylo přehráno 30 náhodných nahrávek řeči a 30 náhodných hudebních nahrávek. Každá nahrávka měla číslo a posluchač měl k danému číslu přiřadit jednu z šesti nabízených emocí.

7.6.1 Vyhodnocení poslechového testu u řečových nahrávek

Nejlépe klasifikovatelnou emocí v nahrávkách u poslechového testu byl vztek, u kterého bylo dosaženo osmdesátiprocentní úspěšnosti. Zatímco nejhůře se posluchačům klasifikoval strach. Celková úspěšnost činila 60 procent, což je o trochu lepší výsledek než u klasifikace realizované neuronovými sítěmi. Tento test odhalil mnohé nedokonalosti databáze vyplývající z faktu, že ukázky byly nahrávány neherci.U některých řečníků bylo velmi obtížné identifikovat jakoukoliv emoci. Jiní řečníci při nahrávání



Obr. 7.12: Třívrstvá neuronová síť

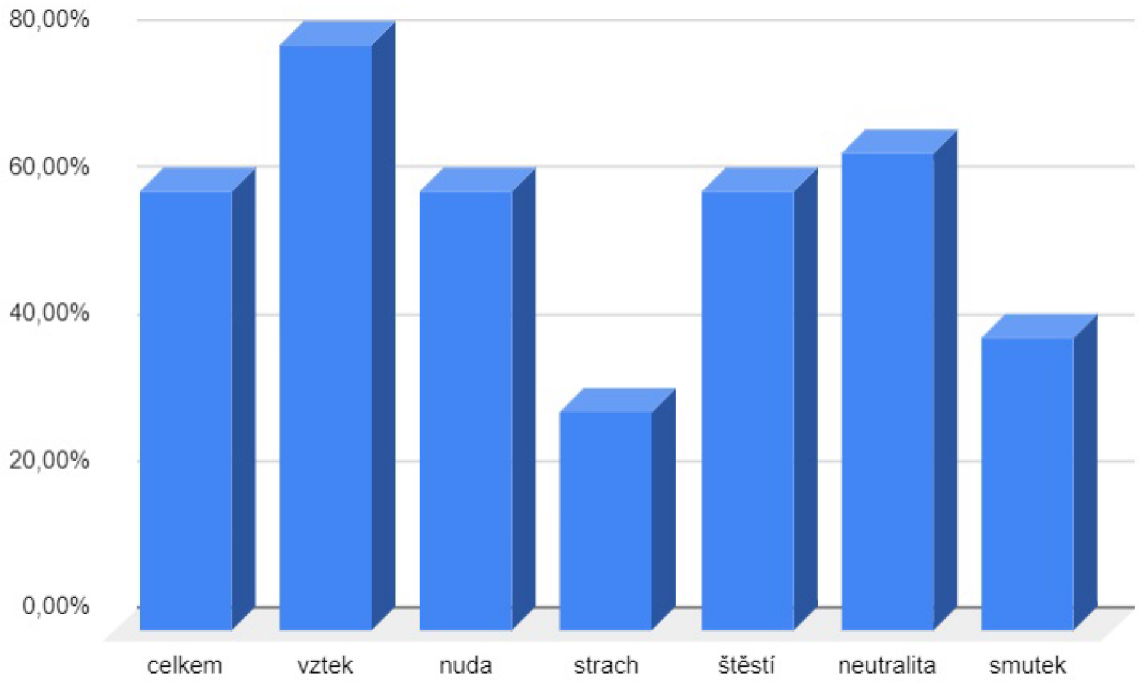
naopak zacházeli do přehnaného afektu, což však klasifikaci posluchačům mnohdy ulehčilo.

