UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE
FAKULTA PRÍRODNÝCH VIED

REAL TIME KLASIFIKÁCIA EMOCIONÁLNEHO STAVU Z REČI

Diplomová práca

2020 **Bc. Timotej Sulka**

UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE

FAKULTA PRÍRODNÝCH VIED

REAL TIME KLASIFIKÁCIA EMOCIONÁLNEHO STAVU Z REČI

DIPLOMOVÁ PRÁCA

Študijný odbor:	9.2.9 Aplikovaná informatika
Študijný program:	Aplikovaná informatika
Školiace pracovisko:	Katedra informatiky
Školiteľ:	PaedDr. Martin Magdin, Ph.D.

Nitra 2020 Bc. Timotej Sulka

Univerzita Konštantína Filozofa v Nitre
Fakulta prírodných vied

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta:	Bc. Timotej Sulka
Študijný program:	aplikovaná informatika (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor:	aplikovaná informatika
Typ záverečnej práce:	Diplomová práca
Jazyk záverečnej práce:	slovenský
Sekundárny jazyk:	anglický

Názov: Real Time klasifikácia emocionálneho stavu z reči

Anotácia: Rozpoznávanie hlasu sa v súčasnosti používa vo viacerých sférach, najmä pri call agentoch v rámci telekomunikačných centier. Človek sám dokáže nikedy vycítiť v akom rozpoložení sa druhá osoba nachádza. V prípade napr. detskej linky záchrany je tento fakt veľmi dôležitý. Preto sa v súčasnom období kladie dôraz na systémy, ktoré by sami v reálnom čase (on-line) dokázali rozpoznať a klasifikovať emocionálny stav človeka z reči.

Cieľ diplomovej práce: Cieľom bude vývoj riešenia automatického rozpoznávania a klasifikovania emocionálneho stavu používateľa za základe valencie (intenzita a rozsah pôsobenia, ako i šírka optimálneho vplyvu faktoru na organizmus) reči. Výstup je očakávaný vo forme plne funkčného prototypu riešenia, ktoré bude možné demonštrovať koncovým používateľom.

Charakter práce: aplikačný.

Požiadavky na obsah podľa charakteru práce: popis riešeného problému, návrh systému/hardvérového riešenia a pod. (modely, ..), metodika vývoja/tvorby, implementácia, popis vytvoreného riešenia, testovanie.

Požiadavky na vedomosti a zručnosti študenta:
- znalosť anglického jazyka za účelom štúdia súvisiacej technickej dokumentácie,
- znalosť programovacích jazykov (preferovaný je jazyk C, Java, Python),
- základná znalosť java scriptingu (client side scripting),
- znalosť problematiky webových služieb (pre potreby integrácie formou API).

Školiteľ:	PaedDr. Martin Magdin, Ph.D.
Oponent:	Mgr. Martin Vozár, PhD.
Katedra:	KI - Katedra informatiky
Dátum zadania:	28.09.2018
Dátum schválenia:	24.10.2018

RNDr. Ján Skalka, PhD., v. r.
vedúci/a katedry

POĎAKOVANIE

Ďakujem mojej rodine a priateľom za dôležitú podporu, môjmu školiteľovi PaedDr. Martinovi Magdinovi, Ph. D. za pripomienky a poskytnuté rady, a A.E. za jej prínos vede.

ABSTRAKT

SULKA, Timotej: Real Time klasifikácia emocionálneho stavu z reči. [Diplomová práca]. Univerzita Konštantína Filozofa v Nitre. Fakulta prírodných vied. Školiteľ: PaedDr. Martin Magdin, Ph.D. Stupeň odbornej kvalifikácie: Magister odboru Aplikovaná informatika. Nitra: FPV, 2020. 67 s.

Táto práca sa zaoberá problematikou klasifikácie emocionálneho stavu z reči, a bližšie sa zameriava na vytvorenie riešenia, ktoré umožní klasifikáciu emocionálneho stavu z reči v reálnom čase. V teoretickej časti sumarizuje charakteristiky reči, nástroje na extrakciu týchto charakteristík, rôzne architektúry neurónových sietí, ich využitie v klasifikácii emócií a existujúce riešenia klasifikácie emócií v reálnom čase. Praktická časť opisuje postup predspracovania údajov, experimenty s neurónovými sieťami a vlastnú aplikáciu pre klasifikáciu emócií v reálnom čase naprogramovanú v jazyku Python. Záver práce identifikuje časti riešenia, ktoré môžu byť zlepšené, a poskytuje návrhy pre ďalšie rozširovanie.

Kľúčové slová: Reč. Klasifikácia emócií. Real time. Neurónové siete.

ABSTRACT

SULKA, Timotej: Real Time Classification of Emotional States from Speech. [Diploma thesis]. Constantine the Philosopher University in Nitra. Faculty of Natural Sciences. Supervisor: PaedDr. Martin Magdin, Ph.D. Degree of Qualification: Master of Applied Informatics. Nitra: FNS, 2020. 67 p.

Aim of this diploma thesis is speech emotion recogniton, more specifically its goal is introduction of a solution, that will be able to recognize emotions from speech in real time. In theoretical part we provide summary of speech features, tools for feature extraction, various neural network architectures, their use in emotion recognition and existing solutions of real time speech emotion recognition. Practical part describes data preprocessing, experiments with neural networks and our own application for real time emotion recognition written in Python language. Final part of this thesis identifies steps of solution, which can be improved and provides suggestions for further development.

Keywords: Speech. Emotion classification. Real time. Neural networks.

OBSAH

Úvod **7**

1 Analýza súčasného stavu **8**

1.1 Charakteristiky reči 8

1.1.1 Prozodické charakteristiky 8

1.1.2 Kvalitatívne charakteristiky 9

1.1.3 Spektrálne charakteristiky 10

1.1.4 Vizualizácia zvuku pomocou spektrogramov 14

1.2 Nástroje na spracovanie zvukových signálov 15

1.3 Využitie neurónových sietí na klasifikáciu emócií 16

1.3.1 Dopredné neurónové siete 17

1.3.2 Rekurentné neurónové siete 18

1.3.3 Konvolučné neurónové siete 19

1.4 Klasifikácia emócií v reálnom čase 21

2 Ciele diplomovej práce **24**

Čiastkové ciele 24

3 Postup vlastného riešenia a návrh aplikácie **25**

3.1 Nadviazanie na predchádzajúcu prácu 25

3.2 Analýza vstupných údajov 26

3.3 Predspracovanie údajov 28

3.3.1 Vyrovnanie uniformity tried 29

3.3.2 Orezanie ticha a normalizácia 30

3.3.3 Augmentácia údajov 31

3.4 Experimenty s neurónovými sieťami 33

3.4.1 Experiment 1 – dopredná neurónová sieť 34

3.4.2 Experiment 2 – 1D konvolučná sieť 38

3.4.3 Experiment 3 – 2D konvolučná sieť 42

3.4.4 Zhodnotenie experimentov a výber modelu 45

3.5 Program EmoRec 2 46

3.5.1 Obsluha programu EmoRec 2 47

3.5.2 Štruktúra programu EmoRec 2 50

3.5.3 Princíp fungovania programu EmoRec 2 52

4 Možnosti vylepšovania a rozširovania **54**

4.1 Vstupné údaje 54

4.2 Spracovanie údajov 55

4.3 Experimenty 56
4.4 Aplikácia 56
Záver 59
Zoznam bibliografických odkazov. 60
Zoznam príloh 67

ÚVOD

Klasifikácia emocionálneho stavu používateľa na základe hlasu je náročná úloha, ktorá priťahuje pozornosť odborníkov, a je predmetom mnohých akademických prác. Súčasný výskum sa sústreďuje predovšetkým na experimenty s rôznymi klasifikátormi a charakteristikami, a hľadanie ich najlepšej kombinácie. Problematické je pomerne malé množstvo dostupných nahrávok emócií, ktoré je potenciálne možné využiť pri vytváraní klasifikátora, a takisto fakt, že ľudia majú v reálnych situáciách tendenciu emócie potláčať a neprejavovať ich naplno. Ďalšou prekážkou vo vytvorení univerzálneho riešenia je ľudský hlas, ktorý môže byť ovplyvnený mnohými faktormi – napr. pohlavie, vek, zdravotný stav, atď.

Hlavnou motiváciou pokusov o strojové rozpoznávanie emócií sú široké možnosti využitia v prípade úspechu. Medzi oblasti kde by rozpoznávanie emócií mohlo nájsť uplatnenie patrí napríklad medicína, núdzové linky, inteligentné výukové systémy, komunikácia so zákazníkom v call centrách, rozšírenie možností interakcie s počítačom a rôzne formy umelej inteligencie.

Skutočnou výzvou je vytvorenie systému, ktorý dokáže klasifikovať emócie z hlasu používateľa v reálnom čase, a bude možné ho použiť aj v praxi. Takýto systém musí byť nielen presný, ale aj veľmi rýchly – musí reagovať na zmenu hlasového prejavu používateľa, a poskytovať okamžitú spätnú väzbu. Dá sa predpokladať, že podmienky pri praktickom použití nebudú vždy ideálne, a systém sa bude musieť vysporiadať aj so vstupom so zníženou kvalitou (šum, rušné prostredie, nekvalitný mikrofón). Toto všetko náročnosť klasifikácie ešte zvyšuje. V tejto práci sa pokúsime navrhnúť riešenie, ktoré sa k použitiu v praxi čo najviac priblíži, prípadne sa bude môcť stať súčasťou systému, ktorý kombinuje klasifikáciu pomocou rôznych biometrík.

1 ANALÝZA SÚČASNÉHO STAVU

Táto kapitola tvorí teoretický základ celej práce. Zaoberá sa charakteristikami reči a ich extrakciou. Ďalej sa zaoberá rôznymi typmi architektúr neurónových sietí a poskytuje sumarizáciu prác, v ktorých boli neurónové siete využité na klasifikáciu emocionálneho stavu z reči. V závere sa zameriava na pokusy o klasifikáciu v reálnom čase a existujúce riešenia fungujúce v praxi. Definované pojmy a získané znalosti z tejto časti nám neskôr pomôžu pri vytvorení vlastného riešenia

1.1 CHARAKTERISTIKY REČI

Dôležitým krokom pri návrhu systému, ktorý rozpoznáva emócie, je extrakcia vhodných vlastností z hlasu, ktoré efektívne charakterizujú rozličné emócie. Niektorí odborníci preferujú delenie signálu na krátke intervaly, z ktorých je extrahovaný vektor lokálnych charakteristík, iní radšej používajú globálne charakteristiky. Pretože signál reči nie je stacionárny, je bežné ho pri spracovaní rozdeliť na malé časti, ktoré sa nazývajú rámce. V jednom rámci je už signál považovaný za približne stacionárny, a z každého rámca sú extrahované charakteristiky ako napr. energia alebo frekvencia. Takto získané charakteristiky sa označujú ako lokálne (El Ayadi, Kamel, Karray, 2011).

Globálne charakteristiky sú štatistiky vypočítané zo všetkých extrahovaných charakteristík. Z lokálnych charakteristík sa vypočíta napr. maximum, minimum, rozptyl, priemer, smerodajná odchýlka, špicatosť, šikmosť, a ďalšie podobné hodnoty. Tieto hodnoty sú potom skombinované do jedného vektora globálnych charakteristík (Gao et al., 2017). Väčšina odborníkov sa zhoduje, že používanie globálnych charakteristík je výhodnejšie, pretože klasifikácia je s nimi rýchlejšia a presnejšia. Ich výhodou oproti lokálnym charakteristikám je ich nízky počet. Avšak, odborníci taktiež tvrdia, že globálne charakteristiky sú efektívne len pri rozlišovaní medzi energickými a málo energickými emóciami (napr. hnev a smútok), ale zlyhávajú v rozlišovaní emócií, ktoré sa prejavujú podobne energicky (napr. hnev a radosť) (El Ayadi, Kamel, Karray, 2011).

1.1.1 Prozodické charakteristiky

Prozódia je charakterizovaná rytmom, intonáciou, dôrazom a pauzami v reči (Zewoudie, Luque, Hernando, 2016). Prozodické charakteristiky sú viazané na dlhšie zvukové jednotky – slabiky, slová, frázy a vety (Rao, Koolagudi, Vempada, 2012). Prozodické charakteristiky sú bežne používané v systémoch, ktoré rozpoznávajú emócie.

Predpokladá sa, že tieto charakteristiky v sebe nesú užitočné informácie pre rozpoznávanie emócií (Sato, Obuchi, 2007). Staršie výskumy boli založené na skúmaní vplyvu základnej frekvencie. Okrem základnej frekvencie sú predmetom skúmania aj iné prozodické vlastnosti – napr. tempo reči, relatívne trvanie a intenzita (Gangamohan, Kadiri, Yegnanarayana, 2016).

Základná frekvencia (F_0) je frekvencia, ktorou kmitajú hlasivky. Prevrátenou hodnotou frekvencie je perióda – čas, ktorý uplynie medzi pulzmi hlasiviek (od otvorenia hlasiviek po ďalšie otvorenie) (Wilder, 1975). Základná frekvencia je ovplyvnená vekom a pohlavím. Pri mužoch sa spravidla pohybuje od 100 do 150 Hz, a u žien od 170 do 220 Hz (Zewoudie, Luque, Hernando, 2016).

Miera prechodu nulou (zerocrossing rate) je silne závislá od základnej frekvencie. Je jednoduchým meraním frekvenčného obsahu signálu. Prechod nulou nastáva, ak je znamienko dvoch za sebou idúcich vzoriek opačné. Miera prechodu nulou potom vyjadruje koľkokrát za určený interval (resp. rámec) signál prejde cez hodnotu 0 (Bachu et al., 2010).

Intenzita môže byť v reči použitá na vyjadrenie dôrazu a emócií (Zewoudie, Luque, Hernando, 2016). Môže byť meraná hladinou akustického tlaku v decibeloch. Hladina akustického tlaku (SPL) je 20 násobok dekadického logaritmu pomeru nameranej efektívnej hodnoty akustického tlaku a referenčnej hodnoty (Beranek, Mellow, 2019).

$$SPL = 20 \log_{10} \frac{p_{rms}}{p_{ref}}$$

1.1.2 Kvalitatívne charakteristiky

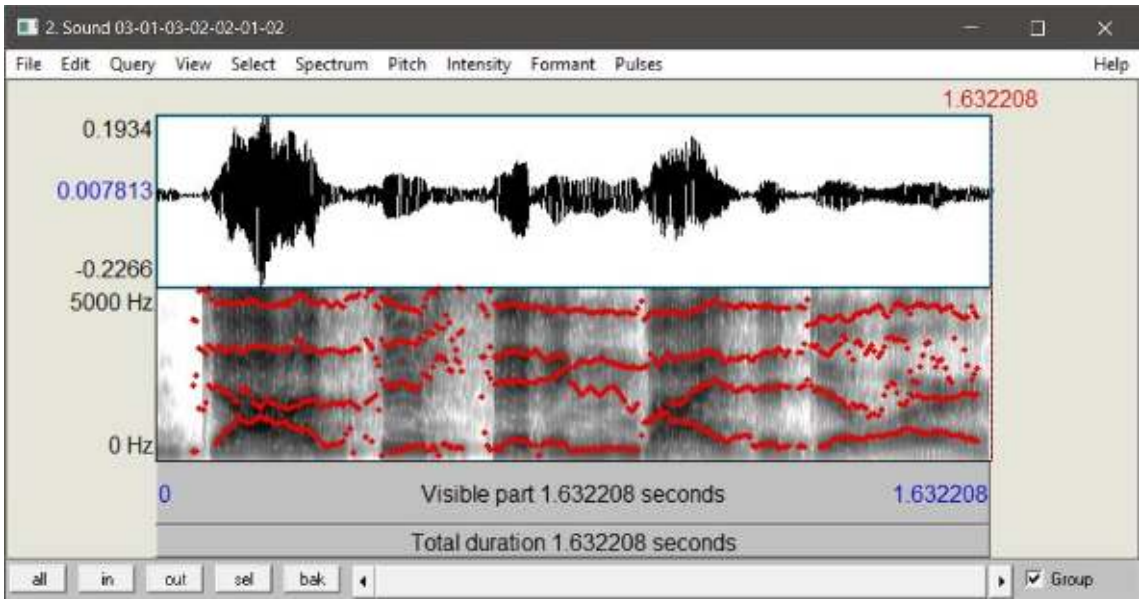
Predpokladá sa, že emocionálny obsah v reči súvisí s kvalitou hlasu. Experimenty s vnímaním ľudského hlasu ukázali silné prepojenie medzi kvalitou hlasu a vnímanou emóciou (El Ayadi, Kamel, Karray, 2011). V širšom význame pojem kvalita hlasu znamená charakteristické zafarbenie reči jednotlivca. Zmenou kvalitatívnych charakteristík je možné dať najavo dôležitú informáciu, napr. úmysly, emócie a postoje (Gangamohan, Kadiri, Yegnanarayana, 2016). Kvalitatívne charakteristiky sú úzko späté s prozodickými charakteristikami (Keller, 2004).

Medzi kvalitatívne charakteristiky patrí jitter, shimmer a ďalšie mikroprozodické javy. Odrážajú vlastnosti hlasu ako napr. dýchavičnosť a chrapľavosť (Batliner et al.,

2011). Perturbácia[1] základnej frekvencie (jitter) označuje kolísanie v periódach základnej frekvencie. Na výpočet tejto perturbácie existuje množstvo metód. Najjednoduchšou je priemerný jitter, ktorý je definovaný ako priemerný absolútny rozdiel dĺžky za sebou idúcich periód. Jitter sa bežne vyjadruje v percentách. Amplitúdová perturbácia (shimmer) je definovaná ako kolísanie v amplitúdach priľahlých periód. Podobne ako pre jitter, aj pre shimmer existuje množstvo rôznych spôsobov výpočtu. Najbežnejším je priemerný shimmer – priemerný absolútny rozdiel v amplitúdach za sebou idúcich periód (Hillenbrand, 2011).

1.1.3 Spektrálne charakteristiky

Spektrálne charakteristiky opisujú spektrum reči, ktoré je vyššie ako základná frekvencia – napríklad harmonické a formantové frekvencie (Steidl, 2008). Harmonické frekvencie sú celočíselné násobky základnej frekvencie – druhá harmonická frekvencia je $2*F_0$, tretia harmonická frekvencia je $3*F_0$, atď. (Wilder, 1975). Formantové frekvencie sú zosilnením určitých frekvencií v spektre. Toto zosilnenie vyplýva z rezonancie rečového ústrojenstva. Formanty sú charakterizované frekvenciou, amplitúdou a šírkou pásma (Steidl, 2008). Formantové frekvencie sú obzvlášť prítomné v samohláskach. V spektre sa formant vyskytuje približne každých 1000 Hz (Abhang, Gawali, Mehrotra, 2016).



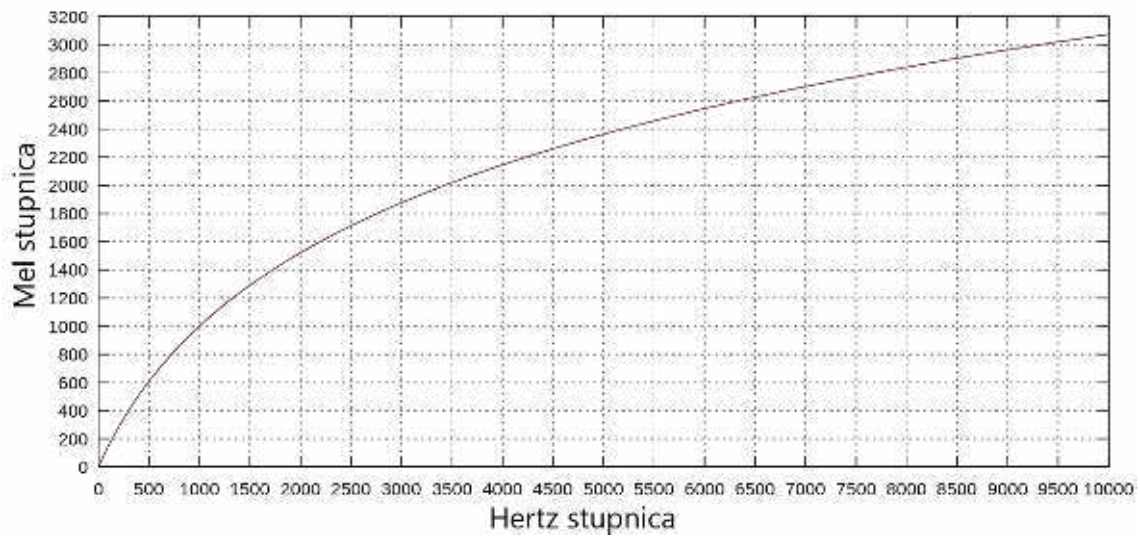
Obrázok 1: Zobrazenie prvých 4 formantov v programe PRAAT.