7.6.2 Vyhodnocení poslechového testu u hudebních nahrávek

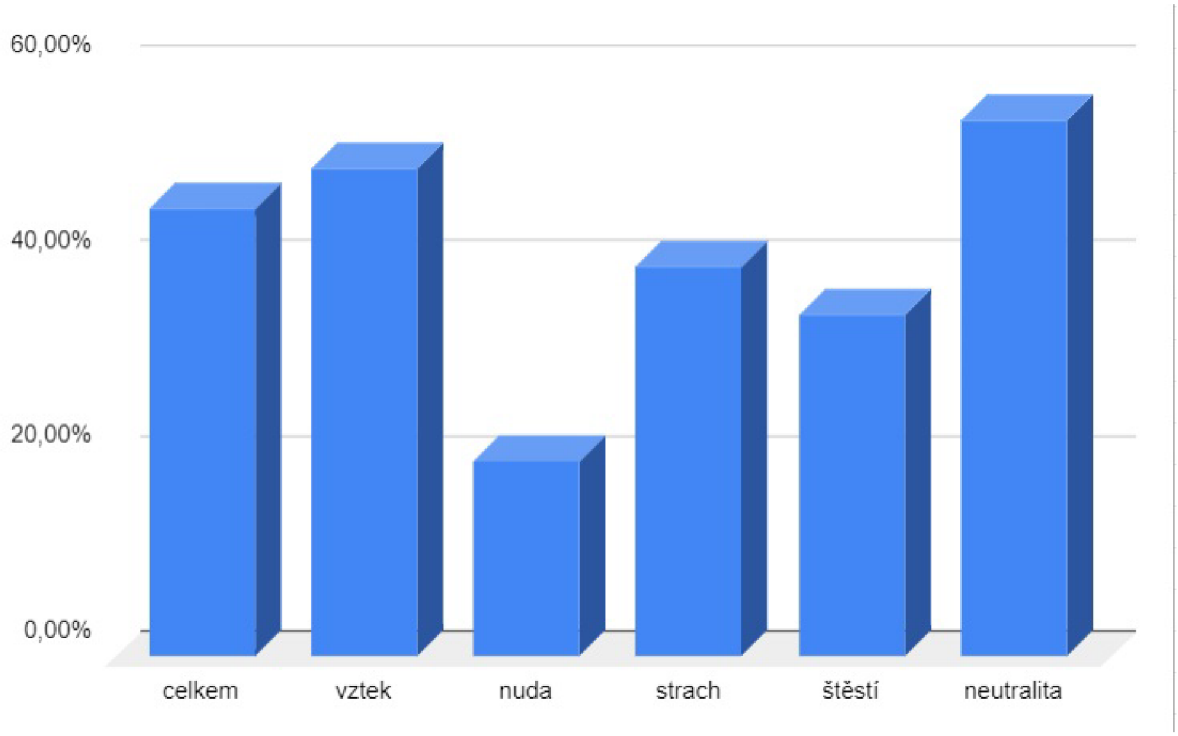
Zde si posluchači vedli o něco hůře než u řeči. Nejvýraznějším problémem bylo klasifikovat nudu. U této emoce byla úspěšnost pouze 20 procent. Celková úspěšnost byla 45.38 procent, což je výrazně menší úspěšnost než u neuronových sítí.

7.6.3 Srovnání výsledků

Zajímavým výstupem statistik je fakt, že při klasifikaci emocí v hudbě posluchači výrazně zaostali oproti umělé neuronové síti, zatímco v případě nahrávek řeči byli jen lehce lepší než neuronová síť. Obzvláště u vyhodnocení poslechových testů se projevila výrazná subjektivita hodnocení jednotlivých posluchačů. Nebyly výjimkou případy, kdy jedné ukázce přiřadil každý z posluchačů úplně jinou emoci. U klasifikace emocí v řeči často dělali posluchači chybu u totožných vzorků.