[1] Rušivý vplyv spôsobujúci nepravidelnosti.

Na rozpoznávanie emócií môžu byť použité aj ďalšie spektrálne charakteristiky, napr. mel frekvenčné kepstrálne[2] koeficienty (MFCC). Alternatívou k MFCC sú lineárne prediktívne kepstrálne koeficienty (LPCC) alebo banka mel filtrov (MFB) (Steidl, 2008).

MFCC a MFB

Mel stupnica je výsledkom experimentov s vnímaním pomocou sluchu. Mel je jednotka založená na spôsobe, akým ľudské ucho vníma frekvencie. Stupnica rozdeľuje frekvencie pod 1 kHz takmer lineárne, vyššie frekvencie sú na stupnici rozdeľované logaritmicky (Mahalakshmi, 2016).



Obrázok 2: Vzťah medzi frekvenciou vyjadrenou v hertzoch a v meloch[3].

MFCC modelujú rozdelenie energie v spektre reči tak, aby sa priblížilo k spôsobu, akým ho vníma ľudské ucho (Menon, Anjusha, 2017). Detailný postup extrakcie MFCC opísali (Rao, Reddy, Maity, 2015). Práve z tohto opisu sme v nasledujúcej časti vychádzali. Prvým krokom je zvýraznenie vyšších frekvencií. Účelom je vyrovnať spektrum znelých[4] zvukov, ktoré sú po nahratí mikrofónom vo vyšších frekvenciách utlmené približne o 6 dB oproti skutočnému spektru. Princíp extrakcie spočíva v rozdelení signálu na rámce, a aplikovaní diskrétnej Fourierovej transformácie, ktorá rámce konvertuje na magnitúdové spektrum. Výsledok Fourierovej transformácie prejde

[2] Kepstrum je inverzia spektra a jeho základná jednotka je kvefrencia (Batliner et al., 2011).

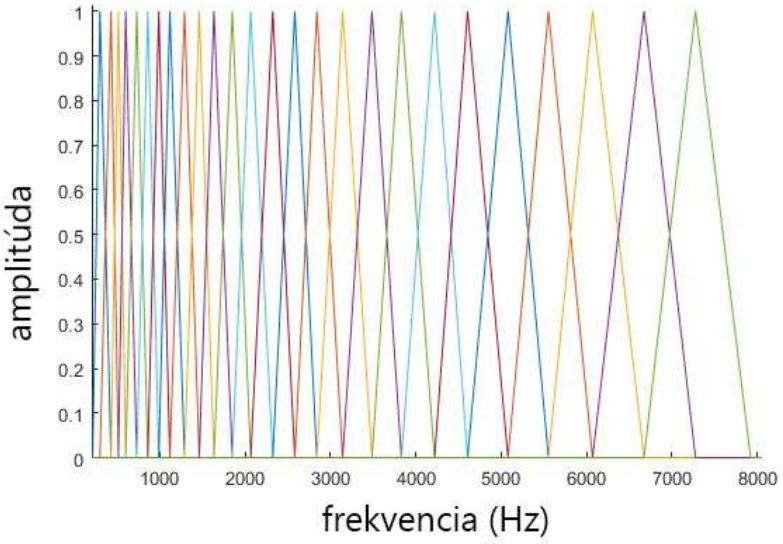
[3] en.wikipedia.org/wiki/Mel_scale#/media/File:Mel-Hz_plot.svg

[4] Znelé zvuky sú tie, ktoré pri ich vytváraní vyžadujú zapojenie hlasiviek.

cez sadu pásmových filtrov, ktorá sa nazýva banka mel filtrov. Najčastejšie používaný filter je trojuholníkový filter, avšak v niektorých prípadoch sa používa aj Hanningov filter. Stred frekvencií filtrov by bol za normálnych okolností rozložený rovnomerne, avšak pre napodobnenie vnímania ľudského ucha je os frekvencie daná nasledujúcou nelineárnou funkciou, v ktorej f je skutočná frekvencia v Hz, a f_{mel} je vnímaná frekvencia:

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700})$$

Magnitúdové spektrum je následne prepočítané na mel spektrum tak, že je vynásobené každým z trojuholníkových mel filtrov.



Obrázok 3: Banka mel filtrov s 26 filtrami[5].

Extrakcia je zakončená diskrétnou kosínusovou transformáciou. Väčšina informácií zo signálu je reprezentovaná v prvých niekoľkých MFCC. Zvyčajne sa používa 8-13 kepstrálnych koeficientov. Kepstrálne koeficienty sa často označujú ako statické charakteristiky, pretože obsahujú iba informáciu z rámca, pre ktorý boli vypočítané. Dynamické charakteristiky je možné získať vypočítaním ich prvej a druhej derivácie. Prvá derivácia prináša informáciu o tempe reči, a druhá derivácia o zrýchleniach.

Namiesto počítania MFCC je možné pre rozpoznávanie emócií použiť výstup priamo z banky filtrov. Touto alternatívou sa zaoberali (Busso, Lee, Narayanan, 2007). Ich experimenty porovnávali skryté Markovove modely, ktoré boli trénované s výstupom z MFB, a modely trénované s MFCC. V experimentoch boli úspešnejšie práve modely

[5] researchgate.net/figure/Mel-filter-bank-with-26-filters_fig2_328819233

trénované pomocou výstupu z MFB, avšak experimenty zahŕňali iba binárnu klasifikáciu, a klasifikáciu medzi 4 emóciami.

LPCC

Lineárne prediktívne kódovanie (LPC) je efektívnou metódou na spracovanie zvuku a reči. Hodí sa na extrakciu parametrov reči, ako napr. formanty a spektrum, pretože veľmi dobre charakterizuje rečové ústrojenstvo (Sunny, Peter, Jacob, 2012). LPC vychádza z myšlienky, že každá vzorka reči *s(n)* môže byť vyjadrená ako lineárna kombinácia predchádzajúcich *p* vzoriek:

$$s(n) \approx a_1 s(n-1) + 2s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p)$$

Predpokladá sa, že a_1, a_2, a_3, ... sú konštantné v celom rámci. Označujú sa ako prediktorové koeficienty alebo koeficienty lineárnej predikcie. Koeficienty sú vypočítané postupným znižovaním rozdielu medzi vzorkami z predikcie a skutočnými vzorkami (Rao, Reddy, Maity, 2015).

Lineárne prediktívne kepstrálne koeficienty (LPCC) sú koeficienty lineárnej predikcie prenesené do kepstra. LPCC sa stali jednou z najpoužívanejších techník pre vyhodnocovanie základných parametrov signálu reči. Extrakcia prebieha podobne ako pri MFCC. Vyššie frekvencie sú zvýraznené, a signál je rozdelený na rámce. Na každý rámec sa použije LPC a vypočítajú sa koeficienty. Posledným krokom je kepstrálna analýza – proces hľadania kepstra sekvencie reči. Pri hľadaní kepstra existujú dva rôzne postupy – FFT kepstrum a LPC kepstrum. V prvom prípade je kepstrum definované ako inverzná rýchla Fourierova transformácia logaritmu magnitúdového spektra reči. Druhá možnosť je odhad kepstrálnych koeficientov pomocou rekurzie – takto získané koeficienty sa nazývajú LPCC (Menon, Anjusha, 2017). (Rao, Reddy, Maity, 2015) uvádzajú postup rekurzie nasledovne:

$$C_0 = \log_e p$$

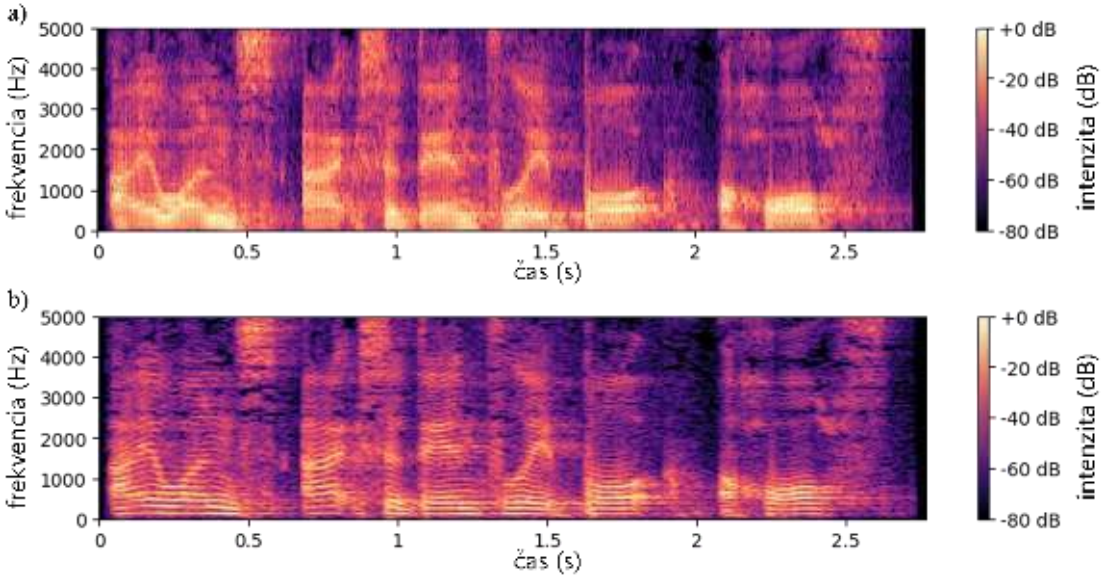
$$C_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad pre\ 1 < m < p$$

$$C_m = \sum_{k=m-p}^{m-1} \frac{k}{m} C_k a_{m-k}, \qquad pre\ m > p$$

1.1.4 Vizualizácia zvuku pomocou spektrogramov

Spektrogram je vizuálna reprezentácia rozdelenia akustickej energie naprieč frekvenciami v čase. Vodorovná os reprezentuje čas, a zvislá os reprezentuje frekvenciu. Intenzita frekvencie v danom čase je znázornená farbou spektrogramu – frekvencie s nízkou magnitúdou sú znázornené tmavou farbou, a svetlé farby znázorňujú frekvencie s vysokou magnitúdou (Hajarolasvadi, Demirel, 2019). V čiernobielych spektrogramoch tmavé farby naopak znázorňujú vysokú magnitúdu (Loizou, 2013). Spektrogram je možné získať použitím krátkodobej Fourierovej transformácie. Nahrávka reči sa najprv rozdelí na krátke rámce rovnakej dĺžky. Použitím Fourierovej transformácie je z každého rámca získané spektrum (frekvencie prítomné v rámci). Spektrogram potom vznikne vizualizáciou zmien spektra v čase (Hajarolasvadi, Demirel, 2019).

Zmenou dĺžky rámca je možné ovplyvňovať časové a frekvenčné rozlíšenie spektrogramu. Ak je rámec relatívne krátky, spektrogram má dobré časové rozlíšenie, ale nízke frekvenčné rozlíšenie. Výsledný spektrogram sa volá širokopásmový spektrogram. Pri zväčšovaní dĺžky rámca sa zlepšuje frekvenčné rozlíšenie na úkor časového rozlíšenia. Spektrogramy získané s použitím relatívne dlhého rámca sa nazývajú úzkopásmové. Kvôli rozdielnemu časovému a frekvenčnému rozlíšeniu sú širokopásmové a úzkopásmové spektrogramy používané na rozdielne úlohy. Širokopásmové spektrogramy sú vhodné na určovanie hranice slov a formantových frekvencií. Úzkopásmové spektrogramy sú preferované pri určovaní základnej frekvencie (Cheung, Lim, 1992).



Obrázok 4: a) širokopásmový spektrogram, b) úzkopásmový spektrogram.

1.2 NÁSTROJE NA SPRACOVANIE ZVUKOVÝCH SIGNÁLOV

V predošlej kapitole sme opísali hlasové charakteristiky, ktoré je možné využiť pri klasifikácii emócií. Teraz uvedieme krátky prehľad niekoľkých dostupných open source nástrojov, ktoré boli za účelom extrakcie charakteristík z reči vytvorené. Informácie o nich sme čerpali z dostupnej dokumentácie.

openSMILE – nástroj napísaný v jazyku C++ slúžiaci na extrakciu množstva prozodických a spektrálnych charakteristík. Zahrnutie knižnice PortAudio rozširuje možnosti tohto nástroja o použitie vstupných zariadení a extrakciu charakteristík v reálnom čase (Eyben et al., 2016).

Praat – holandský program vyvinutý na Amsterdamskej univerzite. Je komplexným a široko používaným nástrojom na analýzu reči. Tento program bol napísaný v jazykoch C a C++. Umožňuje extrahovať veľké množstvo charakteristík (napr. základnú frekvenciu, formantové frekvencie, intenzitu, jitter, shimmer, MFCC, a veľa ďalších). Ďalej umožňuje spektrálnu analýzu (zobrazenie spektrogramov), segmentáciu, syntézu reči, a obsahuje aj nástroje pre strojové učenie (napr. k-NN klasifikátor a neurónové siete) (Boersma, 2001).

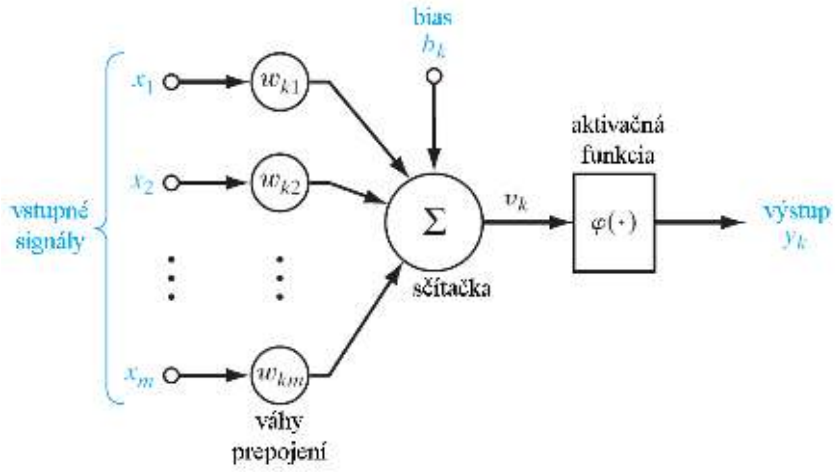
Parselmouth – Python knižnica, ktorá si za svoj cieľ kladie vytvorenie rozhrania, ktoré prenesie kompletnú funkcionalitu programu Praat do jazyka Python. Táto knižnica je založená na priamom prístupe do C a C++ kódu programu Praat, a jej použitie preto zaručuje rovnaké výsledky, ako pri použití programu Praat. Nevýhodou tejto knižnice je, že je ešte stále vo vývoji (súčasná verzia v čase písania tejto práce je 0.33), a jej funkcionalita je teda značne obmedzená (Jadoul, Thompson, De Boer, 2018).

Librosa – Python knižnica, ktorá poskytuje množstvo komplexných funkcií. Zameriava sa predovšetkým na extrakciu spektrálnych charakteristík, ale dokáže extrahovať aj charakteristiky súvisiace s rytmikou a tempom. Pomocou tejto knižnice je možné vytvoriť vizuálnu reprezentáciu zvukových signálov – spektrogramy, chromagramy a tempogramy. Takisto obsahuje funkcie na ďalšie spracovanie charakteristík – napr. počítanie ich derivácií. Užitočné sú aj funkcie orezania ticha, segmentácie, manipulácie s výškou frekvencie vstupného signálu a mnohé iné (McFee at al., 2015).

pyAudioAnalysis – Python knižnica, ktorá dokáže z nahrávok extrahovať 34 charakteristík – energiu, zerocrossing, MFCC, a ďalšie charakteristiky. Knižnica umožňuje aj vizualizáciu pomocou spektrogramu a chromagramu. Okrem extrakcie charakteristík poskytuje funkcionalitu aj na nahrávanie, segmentáciu zvukového signálu, a takisto implementuje niekoľko klasifikátorov (napr. k-NN, SVM, rozhodovacie stromy) pre klasifikačné a regresné úlohy (Giannakopoulos, 2015).

1.3 VYUŽITIE NEURÓNOVÝCH SIETÍ NA KLASIFIKÁCIU EMÓCIÍ

Ľudský mozog je tvorený veľkým počtom navzájom prepojených neurónov. Neurón je bunka, ktorá dokáže vykonať jednoduchú úlohu – napríklad reagovať na vstupný signál. Ak sa prepojí množstvo neurónov, môžu veľmi rýchlo a presne vykonávať zložité úlohy, ako rozpoznávanie reči a obrázkov. Umelé neurónové siete sú vytvorené na základe biologických neurónových sietí. Umelá neurónová sieť je prepojením uzlov, ktoré sú analogické s neurónmi (Zou, Han, So, 2008).



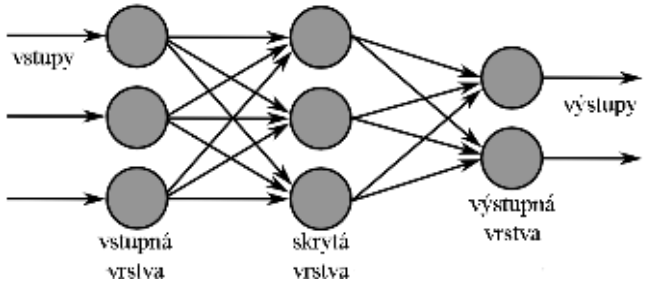
Obrázok 5: Model neurónu (Haykin, 2010).

Na obrázku vyššie (Obrázok 5) je znázornený model neurónu, ktorý je základnou stavebnou jednotkou neurónových sietí. Tento neurón budeme ďalej označovať ako *k*. Tri základné prvky neurónu sú synapsie, sčítačka a aktivačná funkcia. Synapsie (prepojenia) sú charakterizované ich váhou, resp. silou prepojenia. Signál x_j na vstupe synapsie, ktorá vedie k neurónu *k*, je vynásobený váhou spojenia w_{kj}. Váha umelých neurónov, na rozdiel od biologických, môže byť aj záporná. Sčítačka sčíta všetky vstupné signály vynásobené k nim prislúchajúcimi váhami prepojení. Aktivačná funkcia slúži na obmedzenie možného rozsahu výstupného signálu na konečné hodnoty, najčastejšie na interval [0,1] alebo [-1,1]. Model na obrázku zahŕňa aj externe aplikovaný bias, označený ako b_k. Bias,

v závislosti od jeho hodnoty, zvyšuje alebo znižuje vstup do aktivačnej funkcie (Haykin, 2010).

1.3.1 Dopredné neurónové siete

Vo vrstvených neurónových sieťach sú neuróny usporiadané do vrstiev. Najjednoduchšou vrstvenou sieťou je sieť, v ktorej je vstupná vrstva priamo spojená s výstupnou vrstvou. Spojenie je jednosmerné – takáto sieť sa označuje ako dopredná. Ďalej sa označuje ako jednovrstvová – vo vstupnej vrstve neprebiehajú žiadne výpočty, preto sa nepočíta do celkového počtu vrstiev. Viacvrstvové dopredné neurónové siete sa od jednovrstvových odlišujú prítomnosťou jednej alebo viacerých skrytých vrstiev. Tento pojem označuje vrstvu, ktorá nie je vstupom ani výstupom. Po pridaní skrytých vrstiev môže neurónová sieť zo vstupu odvodiť štatistiky vyššieho rádu – zjednodušene povedané, môže získať širšiu perspektívu. Každá vrstva obvykle ako svoj vstup používa iba výstup z predchádzajúcej vrstvy. Ak je v každej vrstve každý neurón prepojený s každým neurónom z ďalšej vrstvy, hovoríme o plne prepojenej neurónovej sieti. Čiastočne prepojená neurónová sieť označuje sieť, v ktorej niektoré synaptické prepojenia chýbajú (Haykin, 2010).



Obrázok 6: Dopredná plne prepojená viacvrstvová neurónová sieť.[6]

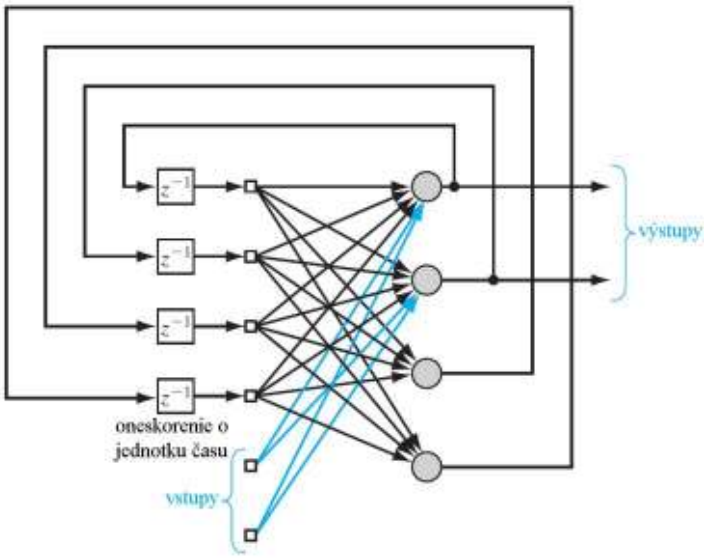
(Han, Yu, Tashev, 2014) využili architektúru dopredných neurónových sietí na klasifikáciu emócií. Z nahrávok 5 emócií boli extrahované globálne štatistiky. Najlepší dosiahnutý výsledok klasifikácie bol 54,3%. (Shih, Chen, Wang, 2017) sa pokúsili navrhnúť algoritmus trénovania neurónovej siete, ktorého cieľom je vylepšiť učenie, ak počet údajov v triedach nie je uniformný. Pri trénovaní použili kombináciu lokálnych a globálnych charakteristík – výsledný vektor mal 384 atribútov. Ich dopredná neurónová sieť obsahovala jednu skrytú vrstvu, bola trénovaná na nahrávkach detí, a dokázala

[6] commons.wikimedia.org/wiki/File:MultiLayerNeuralNetworkBigger_english.png

klasifikovať 5 rôznych emócií so 45,3% úspešnosťou. Klasifikátory, ktoré navrhnutý algoritmus nepoužívali, dosahovali úspešnosť nižšiu o viac ako 10%. (Shaw, Vardhan, Saxena, 2016) klasifikovali so svojou doprednou sieťou 4 emócie (hnev, šťastie, smútok, neutrálna emócia) s 86,87% úspešnosťou. Na trénovanie modelu používali globálne charakteristiky vypočítané z energie, základnej frekvencie, formantových frekvencií a MFCC. Doprednú sieť na klasifikáciu 3 emócií (hnev, šťastie, smútok) v tamilčine a telugčine (indické jazyky) použili (Renjith, Manju, 2017). Zo vzoriek reči extrahovali LPCC a Hurstove parametre. Model dosiahol 75,27% úspešnosť pre tamilčinu a 76,41% pre telugčinu. (Huang et al., 2016) sa pokúsili o klasifikáciu medzi 8 emóciami. Vo svojej práci používali 64 charakteristík, z ktorých boli vypočítané ďalšie globálne štatistiky a derivácie. Uvedená úspešnosť klasifikácie je 44,22%.

1.3.2 Rekurentné neurónové siete

Rekurentné siete sa od dopredných odlišujú tým, že obsahujú aspoň jedno spätné prepojenie. Rekurentná sieť môže napríklad pozostávať z jednej vrstvy neurónov, v ktorej je výstup každého neurónu pripojený späť na vstup všetkých ostatných neurónov, prípadne aj späť na svoj vlastný vstup. Prítomnosť spätných prepojení má veľký vplyv na schopnosť učenia a výkon. Spätné prepojenia zahŕňajú vetvy, ktoré sú tvorené prvkami oneskorenými v čase. Ich dôsledkom je nelineárne dynamické správanie (Haykin, 2010). Rekurentnými sieťami sú napríklad Hopfieldova sieť alebo Kohonenove mapy (Zou, Han, So, 2008).



Obrázok 7: Rekurentná neurónová sieť so skrytou vrstvou (Haykin, 2010).

Rekurentnú sieť na rozpoznávanie 4 rôznych emócií (hnev, vzrušenie, smútok, neutrálna emócia) použili (Chernykh, Prikhodko, 2018). Z reči extrahovali 34 lokálnych charakteristík (13 MFCC, energia, zerocrossing, a ďalšie). Vo svojej práci dosiahli 54% úspešnosť klasifikácie. (Tzinis, Potamianos, 2017) extrahovali 24 lokálnych charakteristík, a k niektorým pridali ich derivácie, čím získali celkovo 47 charakteristík. Rekurentný model obsahoval 2 skryté vrstvy a jeho úlohou bola klasifikácia 4 emócií (hnev, radosť, smútok, neutrálna emócia). Model správne klasifikoval 59,14% emócií. V druhom experimente boli z 3 sekundových segmentov globálne štatistiky, a rovnaký model dosiahol úspešnosť 64,16%. Rekurentné neurónové siete použili aj (Mirsamadi, Barsoum, Zhang, 2017). 4 klasifikované emócie boli rovnaké, ako v predošlej spomínanej práci. Experimentovali s modelmi trénovanými na 257 vektoroch, ktoré boli výstupom z rýchlej Fourierovej transformácie, a s modelmi trénovanými na 32 extrahovaných lokálnych charakteristikách (F_0, 12 MFCC, zerocrossing, a ďalšie). Najlepší výsledok bol dosiahnutý s použitím lokálnych charakteristík – 63,5%. (Feng, Yang, 2018) pomocou waveletovej analýzy získali až 136 charakteristík, a ich rekurentný model v klasifikácii 7 emócií dosiahol úspešnosť 86%.

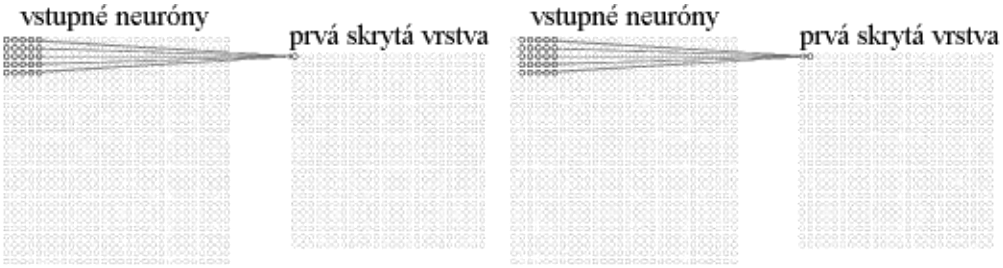
1.3.3 Konvolučné neurónové siete

Ďalším typom neurónových sietí sú konvolučné neurónové siete. Neuróny v konvolučných sieťach môžu byť usporiadané v 1D, 2D alebo 3D priestore. Najbežnejším použitím je spracovanie obrázkov (2D siete). 1D konvolúcie sú vhodné na časovú postupnosť údajov, a 3D vstupom môže byť napríklad magnetická rezonancia (Vasilev et al., 2019). Základný princíp konvolučných neurónových sietí tvoria lokálne recepčné polia, zdieľané váhy a pooling[7] (Nielsen, 2015). Posledný stupeň konvolučnej siete môže pozostávať z obyčajnej viacvrstvovej siete – neuróny z poslednej vrstvy konvolučnej siete sú pripojené ako vektor na vstup prvej plne prepojenej vrstvy (Albawi, Mohammed, Al-Azawi, 2017).

Každý neurón v skrytej vrstve je pripojený k malej oblasti vstupných neurónov (napr. oblasť 5x5 – v prípade obrázka 25 pixelov). Tieto oblasti sa nazývajú lokálne recepčné polia. Recepčné pole sa potom postupne posúva po celej vstupnej matici –

[7] Pooling sa niekedy označuje aj ako subsampling (podvzorkovanie).

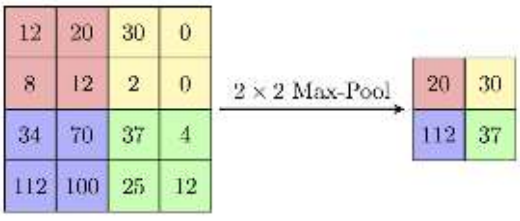
neurón v skrytej vrstve sa učí analyzovať to lokálne recepčné pole, ku ktorému je pripojený.



Obrázok 8: Lokálne recepčné pole a posun o jeden neurón (Nielsen, 2015).

Váhy a bias sú pre každý neurón v skrytej vrstve spoločné – všetky neuróny slúžia na detekciu rovnakej vlastnosti v rôznych častiach vstupu, a tvoria príznakovú mapu. Ak neurón zo skrytej vrstvy dokáže v jeho recepčnom poli rozoznať vlastnosť (napr. vertikálnu hranu), rovnaká schopnosť môže byť užitočná aj na iných miestach. Zdieľané váhy a bias sa spoločne zvyknú označovať ako kernel alebo filter. Príznaková mapa je vypočítaná posúvaním filtra. Počet vlastností, ktoré dokáže konvolučná sieť detegovať je daný počtom filtrov (napr. s 3 filtrami je možné vytvoriť príznakové mapy 3 vlastností) (Nielsen, 2015). Filter má vždy rovnaký počet rozmerov ako vstupné údaje (Vasilev et al., 2019).

Poolingové vrstvy sa obvykle umiestňujú za konvolučné vrstvy. Ich účelom je zjednodušenie výstupu z konvolučnej vrstvy. Poolingová vrstva vezme príznakovú mapu z konvolučnej vrstvy, a vytvorí z nej zmenšenú mapu. Každá jednotka v pooling vrstve zovšeobecňuje určitú oblasť (napr. 2x2 neurónov) z predchádzajúcej vrstvy. Jedným z niekoľkých bežných spôsobov poolingu je max-pooling – výstup jednotky poolingu je hodnota najvyššej aktivácie v oblasti. Konvolučná vrstva obvykle zahŕňa viac príznakových máp – pooling sa aplikuje na každú mapu zvlášť (Nielsen, 2015).



Obrázok 9: Princíp max-poolingu.[8]

[8] researchgate.net/figure/Exemple-dune-operation-de-max-pooling_fig4_337635267

(Satt, Rozenberg, Hoory, 2017) použili spektrogramy na natrénovanie konvolučnej siete. V klasifikácii 4 emócií (hnev, šťastie, neutrálna emócia, smútok) čisto konvolučný model dosiahol úspešnosť 66,1%, a model, ktorý kombinoval konvolučnú sieť s jednou rekurentnou vrstvou emócie správne určil v 68,8% prípadov. Klasifikáciu pomocou spektrogramov zvolili aj (Etienne et al., 2018), a ich model pozostávajúci zo 4 konvolučných vrstiev, ktoré boli nasledované rekurentnou vrstvou dosiahol 64,5% úspešnosť klasifikácie medzi 4 emóciami. Pozoruhodný výsledok s použitím hlbokých retinálnych konvolučných sietí dosiahli (Niu et al., 2017). Vstupom sietí boli taktiež spektrogramy, a v klasifikácii 8 emócií dosiahli priemernú úspešnosť viac ako 99%. (Qayyum, Arefeen, Shahnaz, 2019) z nahrávok reči extrahovali 12 MFCC a porovnali niekoľko rôznych klasifikátorov. Najlepšie výsledky dosiahla 1D konvolučná sieť – 7 emócií dokázala klasifikovať s 83,61% úspešnosťou. O porovnanie úspešnosti klasifikácie s použitím rôznych charakteristík a klasifikátorov sa pokúsili aj (Thakare et al., 2019). V porovnaní sa ako najlepší ukázal model 5 vrstvovej konvolučnej neurónovej siete trénovaný s MFCC (počet koeficientov nie je v práci špecifikovaný), ktorý v klasifikácii 7 emócií dosiahol 72,92% úspešnosť. Rovnaký model bol natrénovaný aj s výstupom z mel banky filtrov, v tomto prípade dosiahol o niečo nižšiu úspešnosť – 65,63%. Iný prístup zvolili (Latif et al., 2019). Namiesto extrakcie charakteristík sa vstupom do konvolučnej siete stali nespracované nahrávky. Architektúra siete obsahovala paralelné konvolučné vrstvy, rekurentnú vrstvu, a zakončená bola plne prepojenou doprednou vrstvou. Ich model dosiahol 60,23% úspešnosť, a klasifikoval 4 rôzne emócie (hnev, radosť, smútok, neutrálna emócia).

1.4 KLASIFIKÁCIA EMÓCIÍ V REÁLNOM ČASE

Napriek tomu, že v súčasnosti existuje mnoho prác zaoberajúcich sa strojovým rozpoznávaním emócií, len veľmi málo z nich venuje pozornosť klasifikácii v reálnom čase, a sústreďujú sa predovšetkým na hľadanie najvhodnejších charakteristík a klasifikátorov. V predchádzajúcej kapitole sme sa zaoberali takýmito pokusmi, ktoré používali neurónové siete. Teraz rozoberieme významné pokusy o klasifikáciu v reálnom čase, a riešenia, ktoré už sú používané v praxi.

(Kim et al., 2007) navrhli spôsob binárnej klasifikácie (hnev a neutrálna emócia) v reálnom čase. Ich aplikácia, naprogramovaná v jazyku C++, pozostáva zo 4 hlavných častí – prúd vstupných údajov, extrakcia charakteristík, modely strojového učenia

a zlúčenie výstupu modelov. Vstup je rozdelený na krátke segmenty (1-5 sekúnd). Zo segmentov sú následne extrahované charakteristiky – 12 MFCC, základná frekvencia a energia. MFCC sú ďalej vstupom pre Gaussove zmiešavacie modely, a z prozodických charakteristík sa vypočítajú globálne charakteristiky, ktoré sú spracované k-NN modelom. Výsledné pravdepodobnosti sú na záver zlúčené matematickým algoritmom. Pri experimentovaní s rôznym nastavením parametrov bola najnižšia dosiahnutá chybovosť v určovaní 4,75%.

(Stolar et al, 2017) vytvorili vlastné experimentálne riešenie v prostredí Matlab. Riešenie dokáže klasifikovať 7 rôznych emócií. Prúd vstupných údajov sa delí na 1 sekundové bloky. Z blokov sú vypočítané spektrogramy, ktoré sú následne prevedené do formy farebných obrázkov, a rozdelené na separátne RGB kanály. (Stolar et al, 2017) experimentovali s konvolučnými neurónovými sieťami, a kombináciou konvolučnej siete a SVM klasifikátora. Napriek tomu, že toto riešenie zvláda klasifikáciu v reálnom čase, používa modely, ktoré boli trénované separátne pre mužov a ženy. Použitý model konvolučnej siete pri testovaní dosiahol úspešnosť 79,68% pre ženy a 76,79% pre mužov, kombinácia konvolučnej siete a SVM bola v klasifikácii menej presná – 73,84% pre ženy a 72,73% pre mužov.

(Cen et al., 2015) sa zaoberali klasifikáciou emócií v reálnom čase a jej využitím v online učení. Zatiaľ čo predchádzajúce riešenia používali segmentáciu na fixnú dĺžku, v tomto prípade algoritmus na základe energie deteguje hlasovú aktivitu, a reč delí na segmenty, ktoré môžu byť dlhé 3 alebo viac sekúnd. Z každého segmentu je extrahovaný vektor 132 charakteristík (18 PLP, 12 MFCC, 13 LPCC, a ich prvé a druhé derivácie). Použitý je SVM klasifikátor, ktorý dokáže klasifikovať 4 emócie (hnev, šťastie, smútok, neutrálna emócia). Aplikácia s grafickým rozhraním umožňuje vstup z mikrofónu aj zo súboru. Klasifikátor dosiahol 90% úspešnosť pri analyzovaní hotových nahrávok, a 78,78% úspešnosť pri teste vyhodnocovania v reálnom čase.

Pre online učenie bol vytvorený aj framework FILTWAM. Jeho cieľom je poskytovať okamžitú a adekvátnu spätnú väzbu na základe údajov z mikrofónu a webkamery počas toho, ako učiaci sa interaguje s online učebnými materiálmi. Vstup je na základe páuz dlhších ako 1 sekunda rozdelený na segmenty, z ktorých sú následne extrahované charakteristiky (podrobnosti o charakteristikách autori neuvádzajú). 7 emócií je klasifikovaných pomocou SMO klasifikátora. Úspešnosť bola vyhodnotená

v štúdii s 12 účastníkmi, a dosiahla 67%. Ďalšie pokusy naznačujú, že úspešnosť sa významne zvýši, ak sú pre klasifikáciu okrem reči zároveň použité zábery tváre (Bahreini, Nadolski, Westera, 2016).

Pravdepodobne najkomplexnejším existujúcim nástrojom je Social Signal Interpretation framework (SSI). SSI je open source framework naprogramovaný v jazykoch C a C++. Jeho účelom je spracovanie rôznych signálov (napr. video, reč, tep) v reálnom čase. Podporuje veľké množstvo senzorov, a obsahuje filtre, algoritmy na extrakciu charakteristík, a nástroje na strojové učenie a rozpoznávanie vzorov. Jednou z jeho súčastí je EmoVoice[9] (Wagner et al., 2013). Princíp fungovania EmoVoice opísali (Vogt, André, Bee, 2008), avšak niektoré informácie už nie sú aktuálne, preto ich doplníme informáciami z online dokumentácie[10]. EmoVoice je nástroj na offline aj online rozpoznávanie emócií z reči v reálnom čase, a poskytuje funkcionalitu na segmentáciu, extrakciu charakteristík a klasifikáciu. Reč je segmentovaná na základe detekcie hlasovej aktivity – pauzy dlhšie ako 200 ms oddeľujú segmenty. Okrem toho je možné nastaviť maximálnu dĺžku segmentu, obvykle 2-3 sekundy. Tento nástroj z hlasu dokáže extrahovať veľké množstvo rôznych charakteristík, a zároveň poskytuje možnosť výberu najrelevantnejších z nich (napr. analýzou korelácie). EmoVoice v sebe integruje 2 klasifikátory – súčasná online dokumentácia uvádza SVM a lineárny SVM. (Vogt, André, Bee, 2008) odporúčajú zvolený klasifikátor natrénovať v závislosti od situácie, v ktorej bude používaný, a poskytujú rozhranie, ktoré túto úlohu značne uľahčuje.

[9] github.com/hcmlab/emovoice

[10] rawgit.com/hcmlab/emovoice/master/docs/index.html#audio-signal

2 CIELE DIPLOMOVEJ PRÁCE

Hlavným cieľom tejto diplomovej práce je vytvoriť vlastné riešenie vo forme aplikácie, ktorá zo vstupného zariadenia (napr. z mikrofónu) dokáže snímať reč používateľa, v reálnom čase s čo najmenším oneskorením tento vstup spracovávať, a vykonávať predikciu emocionálneho stavu. Spracovanie by malo zahŕňať kroky, v ktorých je vstup vhodne upravený, a sú z neho extrahované charakteristiky, na základe ktorých potom klasifikátor (neurónová sieť) vykoná predikciu emócie.