Obr. 7.13: Výsledky testů u nahrávek řeči



Obr. 7.14: Výsledky testů u nahrávek hudby

8 Závěr

Ve své práci jsem se snažil vytvořit databázi řečových nahrávek a hudebních nahrávek, z nichž jsem následně určil vhodné parametry pro zařazení do datasetu určeného pro trénink ,respektive klasifikaci, emocí. Zde nastal první zásadní problém. Tím je nedostatečná velikost obou datasetů. Pro umělé neuronové sítě není dataset obsahující řádově pár stovek vzorků dostatečně velký. Takže už z tohoto důvodu nešlo očekávat příliš oslnivé výsledky. Určení parametrů pro řečovou databázi vycházelo z celkem relevantních předpokladů, ale kromě malé velikosti datasetu se zde projevil i další zásadní problém, který jenom potvrdily poslechové testy. Věrohodnost mnohých nahrávek není příliš valná. Samozřejmě je nutné brát v potaz fakt, že lidé, kteří se zúčastnili nahrávání, nejsou profesionální herci. Nikomu ze zúčastněných řečníků rozhodně nelze upřít snahu. V každém případě to vysvětluje i souvislost mezi výsledky klasifikace neuronovýma sítěma a výsledky klasifikace posluchači. Výsledky zde byly velmi podobné, což jasně potvrzuje, že při lépe zvládnutém nahrávání by se i s menším datasetem dalo dosáhnout lepší přesnosti. U hudební databáze by možná ještě stálo za to trochu popřemýšlet o možných parametrech. Při klasifikaci bylo u hudební databáze dosáhnuto mnohem lepších výsledků, avšak tyto výsledky jsou v silném kontrastu oproti výsledkům poslechového testu. Úvahy nad možnostmi parametrizace hudby mě dohnaly k mnohým nápadům, které by sami o sobě mohly sloužit jako zadání jiné závěrečné práce. Především vytvoření systému pro rozpoznávání užívaných tónických prostředků v kompozici, detailní rozpoznávání toho, jaký modus je v melodii použit. V dnešní době není žádným problémem automatické rozlišení durové a mollové tóniny. Dostat se v této oblasti ještě dál a zkoumat mody by bylo velice zajímavou výzvou. Jeden z parametrů je mnou definovaný koeficient modu, což rozhodně napomáhá lepším výsledkům. Dalším zásadním faktem je subjektivita vnímání emocí v hudbě. Veškerá parametrizace vycházela z mého úsudku, který však u mnohých posluchačů může být výrazně odlišný.

Literatura

[1] *STUCHLÍKOVÁ, I.vadání Základy psychologie emocí. Praha: Portál 2002. ISBN: 8071785539 s. 11*

[2] *ČERNÝ, Lukáš. Rozpoznávání a klasifikace emocí na základě analýzy řeči Brno, 2010 [cit. 2021-12-11]* Dostupné z: <http://hdl.handle.net/11012/10673> [online] Diplomová práce. Vysoké učení technické v Brně. Fakulta elektrotechniky a komunikačních technologií. Ústav telekomunikací. Vedoucí práce Zdeněk Smékal.

[3] *NAKONEČNÝ, Milan. Lidské emoce. Praha: Academia, 2000. ISBN 80-200-0763-6. S. 42.*

[4] *https://medicalxpress.com/news/2018-04-scientists-disconfirm-belief-human html*

[5] *https://en.wikipedia.org/wiki/Robert_Plutchik*

[6] *PSUTKA J. Komunikace s počítačem mluvenou řečí ACADEMIA 1995*

[7] *PSUTKA J. Mluvíme s počítačem česky ACADEMIA 2006*

[8] *SIGMUND M. Analýza řečových signálů VUT 2000*

[9] *KRČMOVÁ, M. Fonetika.* [online]. Dostupné z WWW: <http://is.muni.cz/do/1499/el/estud/ff/js07/fonetika/materialy/index.html/>

[10] *https://cs.wikipedia.org/wiki/Hudba*

[11] *MLEJNEK R. Emoce v hudbě diplomová práce Univerzita Karlova 2007*

[12] *LIŠKOVÁ A. Emoce a city jako součást prožívání hudby bakalářská práce Západočeské univerzita 2014*

[13] *NAVRÁTIL,M. Rozpoznávání emočních stavů pomocí analýzy řečového signálu. Diplomová práce, 2008..*

[14] *FANČAL, Petr. Analýza zvukové interpretace hudby metodami číslicového zpracování signálu [online]. Brno, 2017.* Dostupné z: <http://hdl.handle.net/11012/65657> Diplomová práce. Vysoké učení technické v Brně. Fakulta elektrotechniky a komunikačních technologií. Ústav telekomunikací. Vedoucí práce Zdeněk Smékal.

[15] *TUČKOVÁ J. Úvod do teorie aplikací neuronových sítí ČVUT 2005*

[16] *https://cs.wikipedia.org/wiki/Neuron*

[17] *TUČKOVÁ J. Vybrané aplikace umělých neuronových sítí při zpracování signálu*

[18] *KAŇA, Michal Kohonenova síť: bakalářská práce. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav automatizace a měřící techniky, 2011. Vedoucí práce byl doc. Ing. Václav Jirsík, CSc*