ČIASTKOVÉ CIELE

- sumarizácia charakteristík, ktoré môžu byť extrahované z hlasu,
- sumarizácia architektúr neurónových sietí a ich využitia na klasifikáciu emocionálneho stavu z reči,
- sumarizácia existujúcich riešení pre klasifikáciu emocionálneho stavu v reálnom čase,
- analýza nahrávok emócií, ktoré budú neskôr použité na trénovanie vlastného klasifikátora,
- predspracovanie nahrávok a extrakcia charakteristík,
- testovanie rôznych typov architektúr neurónových sietí,
- výber najlepšieho modelu s ohľadom na rýchlosť a presnosť klasifikácie,
- návrh a vytvorenie vlastnej aplikácie s použitím vybraného modelu,
- zhodnotenie výsledkov a identifikácia krokov, ktoré môžu byť zlepšené,
- návrh postupu ďalšieho rozširovania.

3 POSTUP VLASTNÉHO RIEŠENIA A NÁVRH APLIKÁCIE

V praktickej časti tejto práce sme sa zaoberali vhodným postupom predspracovania dát. V experimentoch sme preskúmali spracovanie rôznych hlasových charakteristík, a hľadali sme vhodné architektúry neurónových sietí. Po testovaní sa najlepší model stal základom pre aplikáciu EmoRec 2, ktorú sme navrhli a naprogramovali ako vlastné riešenie rozpoznávania emócií v reálnom čase.

3.1 NADVIAZANIE NA PREDCHÁDZAJÚCU PRÁCU

V našej bakalárskej práci sme sa zaoberali pojmom emócia, a preskúmali sme schémy emócií navrhnuté rôznymi psychológmi. Rozhodli sme sa používať Ekmanovu schému, v ktorej je definovaných 7 emócií – hnev, znechutenie, strach, radosť, smútok, prekvapenie a neutrálna emócia. Ďalej sme sa zaoberali databázami reči, ktoré obsahujú nahrávky rôznych emócií, a ich dostupnosťou. Tieto znalosti nám teraz pomôžu pripraviť bázu údajov, s ktorou natrénujeme a otestujeme neurónovú sieť. Opísali sme niektoré charakteristiky merateľné na hlase, a využili sme program PRAAT na ich extrakciu z nahrávok. Potom sme sa pomocou ANOVA analýzy pokúsili vybrať charakteristiky s najvyššou varianciou – teda tie, ktoré najlepšie dokážu rozlíšiť emócie. Vytvorili sme program EmoRec, ktorý na rozlišovanie emócií používal jednoduchú formu k-NN algoritmu, avšak v testoch dosahoval iba nízku úspešnosť – 20 - 35% (Sulka, 2018).

Táto práca je jej priamym pokračovaním. Aj tentokrát budeme používať Ekmanovu schému. Zamerali sme sa na preskúmanie širšieho množstva charakteristík a spôsoby ich spracovania. Namiesto k-NN algoritmu sme tentokrát experimentovali s neurónovými sieťami. Oproti bakalárskej práci sme sa pokúsili vytvoriť robustnejšie riešenie, ktoré bude disponovať možnosťou určovania emócií v reálnom čase, a tým sa ešte viac priblíži k možnému budúcemu využitiu v praxi. Dôležitým krokom ku klasifikácii v reálnom čase bolo nahradenie programu PRAAT knižnicami pre jazyk Python, ktoré poskytujú rovnakú, alebo podobnú funkcionalitu, a zároveň sú jednoducho použiteľné. Možnosť extrakcie charakteristík priamo v našom vlastnom programe je veľkou výhodou oproti riešeniu, ktoré sme použili v bakalárskej práci, a práve z tohto dôvodu sme tentokrát zvolili jazyk Python. Hlavný program a všetky ostatné skripty boli vytvorené v IDE Visual Studio 2017.

3.2 ANALÝZA VSTUPNÝCH ÚDAJOV

Rozhodli sme sa, že vzhľadom na ich dostupnosť, a široké použitie v iných odborných prácach, na vypracovanie vlastného riešenia použijeme rovnaké databázy reči ako v bakalárskej práci. Týmito databázami sú:

- RAVDESS[11],
- EMO-DB[12],
- SAVEE[13].

Nahrávky vo všetkých troch databázach boli vytvorené prostredníctvom hercov – zaraďujú sa teda medzi simulované databázy. Pred predspracovaním údajov je vhodné najprv analyzovať aké údaje nám vybrané databázy reči poskytujú, a do akej miery spĺňajú nasledujúce, nami definované, kritériá:

- Triedy (emócie) obsiahnuté v databáze by mali reprezentovať emócie definované v Ekmanovej schéme,
- Počet nahrávok emócií by mal byť čo najväčší, a ich rozloženie v rámci tried uniformné,
- Počet hercov, ktorí sa na vytváraní databázy podieľali, by mal byť čo najväčší,
- Zastúpenie pohlaví a vekových kategórií by malo byť čo najširšie,
- Nahrávky v databáze by mali byť čo najkvalitnejšie,
- Databáza by mala obsahovať nahrávky čo najväčšieho počtu rôznych viet.

Proces predspracovania údajov sme prispôsobili s ohľadom na splnenie (prípadne nesplnenie) týchto kritérií.

RAVDESS

Databáza bola vytvorená na Ryersonovej univerzite v Toronte za pomoci 24 severoamerických hercov, 12 mužov a 12 žien, vo veku od 21 do 33 rokov. Nahrávky majú vzorkovaciu frekvenciu 48 kHz a bitovú hĺbku 16 bitov. Nahraných bolo 8 emócií – hnev, znechutenie, pokoj, strach, radosť, smútok, prekvapenie a neutrálna emócia. Táto databáza je audiovizuálna – obsahuje 7356 súborov medzi ktorými sú zábery tváre, spev a reč. Samotné nahrávky reči tvoria iba časť databázy – 1440 súborov. Herci

[11] smartlaboratory.org/ravdess/

[12] emodb.bilderbar.info/index-1024.html

[13] kahlan.eps.surrey.ac.uk/savee/

na nahrávkach predvádzali vždy jednu z dvoch významovo neutrálnych anglických viet – „Kids are talking by the door.“ a „Dogs are sitting by the door.“, pričom im zakaždým dodali požadovaný emocionálny výraz. Každú emóciu, okrem neutrálnej, herci nahrali s dvomi stupňami výrazu – slabý a výrazný prejav. Počet nahrávok neutrálnej emócie je teda oproti ostatným emóciám iba polovičný – pre každú emóciu existuje 192 nahrávok, pre neutrálnu emóciu iba 96 (Livingstone, Russo, 2018).

EMO-DB

Táto databáza sa zvykne označovať aj ako berlínska. Obsahuje nahrávky v nemeckom jazyku, ktoré nahralo 10 hercov, 5 mužov a 5 žien vo veku od 21 do 35 rokov. Vzorkovacia frekvencia nahrávok je 16 kHz, a bitová hĺbka je 16 bitov. V databáze sa strieda 10 rôznych viet. Pôvodne bolo vytvorených približne 800 nahrávok, a následne boli vyhodnotené testom vnímania s 20 účastníkmi. Každý účastník dostal tieto nahrávky v náhodnom poradí bez označenia emócie, ktorú má nahrávka reprezentovať, a každú nahrávku si mohol vypočuť iba raz. Z nahrávok, ku ktorým priradilo správnu emóciu viac ako 80% účastníkov, a zároveň ich viac ako 60% účastníkov označilo ako prirodzené, bola vytvorená finálna databáza, ktorá obsahuje 535 súborov. Emócie v tejto databáze sú hnev (127 súborov), znechutenie (46 súborov), strach (69 súborov), radosť (71 súborov), smútok (62 súborov), nuda (81 súborov) a neutrálna emócia (79 súborov) (Burkhardt et al., 2005).

SAVEE

Univerzita v Surrey vytvorila databázu SAVEE. Nahrávania sa nezúčastnili profesionálni herci, ale výskumní pracovníci a študenti postgraduálneho štúdia, 4 muži vo veku od 27 do 31 rokov. Nahrávky žien nie sú zahrnuté. Súbory sú nahraté so vzorkovacou frekvenciou 44,1kHz a s bitovou hĺbkou 16 bitov. Na nahrávkach je zachytených 15 viet pre každú emóciu – 3 vety spoločné pre všetky emócie, 2 unikátne vety pre každú emóciu a 10 unikátnych generických[14] foneticky vyvážených viet pre každú emóciu. Databáza sa drží Ekmanovej schémy, a obsahuje hnev, znechutenie, strach, šťastie, smútok, prekvapenie a neutrálnu emóciu. Pre každú emóciu bolo vytvorených 60 zvukových súborov, okrem neutrálnej emócie, pre ktorú bolo vytvorených 120 súborov. Zároveň s nahrávkami reči boli snímané aj tváre účastníkov (Jackson, Haq, 2011).

[14] Tzn. nemajú žiadny zmysel.

Rozbor štruktúry databáz reči ukázal negatívne vlastnosti, na ktoré je potrebné brať ohľad, a ktoré môžu ovplyvniť výsledky trénovania klasifikátora. Žiadna z databáz neobsahuje rovnaký počet nahrávok pre každú emóciu – to znamená, že všetky databázy porušujú uniformitu tried, pričom najvýraznejšie uniformitu porušuje EMO-DB, kde má každá emócia iný počet nahrávok. Všetky databázy majú pomerne málo nahrávok, najviac ich obsahuje databáza RAVDESS - 1440. SAVEE neobsahuje nahrávky žien. Keďže cieľom je vytvoriť klasifikátor, ktorý bude nezávislý od pohlavia, je pre jeho trénovanie nutné použiť databázu, v ktorej sú v rovnakom pomere zastúpené ženy aj muži – toto kritérium spĺňa iba RAVDESS. Žiadna z databáz nereprezentuje širšie spektrum vekových kategórií – databázy boli nahrané mladými ľuďmi. RAVDESS a EMO-DB obsahujú aj nahrávky emócií, ktoré nie sú definované v Ekmanovej schéme. V databáze RAVDESS je okrem 7 Ekmanových emócií navyše aj ôsma emócia – pokoj. EMO-DB neobsahuje emóciu prekvapenie, ale namiesto nej je tu emócia nuda. Nahrávky vo všetkých databázach sú dostatočne kvalitné, pričom RAVDESS poskytuje najkvalitnejší a najdetailnejší záznam hlasu, ale napriek tomu, že má najviac nahrávok zo všetkých databáz, zlyháva v rôznorodosti viet, ktoré herci predvádzali. Z uvedeného sme vyvodili záver, že kritériá definované na začiatku tejto kapitoly najviac spĺňa databáza RAVDESS.

3.3 PREDSPRACOVANIE ÚDAJOV

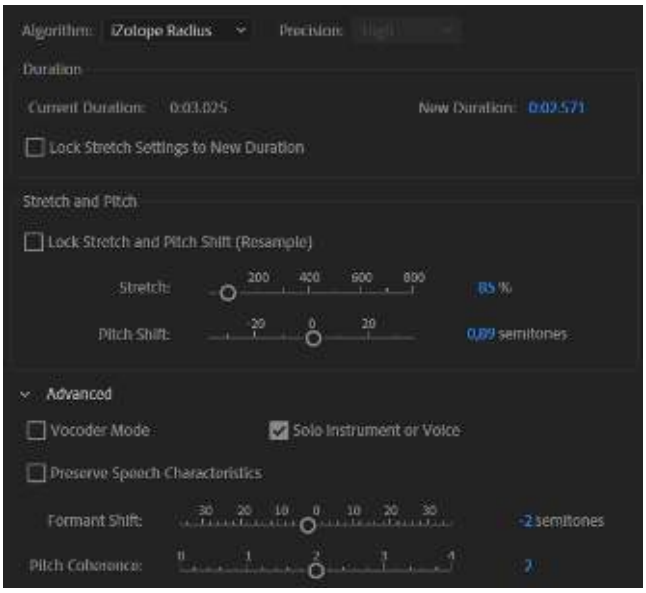
V procese predspracovania sme pripravili súbory na extrakciu rôznych hlasových charakteristík normalizáciou a orezaním ticha, a zároveň sme sa pokúsili napraviť niektoré zistené nedostatky. Časť predspracovania bola vykonaná manuálne, a časť bola zautomatizovaná pomocou skriptu *PreprocessData.py* (Príloha B). Z analýzy v predchádzajúcej kapitole vyplýva, že najväčšie problémy sú málo súborov a neuniformné triedy. Získavanie nových nahrávok (prípadne vytváranie vlastnej databázy) je časovo náročné, preto sme museli nájsť iný spôsob ako počet súborov rozšíriť. Rozhodli sme sa, že databázy nebudeme kombinovať – viedlo by to k ešte väčšiemu narušeniu uniformity tried. Namiesto toho sme sa pokúsili tieto problémy vyriešiť takými úpravami zvukových súborov, ktoré zmenia niektoré ich vlastnosti, ale zároveň zostane zachovaný ich emocionálny obsah. Takto sme získali nové súbory, ktoré sme pridali k pôvodným. Databázy obsahujú aj súbory, ktoré sú pre nás zbytočné – nahrávky emócie pokoj v databáze RAVDESS a emócie nuda v EMO-DB. Tieto súbory sme z ďalšieho spracovania vynechali.

Ďalej sme sa rozhodli, že na trénovanie neurónovej siete budeme používať iba databázu RAVDESS. Súbory tejto databázy sme zároveň rozdelili na trénovaciu a testovaciu podmnožinu v pomere približne 8:2. Uniformita tried a veľké množstvo údajov sú dôležité predovšetkým pre trénovanie klasifikátora, preto stačilo, že sme tieto problémy vyriešili iba pre trénovaciu podmnožinu databázy RAVDESS. Postup, ktorý sme za týmto účelom navrhli sa spolieha na existenciu súborov, z ktorých môžu byť odvodené nové. Z tohto dôvodu nie je možné vyrovnať uniformitu tried v databázach EMO-DB a SAVEE – v databáze EMO-DB chýba celá jedna trieda (prekvapenie), a databáza SAVEE neobsahuje žiadne nahrávky žien. Tieto databázy do procesu trénovania nezasahovali, a popri testovacej podmnožine databázy RAVDESS sme ich použili na zvýšenie robustnosti testovania – keďže nahrávky v nich nemajú žiadnu väzbu na použité trénovacie údajov, pomohli nám získať lepšiu predstavu o úspešnosti klasifikátorov v reálnych podmienkach.

3.3.1 Vyrovnanie uniformity tried

Po rozdelení databázy RAVDESS nám v trénovacej podmnožine zostalo 76 súborov pre neutrálnu emóciu a 152 súborov pre všetky ostatné emócie. Na začiatku predspracovania bolo našim cieľom získať ďalších 76 súborov neutrálnej emócie, aby bol ich počet rovnaký ako počet súborov ostatných emócií.

V programe Adobe Audition sme pomocou funkcie Stretch and Pitch posunuli formantové frekvencie o dva poltóny nižšie. Využili sme hromadné spracovanie (Batch Process), a túto funkciu sme aplikovali na všetky súbory neutrálnej emócie. Vznikli nové súbory, ktoré sa od starých odlišovali mierne inou farbou a výškou hlasu. Došlo aj k nepatrnej zmene v tempe reči – reč sa mierne zrýchlila a teda nové súbory sú o niečo kratšie ako pôvodné (o niekoľko stoviek milisekúnd). Tento výsledok bol pre nás uspokojivý, pretože súbory sa zmenili dostatočne na to, aby ich bolo možné odlíšiť od pôvodných, ale zároveň neboli zmenené príliš radikálne.



Obrázok 10: Nastavenie úpravy formantových frekvencií v programe Adobe Audition.

3.3.2 Orezanie ticha a normalizácia

Pokračovali sme tým, že sme zo súborov orezali ticho, a normalizovali sme ich hlasitosť na jednotnú úroveň. Táto časť predspracovania bola automatizovaná, a je súčasťou skriptu *PreprocessData.py* (Príloha B), a je dôležitá pre všetky tri databázy. Používali sme funkcie z knižnice Librosa.

Veľa súborov v databázach, s ktorými pracujeme, obsahuje aj pasáže ticha. Keďže sme mali k dispozícii len krátke súbory a častý jav je, že súbor obsahuje napr. 2 sekundy reči a 2 sekundy ticha, tieto tiché pasáže môžu významne ovplyvniť hlasové charakteristiky, ktoré budú merané na celej dĺžke súboru. Avšak, museli sme brať do úvahy fakt, že prípadné dlhšie pauzy medzi slovami môžu byť tiež faktor, ktorý pomôže určiť emóciu, a nemôžu byť zo súboru jednoducho vystrihnuté. Preto sme orezávali iba ticho na začiatku a na konci súboru, teda pred začiatkom reči a po jej konci. Použili sme funkciu *trim()* z knižnice Librosa, ktorá vyžaduje vstupný parameter určujúci o koľko decibelov musí byť hlasitosť nižšia od referenčnej hodnoty aby bola orezaná. Funkcia *trim()* používa ako referenčnú úroveň maximálnu hlasitosť aktuálne spracovávaného súboru. Hodnotu parametra sme experimentálne určili na 28 decibelov.

Ďalším krokom bola normalizácia hlasitosti. Databázy boli nahrané v rozdielnych podmienkach a za použitia rôznej techniky. Niektoré vlastnosti, ako napr. vstupná úroveň signálu, môžu byť technikou výrazne ovplyvnené. Svoju úlohu zohráva aj vzdialenosť herca od mikrofónu – hlasitosť nahrávky s ubúdajúcou vzdialenosťou klesá. Mohlo by to

viesť k potenciálnym problémom s utlmením niektorých energických emócií (napr. hnev alebo radosť) ak je hovoriaci od mikrofónu ďaleko, alebo naopak, k veľkému nárastu intenzity pri emóciách, ktoré sa vyznačujú skôr tichým prejavom (napr. smútok), ak je hovoriaci k mikrofónu príliš blízko. Pokúsili sme sa vytvoriť riešenie, ktoré nebude závislé od použitia konkrétneho hardwaru a fixnej vzdialenosti od mikrofónu.

Rozhodli sme sa použiť podobný prístup ako v našej bakalárskej práci – súbory sme normalizovali pomocou funkcie *normalize()* tak, aby priemerná intenzita každého súboru bola 60 decibelov (pri referenčnej hodnote akustického tlaku 0,00002 Pa). Nevýhodou je, že prídeme o dynamický rozsah hlasitosti, ktorý by potenciálne mohol pomôcť odlíšiť jednotlivé triedy emócií, získame však istotu, že úroveň vstupného signálu bude rovnaká bez ohľadu na typ mikrofónu alebo vzdialenosť hovoriaceho. Predpokladali sme, že neurónová sieť bude mať stále k dispozícii dostatok iných hlasových charakteristík, na základe ktorých sa bude môcť rozhodovať, a preto sme sa intenzitu rozhodli v zmysle princípu „niečo za niečo“ obetovať.

3.3.3 Augmentácia údajov

Je známy fakt, že neurónové siete sa najlepšie učia, ak majú k dispozícii veľké množstvo trénovacích vzoriek. V opačnom prípade majú tendenciu skĺznuť k prílišnému zapamätaniu trénovacích údajov a k slabej generalizácii. Tento problém sa nazýva preučenie, resp. overfitting. Jednou z možností ako preučenie zredukovať je získať viac údajov. Keďže sme nemali možnosť získať úplne nové nahrávky, pokúsili sme sa upraviť súbory, ktoré máme k dispozícii. Úpravy súborov by mali byť zvolené tak, aby nielen rozšírili pôvodný dataset, ale aby odrážali rôzne situácie, s ktorými sa môže natrénovaný klasifikátor neskôr stretnúť. To môže byť napríklad šum v pozadí alebo znížená kvalita. Naopak, nevhodná úprava by bola súbor prehrávaný od konca k začiatku – je veľmi nepravdepodobné, že klasifikátor bude musieť klasifikovať človeka, ktorý rozpráva odzadu. Procesu úpravy súborov sa hovorí augmentácia. Vyrovnanie uniformity tried (pozri 3.3.1) bolo tiež špeciálnym prípadom augmentácie. Znovu sme pracovali iba s trénovacou podmnožinou databázy RAVDESS, a podarilo sa nám ju rozšíriť až na desaťnásobok pôvodnej veľkosti – teraz máme k dispozícii 10640 súborov. Prvé tri augmentácie z nasledujúceho zoznamu boli vytvorené ručne v programe Adobe Audition, ostatné sú výsledkom skriptu *PreprocessData.py* (Príloha B).

Zníženie kvality – zámerom bolo zmeniť súbory tak, aby zneli menej detailne a plechovo, podobne ako napríklad v starom telefóne. Využili sme efekt GuitarSuite a nastavili sme preset Lowest Fidelity.

Skreslenie – v týchto súboroch je prítomné chrčanie, ktoré by v realite mohlo byť spôsobené napríklad poškodeným káblom alebo nekvalitným telefonickým spojením. Použili sme efekt Distortion.

Ozvena – takto upravené súbory znejú ako keby boli nahraté vo veľkej prázdnej miestnosti. V Adobe Audition sme použili Convolution Reverb.

Úprava výšky hlasu – výšku hlasu sme v súbore náhodne zmenili funkciou *pitch_shift()* z knižnice Librosa. Najprv sa náhodne určilo či sa bude výška posúvať smerom dole alebo hore, a potom sa výška posunula náhodne o 1 až 3 poltóny.

Výpadok – táto augmentácia simulovala krátke výpadky niektorých vzoriek zo súboru. Vo výslednom súbore potom boli prítomné seky, typické napríklad pre situáciu keď počas telefonátu cez internet nastane strata packetov. Najprv sme náhodne určili počet výpadkov, ich počet mohol byť 4 až 6. Dĺžka jedného výpadku bola určená na 5% z celkovej dĺžky súboru.

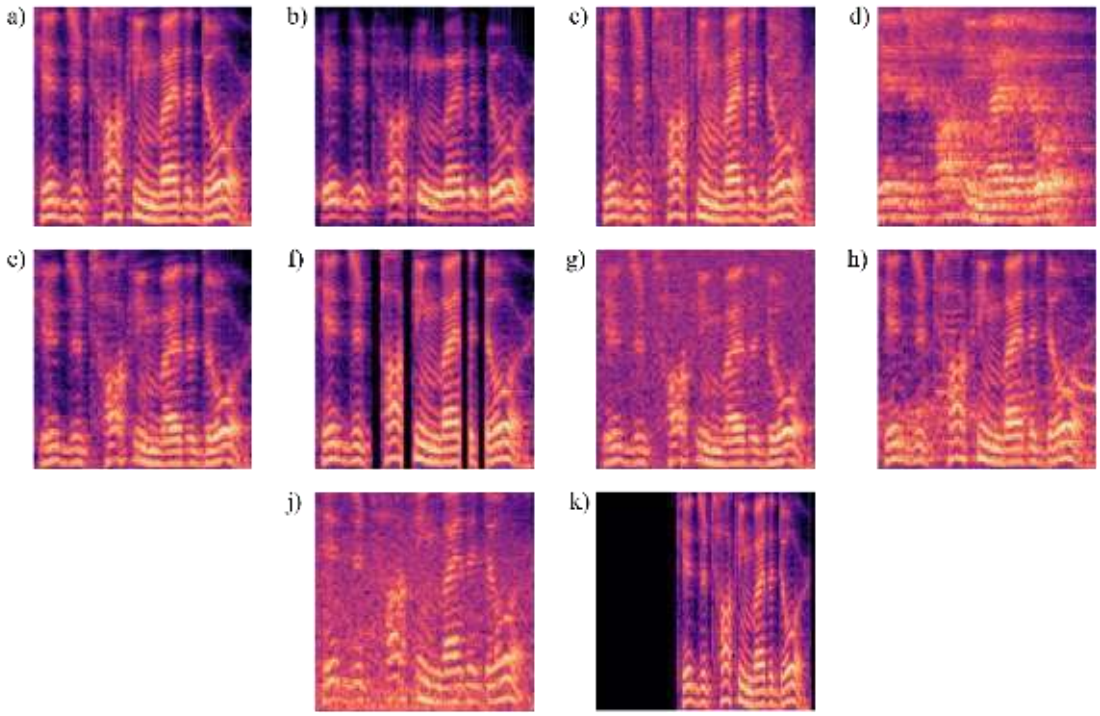
Šum v pozadí – súbory sme zmixovali s bielym šumom. Mixovanie súborov v skripte je úplne jednoduché, stačí načítať dva súbory do NumPy polí, a potom polia sčítať, avšak ich veľkosť musí byť rovnaká. Mali sme k dispozícii jeden dlhý súbor, v ktorom nebolo nič okrem šumu. Z tohto súboru sme vybrali časť tak dlhú, aby sa presne zhodovala s dĺžkou súboru, ktorý sme upravovali, a vzorky sme potom zmiešali. Pred samotným mixovaním sme ešte náhodne rozhodli, či sa zároveň upraví aj výška hlasu.

Konverzácia v pozadí – použili sme rovnaký postup, ako v predchádzajúcom prípade, ale tentokrát sme do pozadia primiešali zvuky rušného prostredia, v ktorom prebieha niekoľko konverzácií naraz. Takéto súbory môžu napríklad reprezentovať bežný ruch na spoločenskej udalosti alebo v reštaurácii.

Ruch ulice v pozadí – aj tentokrát sme nezmeneným postupom zmiešali pôvodný súbor s iným, v tomto prípade boli primiešané zvuky ulice a prechádzajúcich áut.

Pridanie ticha – pri orezaní ticha v predchádzajúcej kapitole sme sa rozhodli orezávať ticho iba na konci a na začiatku súboru. Avšak, pri spracovaní v reálnom čase môže dôjsť k situácii, že na začiatku alebo na konci sa objaví zanedbateľné množstvo

vzoriek (napríklad náhodný zvuk z prostredia), ktoré spôsobia, že súbor už nebude začínať (resp. končiť) tichom, a teda ticho nebude orezané. Preto na začiatok a na koniec náhodne pridáme ticho, tak aby bol súčet pasáží ticha zo začiatku a z konca súboru dve sekundy. Alternatívne k tejto augmentácii sme si mohli ponechať pôvodný neorezaný súbor, avšak dĺžka pasáží ticha v týchto súboroch bola približne rovnaká na začiatku aj na konci. Rozhodli sme sa ticho vygenerovať znovu s pridaným prvkom náhodnej dĺžky – môže sa napríklad pridať 1,8 sekundy ticha na začiatok a 0,2 na koniec.



Obrázok 11: Porovnanie spektrogramov augmentácií jedného súboru.
***a)** pôvodný súbor, **b)** znížená kvalita, **c)** skreslenie, **d)** ozvena, **e)** zmenená výška hlasu, **f)** výpadok, **g)** biely šum, **h)** konverzácie v pozadí, **j)** ruch ulice, **k)** náhodne pridané ticho.*

3.4 EXPERIMENTY S NEURÓNOVÝMI SIEŤAMI

Predspracovaním sme súbory pripravili na extrakciu rôznych hlasových charakteristík, ktoré budú použité ako vstup pre neurónové siete. Našim cieľom v tejto časti bolo nájsť najvhodnejšie charakteristiky a najvhodnejšiu architektúru modelu. Keďže požiadavka na túto diplomovú prácu je spracovanie v reálnom čase, museli sme si pri experimentoch, okrem úspešnosti modelu, všímať aj jednoduchosť extrakcie charakteristík, náročnosť ich ďalšieho spracovania a rýchlosť predikcie neznámych dát.

Na tvorbu a tréning modelu sme použili knižnicu keras s tensorflow backendom. Ďalej sme použili technológiu Nvidia CUDA, pomocou ktorej sme presunuli záťaž trénovania a predikcie na grafickú kartu, ktorou bola Nvidia GeForce GTX 1050.

Vytvorené modely sme otestovali s testovacou podmnožinou databázy RAVDESS, a s databázami EMO-DB a SAVEE. Výsledky predikcie sme v každom experimente spracovali do formy tabuliek, kde sme úspešnosť klasifikácie vyjadrili v percentách. Diagonála tabuliek zobrazuje percento testovacích prípadov, ktoré boli určené správne. Riadky tabuliek reprezentujú emócie, ktoré sme očakávali, a stĺpce emócie, ktoré určil model. Na doplnenie sme vytvorili aj tabuľky kde sme úspešnosť vyjadrili absolútnym počtom správne a nesprávne určených emócií. Všetky tabuľky sú v prílohe (Príloha F).

3.4.1 Experiment 1 – dopredná neurónová sieť

Príprava:

Prvý experiment je variáciou experimentu z našej bakalárskej práce s miernymi obmenami. V bakalárskej práci sme pomocou programu PRAAT zmerali rôzne charakteristiky – základnú frekvenciu, prvé 3 formantové frekvencie, intenzitu, jitter a shimmer. Z týchto charakteristík sme potom vypočítali globálne charakteristiky. Teraz sme tieto charakteristiky rozšírili na 5 formantových frekvencií, a pridali sme ďalšiu charakteristiku – harmonicitu. Na meranie sme použili knižnicu parselmouth.

Správne určenie formantových frekvencií sa ukázalo ako problematické, pretože najprv je potrebné určiť maximálnu frekvenciu – hornú hranicu frekvenčného rozsahu, v ktorom budú formantové frekvencie vyhľadávané. Problémom je, že vhodná hodnota maximálnej frekvencie je spravidla 5000 Hz pre mužov, 5500 Hz pre ženy a 8000 Hz pre deti[15]. Ak by sme sa zamerali iba na dospelých ľudí, potom by pre dosiahnutie čo najväčšej presnosti bolo najprv nutné správne určiť pohlavie. V tejto fáze sme už museli premýšľať aj nad možnými budúcimi dôsledkami a podobou finálneho programu. Pri trénovaní by sme mohli súbory veľmi jednoducho roztriediť na mužov a ženy, avšak pri klasifikácii v reálnom čase nie je možné určiť pohlavie dopredu. Potenciálnym riešením by mohlo byť natrénovanie ďalšieho modelu, ktorý najprv určí pohlavie, a ďalší postup by sa prispôsobil tejto predikcii. To by znamenalo nutnosť rozšírenia výskumu

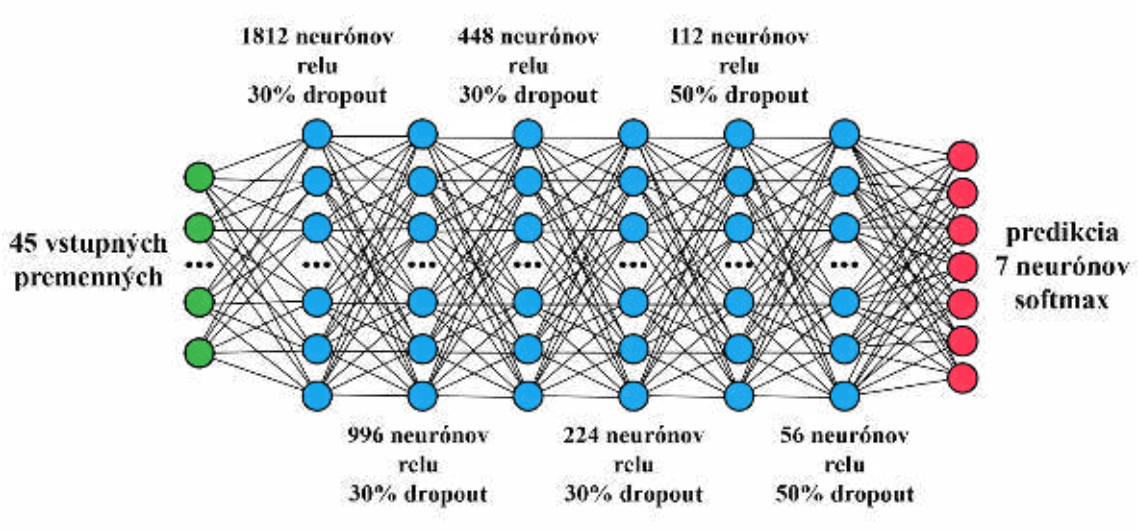
[15] fon.hum.uva.nl/praat/manual/Sound__To_Formant__burg____.html

o ďalšiu úlohu, a takisto aj zvýšenie počtu úkonov vedúcich k predikcii emócie, čo by mohlo v reálnom čase spôsobiť problém. Druhým možným riešením by bolo manuálne nastavenie pohlavia v programe, čím by sa práca s ním stala menej pohodlnou, a jeho možnosti by sa značne limitovali – napr. pri vystriedaní hovoriaceho pri mikrofóne by bolo nutné manuálne prepnúť pohlavie. Ani jedna z týchto možností pre nás nebola prijateľná. Namiesto toho sme zvolili jednoduchší postup – maximálnu frekvenciu sme nastavili na 5500 Hz, čo poskytne dostatočný frekvenčný rozsah pre ženy aj mužov, avšak formanty sa v prípade mužov môžu mierne líšiť od ich skutočnej hodnoty.

Po extrakcii charakteristík sme z nich určili maximá, minimá, smerodajné odchýlky, mediány a ďalšie hodnoty – spolu až 64 globálnych charakteristík pre každý súbor. Každú premennú sme preškálovali na hodnoty medzi 0 a 1. Škálovali sme funkciou *MinMaxScaler()* z knižnice scikit-learn. Počet premenných sme sa pokúsili ešte zredukovať identifikáciou zbytočných premenných. Vylúčili sme charakteristiky s mierou vzájomnej korelácie nad 90%. Vysoká miera korelácie medzi dvomi charakteristikami indikuje, že tieto charakteristiky nesú rovnakú informáciu, a teda nám stačí jedna z nich. Mieru korelácie sme určili funkciou *corr()* z knižnice pandas, ktorá ako metódu štandardne používa Pearsonov korelačný koeficient. Na záver prípravy nám zostalo 45 premenných. Úplný zoznam premenných je možné nájsť v súbore *Ex1_vars.pdf* (Príloha C). Príprava bola zautomatizovaná skriptom *Ex1_prep.py* (Príloha C).

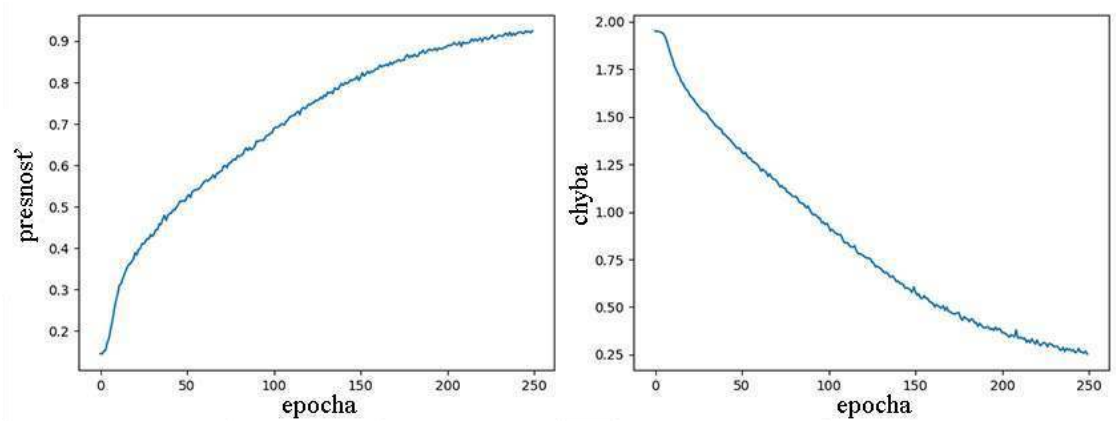
Model:

V tomto experimente sme vytvorili model doprednej neurónovej siete. Architektúra siete a aktivačné funkcie sú zobrazené na obrázku nižšie (Obrázok 12). Preučenie sme sa pokúsili zredukovať dropout technikou – časť neurónov náhodne vypadne, a sieť je nútená k lepšej generalízácii. Na vytvorenie a trénovanie modelu sme použili skript *Ex1_NN.py* (príloha C).



Obrázok 12: Architektúra doprednej neurónovej siete.

Tréning siete prebiehal počas 250 epoch. Použili sme optimizér Adam s rýchlosťou učenia 0.00005. Ako chybovú funkciu sme použili kategorickú krížovú entropiu. Na obrázku nižšie (Obrázok 13) je znázornený priebeh trénovania a znižovanie chyby.



Obrázok 13: Experiment 1 - priebeh trénovania a znižovanie chyby.

Výsledky:

Model v teste s databázou RAVDESS dosiahol celkovú úspešnosť 61,54%. Všetky emócie boli správne určené vo viac ako 50% prípadov, pričom najpresnejšie boli určené emócie hnev a prekvapenie – až 70%. Najmenej presná bola predikcia strachu – 52,50%, pričom strach bol v 15% prípadov nesprávne určený ako smútok. Najväčšiu mieru zámeny za inú individuálnu kategóriu sme pozorovali pri neutrálnej emócií, ktorá bola až v 35% prípadoch nesprávne klasifikovaná ako smútok.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	70,00	10,00	5,00	7,50	0,00	2,50	5,00
	znechutenie	17,50	62,50	2,50	0,00	2,50	2,50	12,50
	strach	7,50	7,50	52,50	7,50	2,50	15,00	7,50
	šťastie	2,50	7,50	2,50	60,00	12,50	10,00	5,00
	neutrál	0,00	0,00	0,00	0,00	60,00	35,00	5,00
	smútok	0,00	12,50	12,50	5,00	12,50	55,00	2,50
	prekvapenie	2,50	12,50	5,00	2,50	2,50	5,00	70,00
	ÚSPEŠNOSŤ	**61,54**						

Tabuľka 1: Experiment 1 - test s databázou RAVDESS.

Test s databázou nemeckých nahrávok dopadol menej uspokojivo ako predchádzajúci test. Model predikoval správnu emóciu v 40,31% nahrávok. Najviac úspešných klasifikácií bolo dosiahnutých v prípade emócie smútok – 79,03%, avšak iné emócie boli so zvýšenou mierou zamieňané práve za túto emóciu. Najmenšiu presnosť sme pozorovali pri znechutení – 26,09%. Nahrávky prekvapenia sa v tejto databáze nenachádzajú, preto je posledný riadok tabuľky vyplnený nulovými hodnotami.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	30,71	15,75	11,02	30,71	0,79	7,09	3,94
	znechutenie	0,00	26,09	13,04	13,04	0,00	39,13	8,70
	strach	1,45	5,80	40,58	5,80	5,80	23,19	17,39
	šťastie	2,82	12,68	5,63	42,25	1,41	26,76	8,45
	neutrál	2,53	24,05	1,27	5,06	31,65	35,44	0,00
	smútok	0,00	3,23	4,84	0,00	9,68	79,03	3,23
	prekvapenie	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	ÚSPEŠNOSŤ	**40,31**						

Tabuľka 2: Experiment 1 - test s databázou EMO-DB.

Najhoršie dopadol test s databázou SAVEE – celková úspešnosť bola iba 24,38%. Väčšina predikcií sa rozložila do 3 kategórií – neutrálna emócia, smútok a znechutenie. Zvyšné 4 emócie boli modelom vnímané iba minimálne, pričom strach nebol správne určený ani raz, a zároveň ani nebola žiadna iná emócia nesprávne klasifikovaná ako táto emócia.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	1,67	20,00	0,00	15,00	36,67	25,00	1,67
	znechutenie	1,67	21,67	0,00	6,67	38,33	30,00	1,67
	strach	1,67	23,33	0,00	8,33	16,67	48,33	1,67
	šťastie	6,67	16,67	0,00	10,00	23,33	40,00	3,33
	neutrál	2,50	7,50	0,00	12,50	53,33	24,17	0,00
	smútok	1,67	21,67	0,00	15,00	15,00	46,67	0,00
	prekvapenie	6,67	21,67	0,00	13,33	8,33	41,67	8,33
	ÚSPEŠNOSŤ	**24,38**						

Tabuľka 3: Experiment 1 - test s databázou SAVEE.

3.4.2 Experiment 2 – 1D konvolučná sieť

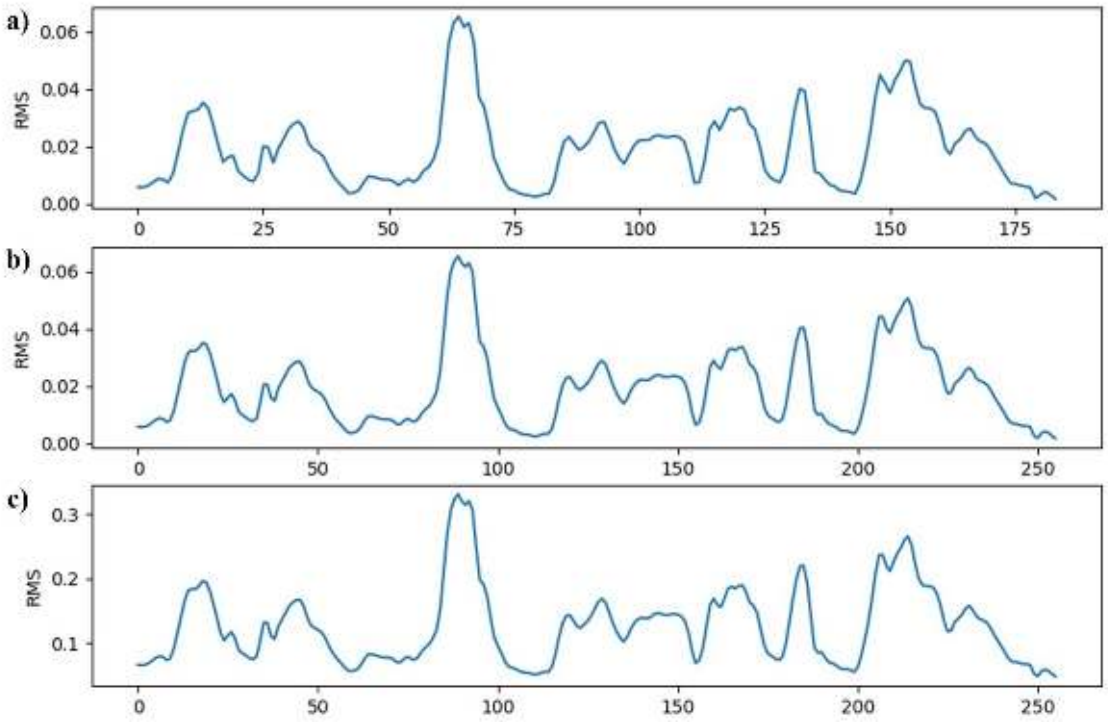
Príprava:

V tomto experimente sme zmerali 13 MFCC charakteristík, a pridali sme k nim efektívnu hodnotu energie a zerocrossing. Tentokrát sme z charakteristík nepočítali globálne charakteristiky, ale použili sme ich lokálne hodnoty merané v čase. Na extrakciu sme použili funkcie *mfcc()*, *rms()* a *zero_crossing_rate()* z knižnice Librosa. Nastavili sme vstupné parametre týchto funkcií. Dĺžku okna rýchlej Fourierovej transformácie sme nastavili na 800 vzoriek, a počet vzoriek medzi za sebou idúcimi rámcami na 400. Extrakciu charakteristík a ich následné spracovanie sme zautomatizovali v skripte *Ex2_prep.py* (Príloha D).

Problémom tejto metódy je, že 15 vybraných charakteristík je meraných v čase. Výsledkom merania je matica 15xN, kde N závisí od dĺžky vstupného súboru (pri použitých vstupných parametroch je to približne 100 vzoriek na 1 sekundu). Knižnica keras premenlivú veľkosť vstupu do konvolučnej neurónovej siete podporuje, avšak prvá predikcia po zmene veľkosti môže trvať až niekoľko sekúnd – to by nebolo vhodné pri určovaní v reálnom čase. Graf funkcie, ktorá je tvorená lokálnymi hodnotami charakteristiky v čase, je možné podľa potreby prevzorkovať na nami určenú fixnú dĺžku tak, aby bol zároveň zachovaný jej tvar. Tento cieľ je dosiahnuteľný tým, že na grafe spojitej funkcie sa vyberie taký počet bodov, ktorý zodpovedá požadovanej dĺžke, a potom sa pomocou interpolácie vytvorí nový graf, ktorý sa vo vybraných bodoch zhoduje s pôvodným grafom. Na to sme použili triedu *interp1d()*, ktorá je súčasťou knižnice scipy. Rozhodli sme sa použiť kubickú interpoláciu, pretože po pokusoch sa

ukázalo, že zachováva tvar pôvodného grafu vernejšie ako lineárna interpolácia. Grafy charakteristík sme preškálovali na 256 vzoriek.

Na záver sme každú charakteristiku preškálovali na hodnoty medzi 0 a 1. Opäť sme použili min-max škálovanie. Z celej trénovacej podmnožiny databázy RAVDESS sme zistili absolútne maximá a minimá pre každú z 15 charakteristík. Potom sme na ich základe preškálovali trénovaciu aj testovaciu podmnožinu, a rovnako zvyšné dve databázy. Zároveň sme tieto maximá a minimá uložili do súboru, pretože vo finálnom programe musí škálovanie prebiehať rovnakým spôsobom, a na základe rovnakých hodnôt, ako na dátach, s ktorými bol model natrénovaný.

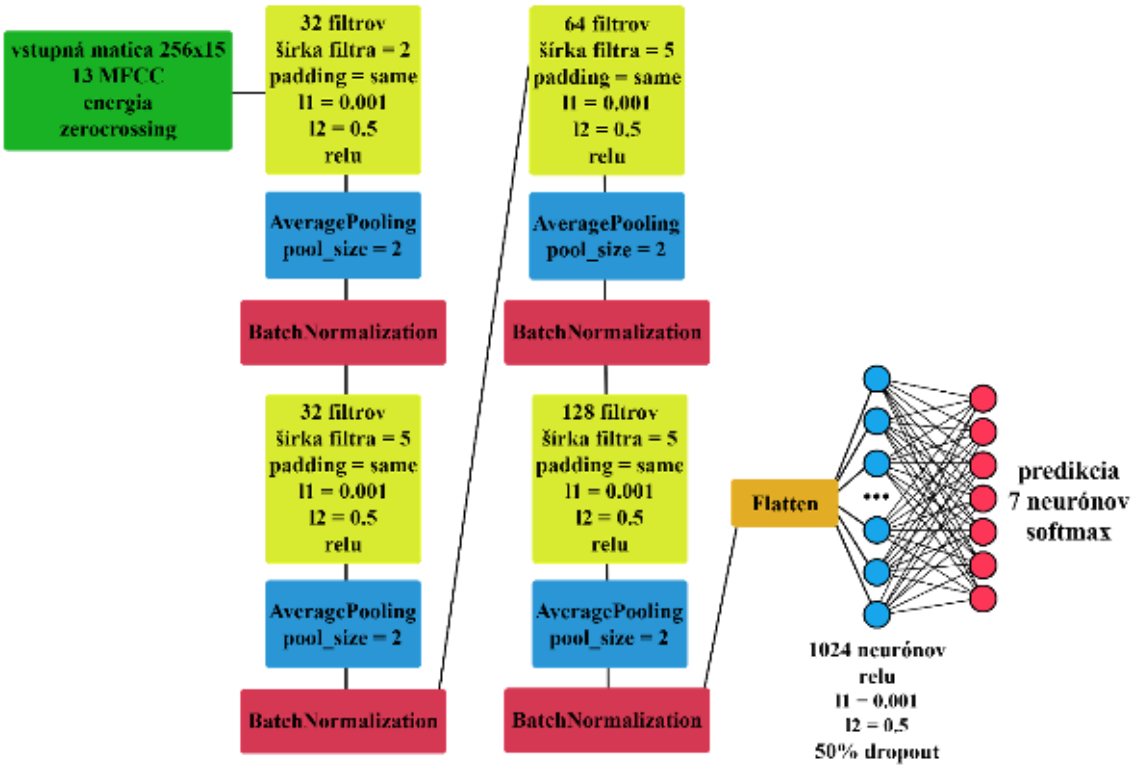


Obrázok 14: Ukážka prevzorkovania a škálovania – tvar funkcie nie je zmenený. a) graf energie pred úpravou, b) graf energie po prevzorkovaní na 256 vzoriek, c) graf energie po min-max škálovaní.

Úplne posledným krokom bolo transponovanie matice charakteristík z tvaru 15x256 na tvar 256x15. Transponovanie bolo potrebné z jednoduchého dôvodu – neurónová sieť vytvorená s použitím knižnice keras pre správne fungovanie vyžaduje maticu, kde stĺpce reprezentujú charakteristiky, a riadky hodnoty týchto charakteristík v čase, zatiaľ čo knižnica librosa poskytuje výstup v opačnej forme.

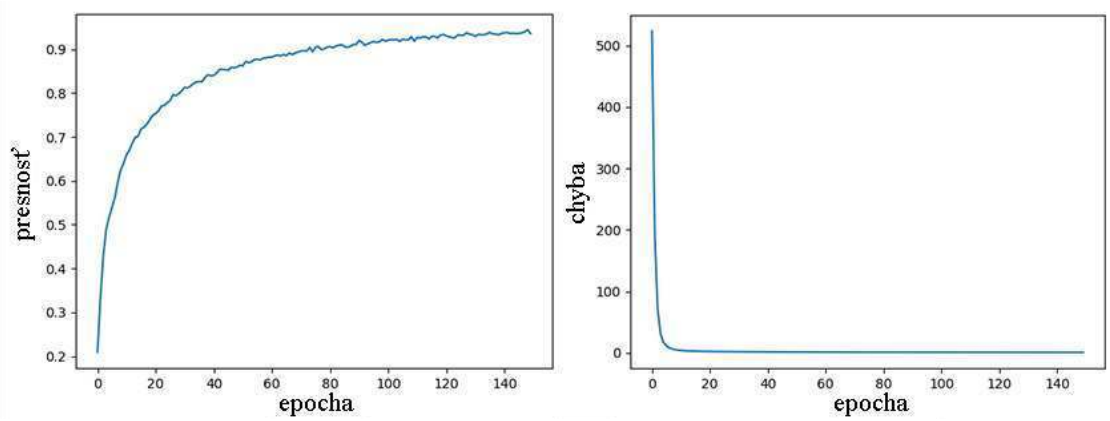
Model:

Vstupom do neurónovej siete sú lokálne hodnoty 15 charakteristík. Na takýto typ vstupu sú vhodné jednorozmerné konvolučné siete. Architektúra siete je na obrázku nižšie (Obrázok 15). Použili sme 4 konvolučné vrsty, za každou z nich nasledovala pooling vrstva a normalizácia. Sieť sme zakončili pridaním jednej plne prepojenej vrstvy. Preučenie sme redukovali l1 a l2 regularizáciou (penalizácia váh), a dropoutmi. Zdrojový kód tohto modelu sa nachádza v skripte *Ex2_NN.py* (Príloha D).



Obrázok 15: Architektúra 1D konvolučnej siete.

Sieť sme trénovali s optimizérom Adam s rýchlosťou učenia 0.00005 počas 150 epoch. Chybovou funkciou bola opäť kategorická krížová entropia. Presnosť a chyba počas jednotlivých epoch je zobrazená na nasledujúcich obrázku nižšie (Obrázok 16).



Obrázok 16: Experiment 2 - priebeh trénovania a klesajúca chyba.

Výsledky:

V testovaní s databázou RAVDESS dokázal tento model dosiahnuť až 78,85% úspešnosť. Úspešnosť určenia žiadnej kategórie neklesla pod 60%, 4 kategórie zo 7 boli správne určené vo viac ako 80% prípadov. Najvýraznejšia chyba nastala pri emócii prekvapenie, ktorá bola v 20% prípadoch určená ako šťastie.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	90,00	2,50	0,00	7,50	0,00	0,00	0,00
	znechutenie	5,00	95,00	0,00	0,00	0,00	0,00	0,00
	strach	2,50	5,00	67,50	7,50	2,50	7,50	7,50
	šťastie	0,00	0,00	0,00	85,00	5,00	2,50	7,50
	neutrál	0,00	0,00	0,00	5,00	80,00	10,00	5,00
	smútok	0,00	5,00	15,00	7,50	5,00	62,50	5,00
	prekvapenie	5,00	0,00	2,50	20,00	0,00	0,00	72,50
	ÚSPEŠNOSŤ	**78,85**						

Tabuľka 4: Experiment 2 - test s databázou RAVDESS.

Model v teste s databázou EMO-DB nedosiahol úspešnosť predchádzajúceho testu. Iba dve emócie boli správne určené vo vyššej miere ako 50% - hnev a smútok. Model dosiahol celkovú úspešnosť 39,65%.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	54,33	12,60	6,30	22,05	0,79	3,94	0,00
	znechutenie	32,61	15,22	10,87	4,35	4,35	19,57	13,04
	strach	15,94	8,70	26,09	15,94	5,80	14,49	13,04
	šťastie	45,07	7,04	7,04	28,17	2,82	7,04	2,82
	neutrál	7,59	0,00	2,53	36,71	29,11	12,66	11,39
	smútok	0,00	3,23	16,13	4,84	4,84	69,35	1,61
	prekvapenie	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	ÚSPEŠNOSŤ	**39,65**						

Tabuľka 5: Experiment 2 - test s databázou EMO-DB.

Test s databázou SAVEE dopadol opäť najhoršie. Väčšina nahrávok bola určená ako smútok alebo neutrálna emócia. Predikcia ostatných tried bola zanedbateľná. Aj emócie, ktoré sú obvykle výrazné a energické (napr. hnev) boli vo väčšine prípadov určené ako smútok.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	8,33	3,33	0,00	8,33	13,33	65,00	1,67
	znechutenie	0,00	3,33	0,00	1,67	20,00	75,00	0,00
	strach	8,33	0,00	0,00	10,00	25,00	55,00	1,67
	šťastie	3,33	8,33	0,00	6,67	18,33	61,67	1,67
	neutrál	0,00	4,17	0,00	0,83	29,17	65,83	0,00
	smútok	0,00	5,00	0,00	0,00	13,33	81,67	0,00
	prekvapenie	3,33	1,67	0,00	3,33	26,67	65,00	0,00
	ÚSPEŠNOSŤ	**19,79**						

Tabuľka 6: Experiment 2 - test s databázou SAVEE.

3.4.3 Experiment 3 – 2D konvolučná sieť

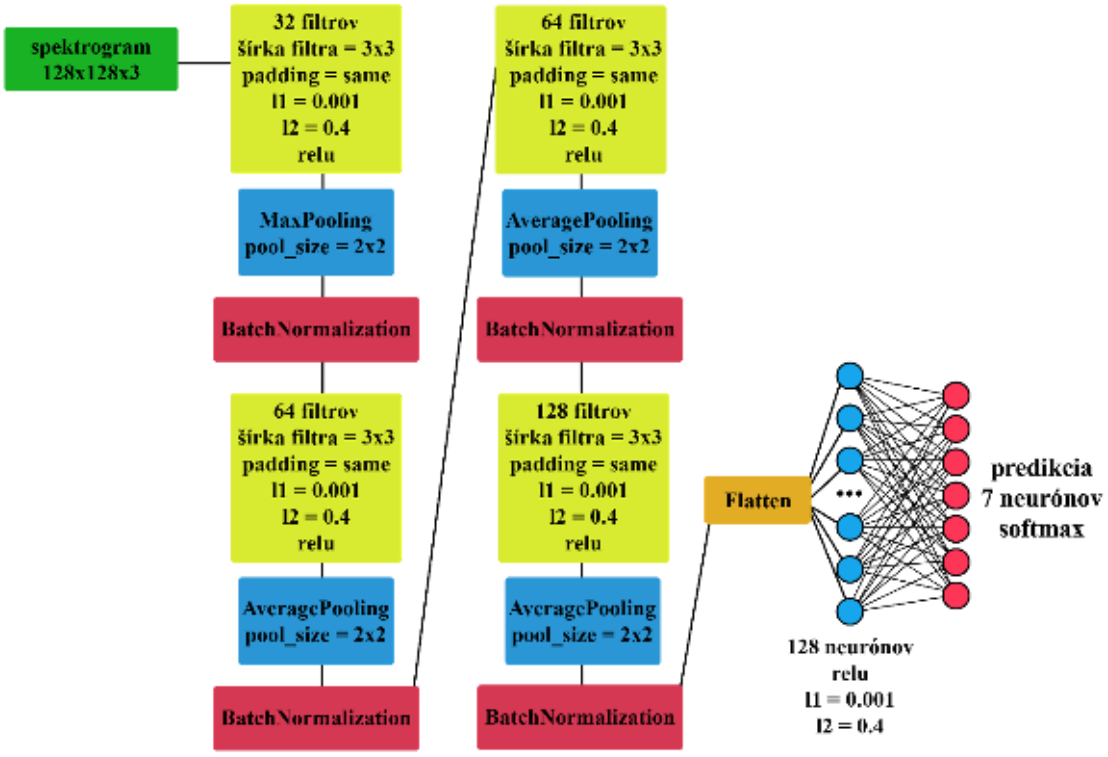
Príprava:

V poslednom experimente sme ako vstup do neurónovej siete použili grafickú reprezentáciu zvukových súborov – spektrogramy. Spektrogramy sme vypočítali funkciou *melspectrogram()* z knižnice Librosa. Tieto spektrogramy na y osi zobrazujú frekvenciu prepočítanú z Hz na mel jednotky. Frekvenčný rozsah pre výpočet spektrogramov sme nastavili od 50 do 4000 Hz. Parametre rýchlej Fourierovej transformácie boli rovnaké ako v experimente 2. Potom sme ich uložili vo formáte .png.

V prípade spektrogramov nastal rovnaký problém ako v predchádzajúcom experimente. Rozmery spektrogramu sú závislé od dĺžky zvukového súboru. Preto sme ich rozmery ešte upravili na 128x128 pixelov. Na vytváranie spektrogramov sme napísali skript *Ex3_prep.py* (Príloha E).

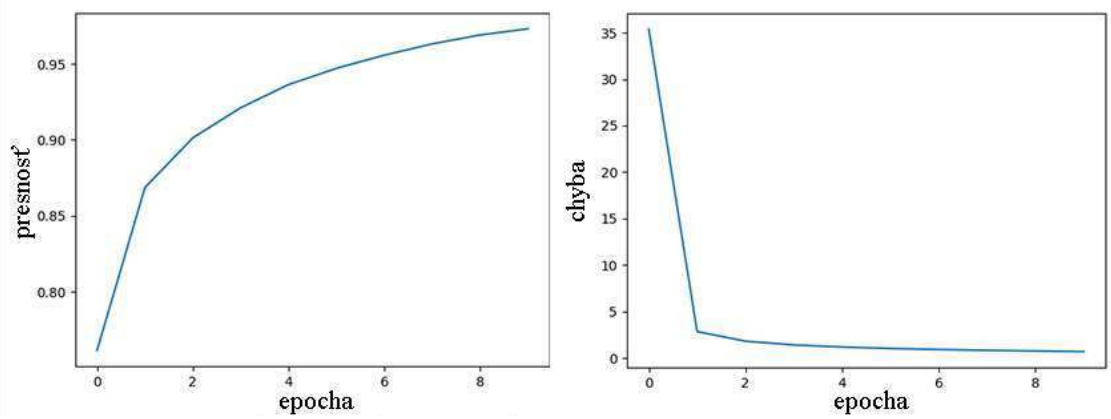
Model:

Vstupom do neurónovej siete je matica s rozmermi 128x128x3. Tretím rozmerom matice sú farebné kanály (RGB). Ukázalo sa, že táto metóda je pamäťovo náročná, a nie je možné všetky trénovacie súbory načítať naraz. Využili sme triedu *ImageDataGenerator* z knižnice keras, a obrázky sme načítavali postupne po 32. Hodnoty pixelov sme normalizovali medzi 0 a 1 nastavením rescale parametra generátora na 1./255. Na zredukovanie preučenia sme použili l1 a l2 regularizáciu. Po každej konvolučnej vrstve nasledovala pooling vrstva a normalizácia. Model sme vytvorili v skripte *Ex3_NN.py* (Príloha E).



Obrázok 17: Architektúra 2D konvolučnej siete.

Model bol trénovaný počas 10 epoch s optimizérom Adam a rýchlosťou učenia 0,00001. Ako chybovú funkciu sme použili sme kategorickú krížovú. Na obrázku (Obrázok 18) je presnosť a chyba počas epoch.



Obrázok 18: Experiment 3 – presnosť trénovania a klesajúca chyba.

Výsledky:

V teste s databázou RAVDESS sme dosiahli úspešnosť 76,54%. Každá z emócií bola správne určená vo viac ako 60% prípadov, ale úspešnosť určenia pri žiadnej z nich nepresiahla 90%.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	87,50	5,00	0,00	2,50	2,50	0,00	2,50
	znechutenie	5,00	87,50	0,00	2,50	0,00	2,50	2,50
	strach	0,00	5,00	75,00	5,00	2,50	10,00	2,50
	šťastie	5,00	0,00	10,00	65,00	7,50	0,00	12,50
	neutrál	0,00	5,00	5,00	15,00	65,00	10,00	0,00
	smútok	0,00	15,00	12,50	5,00	2,50	62,50	2,50
	prekvapenie	2,50	0,00	5,00	2,50	0,00	2,50	87,50
	ÚSPEŠNOSŤ	**76,54**						

Tabuľka 7: Experiment 3 - test s databázou RAVDESS.

V teste s databázou EMO-DB model správne určil emóciu v 29,52% nahrávok. Tento test sa vyznačoval veľkou mierou nesprávne určených emócií. Zvýšenú mieru úspešnosti sme pozorovali iba pri strachu a neutrálnej emócií.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	8,66	1,57	10,24	22,05	1,57	0,00	55,91
	znechutenie	0,00	2,17	28,26	6,52	8,70	6,52	47,83
	strach	0,00	0,00	50,72	0,00	17,39	0,00	31,88
	šťastie	1,41	0,00	12,68	15,49	14,08	0,00	56,34
	neutrál	0,00	0,00	3,80	5,06	77,22	7,59	6,33
	smútok	0,00	9,68	41,94	0,00	17,74	24,19	6,45
	prekvapenie	0,00	0,00	0,00	0,00	0,00	0,00	0,00
	ÚSPEŠNOSŤ	**29,52**						

Tabuľka 8: Experiment 3 - test s databázou EMO-DB.

Model v teste s databázou SAVEE správne klasifikoval 21,46% prípadov. Väčšina bola určená ako smútok, znechutenie alebo prekvapenie. Hnev a strach nebol rozoznaný vôbec.

		Určená emócia						
	(%)	hnev	znechutenie	strach	šťastie	neutrál	smútok	prekvapenie
Skutočná emócia	hnev	1,67	31,67	0,00	16,67	6,67	26,67	16,67
	znechutenie	0,00	33,33	1,67	13,33	6,67	31,67	13,33
	strach	0,00	26,67	0,00	11,67	6,67	45,00	10,00
	šťastie	0,00	16,67	0,00	18,33	3,33	30,00	31,67
	neutrál	0,00	20,83	0,00	2,50	17,50	51,67	7,50
	smútok	0,00	16,67	0,00	11,67	8,33	58,33	5,00
	prekvapenie	1,67	31,67	1,67	3,33	3,33	33,33	25,00
	ÚSPEŠNOSŤ	**21,46**						

Tabuľka 9: Experiment 3 - test s databázou SAVEE.

3.4.4 Zhodnotenie experimentov a výber modelu

Vyskúšali sme 3 metódy klasifikácie emócií. Ich spoločným znakom je, že úspešnosť predikcie bola vždy najvyššia v testoch s testovacou podmnožinou databázy RAVDESS, pričom najlepší výsledok bol dosiahnutý v druhom experimente – 78,85%. Testy s databázou EMO-DB boli úspešné na 40,31%, 39,65% a 29,52%. Vo všetkých troch experimentoch boli najhoršie výsledky dosiahnuté v testoch s databázou SAVEE. Z týchto výsledkov sme vyvodili záver, že charakteristiky, ktoré sme skúmali, v sebe nesú dostatočnú informáciu o emocionálnom kontexte nahrávky, avšak nami navrhnuté modely z nich nedokázali odvodiť dostatočne všeobecné pravidlá, ktoré by boli platné

pre všetky databázy. Pravdepodobnou príčinou je, že na trénovanie sme použili iba podmnožinu databázy RAVDESS. V tejto databáze sa opakujú nahrávky iba 2 viet (pozri 3.2), ktoré sú navyše veľmi podobné. Pre ďalší výskum z toho vyplynula potreba získania väčšieho množstva rôznorodých nahrávok.

Pri výbere najlepšieho modelu a metodiky sme navyše museli brať ohľad aj na čas, za aký je možné extrahovať a spracovať charakteristiky zo zadaného vstupu, a ako dlho trvá následná predikcia emócie. Pre každú vstupnú nahrávku sme sa pokúsili pomocou Python knižnice time zmerať približné časy spracovania a predikcie. Zo získaných časov sme potom vyrátali priemerný čas spracovania a predikcie jednej nahrávky. Výsledky merania sú v tabuľke nižšie (Tabuľka 10).

(s)	spracovanie	predikcia
Experiment 1	0,47	0,00046
Experiment 2	0,04	0,00096
Experiment 3	0,33	0,02073

Tabuľka 10: Priemerná rýchlosť spracovania a predikcie nahrávky.

Ako najvhodnejšia sa ukázala metóda z experimentu 2, pretože je najrýchlejšia, spracovanie vstupu prebieha jednoduchým a bezproblémovým spôsobom, a dosiahla najvyššiu celkovú úspešnosť. V neprospech experimentu 1 hovorí pomalé spracovanie knižnicou parselmouth, najnižšia úspešnosť, a rovnako aj presnosť určenia formantových frekvencií ovplyvnená pohlavím. Ani experiment 3 neprekonal úspešnosť experimentu 2, navyše spracovanie bolo niekoľkonásobne časovo náročnejšie. Preto sme sa v našom vlastnom programe rozhodli použiť model vytvorený v experimente 2.

3.5 PROGRAM EMOREC 2

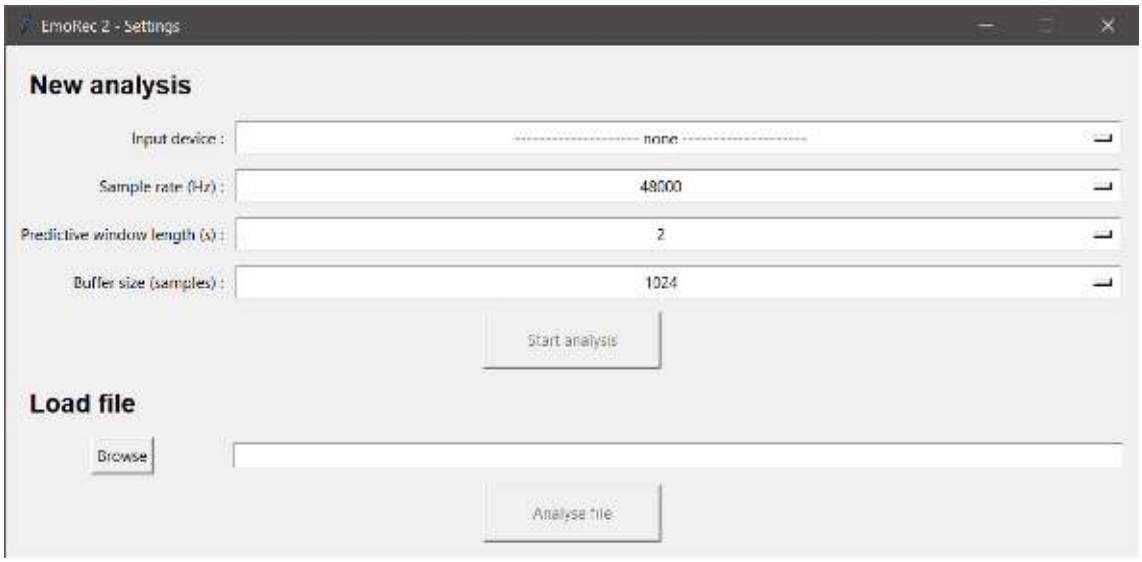
V jazyku Python vo verzii 3.6 sme naprogramovali vlastnú aplikáciu s grafickým rozhraním, ktorej úlohou je v reálnom čase vyhodnocovať prúd údajov zo vstupného zariadenia, napr. z mikrofónu. Na vytvorenie grafického rozhrania sme použili balík tkinter vo verzii 8.6, ktorý je súčasťou štandardnej knižnice jazyka Python. Zdrojový kód tohto programu je k dispozícii v prílohe (Príloha A). Pre správne fungovanie je potrebné najprv nainštalovať tieto knižnice (v zátvorke je uvedená verzia, ktorú sme používali počas tvorby programu):

- matplotlib (3.1.1),
- sounddevice (0.3.13),
- soundfile (0.10.2),
- numpy (1.17.3),
- librosa (0.7.0),
- tensorflow-gpu (1.14.0) alebo tensorflow (1.14.0),
- keras (2.2.5),
- scipy (1.3.1).

S knižnicou tensorflow-gpu je zároveň nutné používať technológiu Nvidia CUDA (používali sme verziu 10.0). Použitie knižnice tensorflow-gpu je odporúčané, pretože umožňuje znížiť zaťaženie procesora počas predikcie.

3.5.1 Obsluha programu EmoRec 2

Prvé okno, ktoré sa zobrazí po spustení programu, je okno s nastaveniami analýzy. Po výbere vstupného zariadenia je možné začať novú analýzu v reálnom čase. Alternatívne je možné načítať a analyzovať hotový súbor vo formáte .wav.

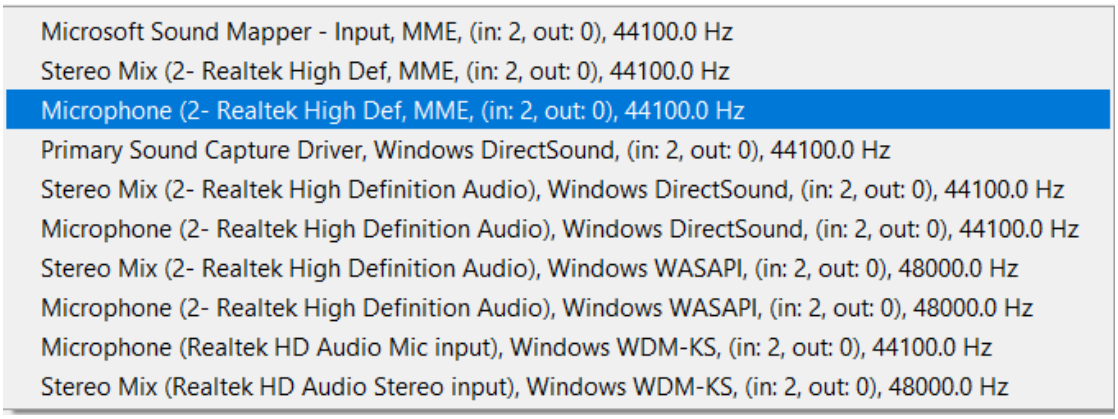


Obrázok 19: Okno s nastaveniami.

V prípade analýzy v reálnom čase je možné nastaviť niekoľko parametrov pre prúd vstupných údajov. Ich hodnoty ovplyvňujú záťaž procesora. Nevhodné nastavenie

môže spôsobiť, že dôjde k pretečeniu zásobníka vstupu, a časť údajov zo vstupného prúdu bude stratená.

Input device – po kliknutí sa zobrazí zoznam audio vstupných zariadení, ktoré sú pripojené k počítaču. Zoznam obsahuje aj informácie o ovládači, ktorý bude zariadenie používať, počte vstupných a výstupných kanálov, a predvolenej vzorkovacej frekvencii zariadenia. Špeciálnym prípadom vstupného zariadenia je Stereo Mix – výberom tohto zariadenia sa vstupom do programu stane prúd údajov, ktorý je operačným systémom odosielaný na výstupné zariadenie (napr. reproduktory). Pomocou Stereo Mix tak môžeme ako vstup do nášho programu použiť napr. zvuk z práve prehrávaného videa alebo telefonátu cez Skype.



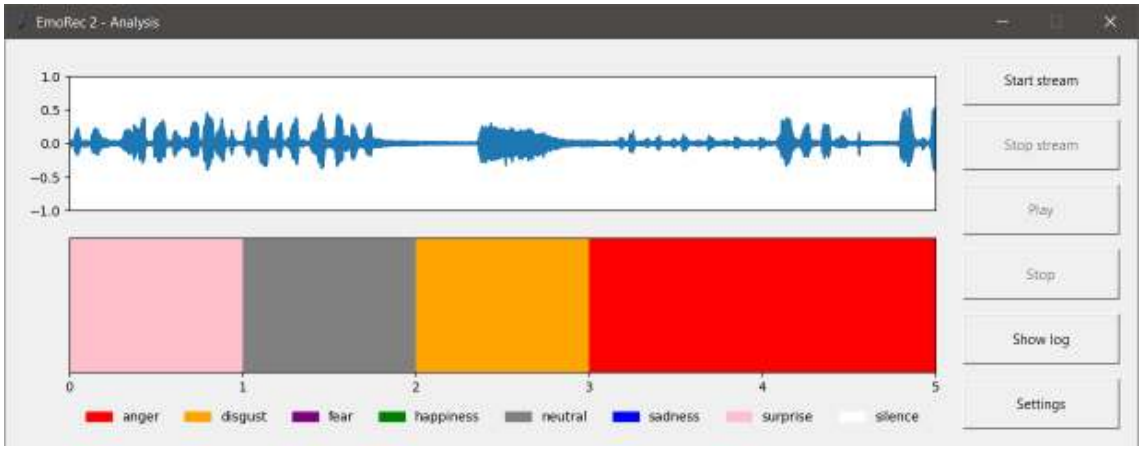
Obrázok 20: Zoznam vstupných zariadení.

Sample rate – pomocou tohto parametra je možné zmeniť vzorkovaciu frekvenciu vstupného zariadenia. Frekvenciu môžeme nastaviť v rozsahu od 44100 Hz do 96000 Hz. Vyššia vzorkovacia frekvencia vernejšie zachováva detaily pri konverzii analógového signálu na digitálny, ale zvyšuje nároky na pamäť, a viac zaťažuje procesor, pretože charakteristiky je potrebné extrahovať z väčšieho množstva vzoriek.

Predictive window length – tento parameter určuje počet sekúnd, počas ktorých sa zbierajú vzorky. Po nazbieraní vzoriek za tento časový úsek sa naraz spracujú ako jeden rámec a vykoná sa predikcia. Zároveň sa začínajú zbierať vzorky pre ďalší rámec. Dĺžku je možné nastaviť od 2 do 10 sekúnd. Kratšie časové okno spôsobuje, že výpočet charakteristík a predikcia sa musí vykonávať častejšie. Pri nastavení väčšieho počtu sekúnd sa naopak zníži frekvencia predikcie, ale zvyšuje sa náročnosť výpočtu charakteristík, a ich spracovania pred predikciou. Toto nastavenie určuje dĺžku rámca aj pri analýze .wav súboru.

Buffer size – nastavuje veľkosť zásobníka ovládača vybraného zariadenia. K dispozícii je veľkosť od 64 do 2048 vzoriek. Nízka kapacita zásobníka ovládača núti našu aplikáciu k častému vyberaniu vzoriek, a ich presunu do vlastného zásobníka. To zvyšuje záťaž procesora. Zvýšenie kapacity zároveň spôsobí aj zvýšenie latencie. Latencia je oneskorenie presunu vzoriek zo zásobníka ovládača vstupného zariadenia do našej aplikácie, a typicky sa pohybuje v desiatkach až stovkách milisekúnd. Nízka latencia je dôležitá predovšetkým pre úlohy ako napr. nahrávanie hudby, a nemá vplyv na výsledky nášho programu, preto odporúčame používať zásobník s vyššou kapacitou.

Po nastavení parametrov alebo výbere vstupného súboru sa zobrazí okno, pomocou ktorého je možné ovládať samotnú analýzu. Na tomto okne sa nachádza aj graf, ktorý zobrazuje priebeh analýzy emócií. Pozostáva z dvoch častí – jedna zobrazuje priebeh vstupného signálu, a druhá pomocou farebného kódu zobrazuje emóciu, ktorá bola pre zodpovedajúci úsek signálu predikovaná. Pri analýze v reálnom čase sa zobrazuje iba posledných 5 rámcov. Dôvodom pre toto obmedzenie je potreba častého prekresľovania grafu, ktoré je pri veľkom množstve vzoriek pomalé. Ak bol načítaný súbor, potom je možné na graf zobraziť všetky rámce, na ktoré bol súbor pri načítaní rozdelený, pretože sa graf vykreslí iba raz.



Obrázok 21: Ovládací panel analýzy.

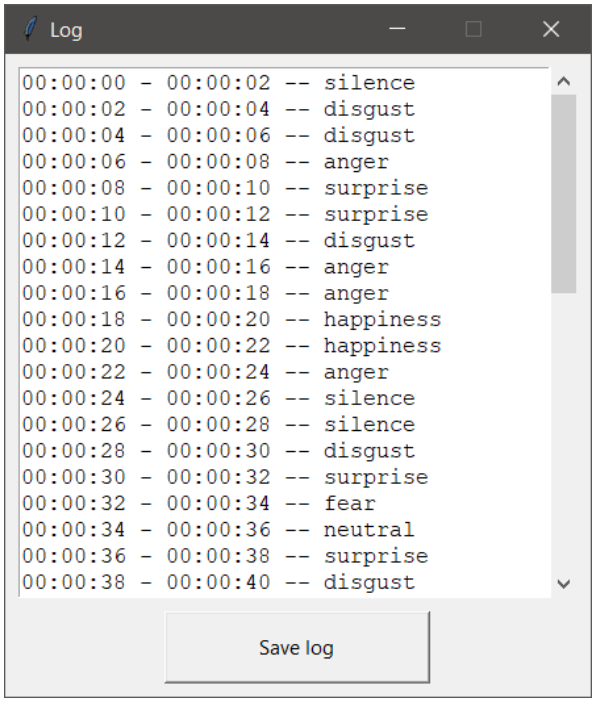
Pomocou tlačidiel v pravej časti okna je možné analýzu ovládať. Niektoré ovládacie prvky sú dynamicky aktivované alebo deaktivované v závislosti od situácie.

Start stream – tlačidlo otvára prúd údajov zo zvoleného vstupného zariadenia. Je aktívne v prípade, že bola zvolená analýza vstupu zo vstupného zariadenia, a prúd údajov nie je otvorený.

Stop stream – uzatvára prúd údajov. Aktivuje sa po otvorení prúdu údajov, a deaktivuje po jeho zatvorení.

Play, Stop – tieto tlačidlá sú aktívne iba v prípade, že bola vybraná analýza .wav súboru. Pomocou týchto tlačidiel je možné súbor prehrať, resp. prehrávanie zastaviť.

Show log – po predikcii sa určená emócia zapíše do logu. Okrem emócie sa zaznamenávajú aj informácie o čase začiatku a konca rámca, pre ktorý bola emócia predikovaná. Po kliknutí na toto tlačidlo sa log otvorí v novom okne, z ktorého je možné celý log skopírovať alebo uložiť vo formáte .csv. Tlačidlo je deaktivované ak je práve otvorený prúd údajov.

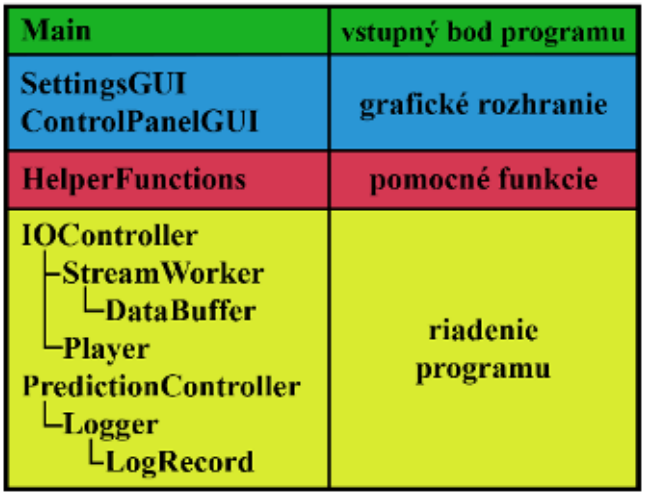


Obrázok 22: Ukážka logu predikcie.

Settings – umožňuje návrat na okno s nastaveniami. Pri otvorenom prúde údajov nie je návrat možný.

3.5.2 Štruktúra programu EmoRec 2

Na obrázku (Obrázok 23) je znázornená vnútorná štruktúra našej aplikácie. Triedy sme navrhli tak, aby sme čo najviac oddelili grafické rozhranie od logiky riadenia programu. Pri riadení sme naviac oddelili správu vstupných zariadení a prúdy údajov od samotného spracovania údajov a predikcie emócií.



Obrázok 23: Štruktúra tried.

Main – táto trieda slúži ako vstupný bod programu. Vytvára inštancie riadiacich tried, a spúšťa grafické rozhranie na hlavnej slučke.

SettingsGUI, ControlPanelGUI – v týchto dvoch triedach je definované rozloženie a funkcie okien nastavení a ovládacieho panelu analýzy.

HelperFunctions – pomocné funkcie pre zamkýnanie a odomkýnanie GUI komponentov. Využívajú ich obidve GUI triedy.

IOController – trieda implementuje funkcie a ďalšie vnorené triedy ovládajúce pripojené vstupné zariadenia, prúd údajov z nich a prehrávanie načítaných súborov.

StreamWorker – vnorená trieda v triede IOController. Jej úlohou je otvárať prúd údajov na vybranom zariadení, udržovať ho otvorený, a zbierať vzorky až kým nie je analýza zastavená.

DataBuffer – vnorená trieda triedy StreamWorker. Je v nej definovaná štruktúra zásobníka pre prichádzajúce vzorky, a mechanizmy pre ich ďalší pohyb.

Player – vnorená trieda v IOController. Úlohou prehrávača je načítanie vzoriek vybraného .wav súboru a jeho prehrávanie.

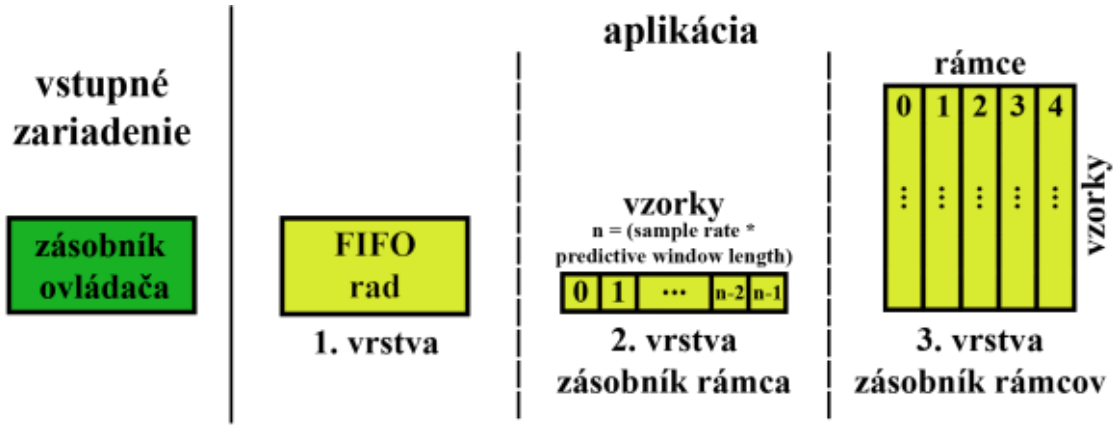
PredictionController – stará sa o načítanie modelu pri spustení programu, spracovanie rámcov, extrakciu charakteristík a samotnú predikciu emócií.

Logger – vnorená trieda v triede PredictionController, ktorej úlohou je uchovávanie záznamov o emóciách, ktoré boli v priebehu analýzy predikované.

LogRecord – vnorená trieda triedy Logger. Je v nej definovaná štruktúra záznamu logu a funkcie konvertujúce záznam na formu pre výpis v GUI alebo pre zápis do .csv súboru.

3.5.3 Princíp fungovania programu EmoRec 2

Analýza v reálnom čase – po výbere vstupného zariadenia a nastavení ostatných parametrov (SettingsGUI) analýza začne po kliknutí na tlačidlo Start stream (ControlPanelGUI). Na novom vlákne (IOController) sa otvorí prúd údajov z vybraného zariadenia (StreamWorker). Prichádzajúce vzorky sa pred spracovaním ukladajú do zásobníka. Za týmto účelom sme, s ohľadom na rýchlosť, navrhli trojvrstvový zásobník (DataBuffer). Prvú vrstvu tvorí FIFO štruktúra, ktorá slúži na priebežný presun vzoriek zo zásobníka ovládača do našej aplikácie. Táto štruktúra je veľmi rýchla, avšak okrem vkladania a výberu vzoriek neumožňuje žiadne iné užitočné operácie. Použili sme implementáciu z knižnice queue. Vzorky sú z prvej vrstvy priebežne vyberané a presúvané do druhej vrstvy, ktorou je zásobník rámca – štandardné indexované pole. Jeho veľkosť závisí od nastavenia vstupných parametrov, a je daná ako *(sample rate * predictive window length)*. Po získaní kompletného rámca sa vzorky presúvajú do tretej vrstvy, ktorou je dvojrozmerné pole. V tejto vrstve zásobníka je uložených posledných 5 rámcov (staršie rámce sa zahadzujú), pričom rámec sa považuje za najmenšiu, ďalej nedeliteľnú, jednotku.



Obrázok 24: Štruktúra zásobníka.

Po vložení nového rámca do tretej vrstvy sa zároveň spustí spracovanie tohto rámca na novom vlákne. Spracovanie (PredictionController) musí prebiehať rovnakým spôsobom, akým boli predspracované dáta, na ktorých bol natrénovaný model v experimente 2 – z rámca je orezané ticho na konci a na začiatku, rámec je

normalizovaný na hlasitosť 60 dB, pomocou knižnice Librosa je vypočítaných 13 mfcc charakteristík, energia a zerocrossing, vykoná sa interpolácia charakteristík, a nakoniec sa matica s charakteristikami transponuje. Model potom z tejto matice predikuje emóciu, a predikcia sa uloží do logu (Logger) spolu s časom začiatku a konca rámca, pričom čas sa počíta od spustenia analýzy. Spracovanie rámca a jeho predikcia sa nevykoná v prípade, že ani jedna vzorka v rámci neprekoná hlasitosť -20 dB (pri referenčnej hodnote, ktorú ako predvolenú používa funkcia *amplitude_to_db()* z knižnice librosa). V tomto prípade sa rámec do logu uloží ako ticho. Nakoniec sa ešte prekreslí graf na grafickom rozhraní (ControlPanelGUI). Na prvú časť grafu sa vykreslí 5 rámcov z tretej vrstvy zásobníka, a na druhú časť sa zobrazí 5 emócií z logu, ktoré boli pre tieto rámce určené.

Analýza súboru – po výbere .wav súboru z disku (SettingsGUI) sa preň vytvorí prehrávač (Player). Prehrávač automaticky zistí vzorkovaciu frekvenciu súboru, a načíta do pamäte všetky vzorky. Pole so vzorkami je rozdelené na rámce (ako dĺžka rámca sa použije hodnota, ktorá je pri načítaní nastavená v poli Predictive window length). Predspracovanie a predikcia postupne prebehnú pre všetky rámce rovnakým spôsobom ako pri analýze v reálnom čase (PredictionController). Aj tentokrát sa výsledky predikcie zapisujú do logu (Logger). Na graf sa potom vykreslia naraz všetky rámce spolu s predikciami (ControlPanelGUI). Prehrávač takisto disponuje možnosťou načítaný súbor prehrať, prípadne prehrávanie zastaviť.

4 MOŽNOSTI VYLEPŠOVANIA A ROZŠIROVANIA

Po vypracovaní vlastného riešenia je vhodné sa ešte raz zamyslieť nad celým postupom, a identifikovať kroky, ktoré poskytujú priestor na vylepšenie alebo k nim existuje alternatívny postup. Zamerali sme sa na 4 hlavné časti – vstupné údaje, ich spracovanie, experimenty a samotnú aplikáciu.

4.1 VSTUPNÉ ÚDAJE

Správanie sa modelov v testoch ukázalo, že najlepšie dokázali rozpoznať emócie v nahrávkach, ktoré pochádzali z rovnakej databázy, na akej boli trénované (RAVDESS). V testoch so zvyšnými dvomi databázami bola úspešnosť predikcie oveľa nižšia. Úspešnosť modelu je silne závislá od kvality vstupných údajov. Model nie je použiteľný v praxi, ak bol trénovaný na nahrávkach, ktoré nie sú dostatočným odrazom prejavov emócií v realite. Ako už bolo spomenuté (3.2 a 3.4.4), nahrávkam, ktoré sme použili na tréning modelov chýba rôznorodosť. Obmedzujú sa iba na dve veľmi podobné anglické vety, sú približne rovnako dlhé, a boli nahrané malou skupinou hercov, ktorá nereprezentuje všetky vekové kategórie. Nahrávok je navyše málo – trénovaciu podmnožinu tvorí 1064 súborov, teda na jednu emóciu pripadá 152 súborov. Ich počet sme dokázali zvýšiť augmentáciou, avšak takto získané súbory sú iba úpravou pôvodných, a teda nie sú riešením ostatných nedostatkov.

Lepším riešením by bolo získanie úplne nových nahrávok. Počas písania tejto práce sme nenašli žiadnu verejne prístupnú databázu, v ktorej by bolo viac súborov ako v databáze RAVDESS, a zároveň by v nej boli obsiahnuté všetky emócie Ekmanovej schémy a obidve pohlavia. Jednou z možností je pokúsiť sa získať napríklad nahrávky z call centier alebo iných zdrojov. Ich výhodou je, že zachytávajú správanie ľudí a emócie v reálnych situáciách. Problémom môže byť ochrana súkromia, prípadne autorské práva, a takisto nízka kvalita nahrávok. Takéto nahrávky by bolo pravdepodobne nutné kategorizovať (teda priradiť emóciu k nahrávke) s pomocou psychológa.

Iným spôsobom ako získať viac nahrávok je vytvorenie vlastnej databázy, napríklad s pomocou divadelných hercov. To by vyžadovalo spoluprácu odborníkov z viacerých oblastí – psychológie, lingvistiky a zvukového inžinierstva. Aby bola novovzniknutá databáza prínosná, musela by obsahovať oveľa väčšie množstvo nahrávok (desiatky až stovky tisíc), ako sa nachádza v existujúcich databázach. Takáto databáza by

ďalej musela obsahovať nahrávky viet rôznej dĺžky, a mala by byť nahratá vo veľmi dobrej kvalite – v prípade potreby je ľahšie degradovať kvalitné súbory, ako sa pokúsiť vylepšiť nekvalitné. Takisto je potrebné dbať na uniformitu tried. Vytvoriť tak veľkú databázu v jednom jazyku môže byť problematické, preto by mohla byť vytvorená databáza niekoľkých príbuzných jazykov – napríklad slovensko-česko-poľská databáza. Proces tvorby novej databázy pravdepodobne bude náročný na čas a financie, avšak dostupnosť takejto databázy by znamenala zásadný posun v problematike strojového rozpoznávania emócií, pretože model by bol pri tréningu nútený nájsť oveľa univerzálnejšie pravidlá, čím by sa predišlo jeho preučeniu, a bol by oveľa viac pripravený na reálne situácie.

4.2 SPRACOVANIE ÚDAJOV

Niektoré kroky vo fáze spracovania údajov pred trénovaním modelu, resp. ich neskoršieho spracovania v aplikácii, môžu byť zlepšené. V predspracovaní sme orezali ticho zo začiatku a konca súborov aby sme sa zbavili častí, ktoré neprinášajú žiadne relevantné informácie o emocionálnom kontexte, naopak, môžu skresliť merania charakteristík. V databázach reči každá nahrávka obsahuje práve jednu vetu, avšak pri spracovaní v reálnom čase nevieme zaručiť, že veta začne aj skončí v tom istom rámci. Často sa stáva, že pauza medzi vetami, ktorú by bolo vhodné orezať, sa nachádza v strede rámca. Lepším riešením pre aplikáciu by bolo nepoužívať pevnú dĺžku rámca, ale spracovávať prichádzajúce vzorky dynamicky, a predikciu emócie vykonávať iba ak dôjde k splneniu určitých podmienok. Takou podmienkou by mohlo byť vykonanie predikcie vždy keď hovoriaci dokončí vetu. Začiatok a koniec vety by mohol byť rozpoznávaný podobne ako pri speech-to-text systémoch.

Aby sme čo najviac eliminovali fakt, že vstupná úroveň signálu môže byť ovplyvnená technikou a vzdialenosťou od mikrofónu, normalizovali sme každý súbor na spoločný priemer – 60 dB. Rovnako sme normalizovali aj rámce počas spracovania v aplikácii. Vedľajším efektom je, že energické emócie majú rovnakú hlasitosť ako emócie, pre ktoré je typický skôr tichý a nevýrazný prejav. Riešenie problému techniky a vzdialenosti, ktoré zároveň nebude mať podobné neželané efekty, je kalibrácia. Model by bol natrénovaný so súbormi bez normalizácie. Zároveň by sme zistili priemernú hlasitosť z nahrávok neutrálnej emócie. Pri spustení programu by sa vykonala kalibrácia – používateľ by bol požiadaný o prečítanie vopred daného textu s neutrálnym výrazom.

Počas tohto kroku by sa zároveň zistil rozdiel intenzity vstupu a nahrávok v databáze, a na základe tohto rozdielu by bola prispôsobená hlasitosť vstupu počas analýzy. Aj toto riešenie má negatíva – práca s aplikáciou sa stane menej pohodlnou, pretože kalibráciu by bolo potrebné vykonať vždy pri zmene techniky alebo vystriedaní používateľa.

4.3 EXPERIMENTY

Problémami experimentu 1 sú nízka úspešnosť a rýchlosť. Extrakcia charakteristík bola v tomto experimente najpomalšia zo všetkých. Príčinou môže byť knižnica parselmouth, ktorá používa pôvodný kód programu PRAAT, avšak tento program nebol navrhnutý pre použitie v reálnom čase. Vhodným pozmenením tohto experimentu by sme mohli dosiahnuť vyššiu rýchlosť extrakcie. V súčasnosti sú z charakteristík počítané hodnoty ako napr. minimá, maximá alebo priemery. Namiesto toho by sme mohli použiť podobný postup ako v experimente 2, a namiesto počítania týchto hodnôt zachovať ich spojitý charakter (priebeh základnej frekvencie, formantov, intenzity, atď. v čase). Na predikciu emócií by sme potom použili 1D konvolučnú neurónovú sieť.

Model v experimente 2 bol najúspešnejší, avšak ešte stále nie je vhodný pre použitie v praxi. Experiment je vhodné zopakovať s väčším počtom mfcc charakteristík (napríklad pridaním ich derivácií), a zistiť ako bude ovplyvnená úspešnosť a časová náročnosť spracovania.

V experimente 3 sme v porovnaní s experimentom 2 dosiahli o niečo nižšiu úspešnosť. Problematickejšia bola rýchlosť – spektrogram je potrebné vypočítať a následne uložiť na pevný disk vo forme obrázku. Obrázok sa potom stáva vstupom do 2D konvolučnej siete. Vynechanie ukladania obrázkov na disk by potenciálne mohlo celý proces urýchliť. Ako vstup do konvolučnej siete by bola použitá matica, ktorá vznikne po vypočítaní spektrogramu, avšak podobne ako pri charakteristikách v experimente 2, takéto matice by sme najprv museli vhodne upraviť tak aby mali vždy rovnakú veľkosť. Aj keby sa takýmto spôsobom podarilo zrýchliť spracovanie zvukových súborov, k zrýchleniu samotnej predikcie pravdepodobne nedôjde.

4.4 APLIKÁCIA

Výhodou jazyka Python je, že preň existuje mnoho užitočných a jednoducho použiteľných knižníc, predovšetkým pre riešenie matematicky zameraných problémov

a strojové učenie. K dispozícií je aj niekoľko knižníc, ktoré dokážu z hlasu extrahovať rôzne charakteristiky. Možnosť ich použitia je pre nás veľmi výhodná, avšak spôsob akým bol Python navrhnutý ho robí menej vhodným pre úlohy, ktoré sú vykonávané v reálnom čase.

Jednou z jeho vlastností je, že je to interpretovaný jazyk. To znamená, že interpreter číta inštrukcie, a postupne ich vykonáva. Je preto o niečo pomalší ako kompilované jazyky, v ktorých je kód pred spustením najprv kompilátorom preložený do strojového kódu. V prípade požiadavky spracovávania údajov v reálnom čase môže aj malé zrýchlenie znamenať výrazné vylepšenie programu.

Ďalší problém spôsobujú vlákna. Knižnica tkinter musí bežať na hlavnej slučke a akákoľvek dlhšia úloha spustená na rovnakej slučke by spôsobila zamrznutie grafického rozhrania. V našom programe musíme konštantne spracovávať prúd údajov zo vstupného zariadenia a vykonávať predikciu emócií. Aby sme zabránili zamrznutiu GUI, musíme tieto úlohy spúšťať na samostatných vláknach, avšak v jazyku Python všetky vytvorené vlákna bežia na tom istom jadre procesora, pretože používa zámok, ktorý sa označuje ako global interpreter lock (GIL). Aj napriek tomu, že sme vytvorili viac vlákien, nemôžu bežať skutočne paralelne – GIL si vlákna odovzdávajú a môže bežať iba vlákno, ktoré GIL aktuálne má. Časovo najnáročnejšou úlohou je prekreslenie grafu na grafickom rozhraní. Ak prekreslenie trvá príliš dlho, môže sa stať, že zásobník ovládača sa naplní vzorkami, ale vlákno, s ktorým bol prúd údajov otvorený nedostane GIL včas, vzorky sa nestihnú presunúť do našej aplikácie, a zásobník pretečie. Riešením by mohla byť knižnica multiprocessing, ktorá umožňuje vytváranie nových procesov bez obmedzenia na jedno jadro procesora, avšak vytvárať nové procesy je pre nás nerentabilné. Vytvoriť nový proces trvá dlhšie ako vytvoriť vlákno existujúceho procesu, procesy nemajú zdieľanú pamäť a bolo by potrebné navrhnúť spôsob komunikácie medzi nimi.

Za zváženie stojí prepísanie aplikácie do iného jazyka. Vhodným jazykom by mohol byť napríklad C#, ktorý je kompilovaný, a nové vlákna sú jadrám procesora prideľované operačným systémom. Nevýhodou je, že C# nedisponuje tak širokým výberom knižníc ako Python. Zaujímavou alternatívou by mohli byť aj jazyky C++ alebo IronPython[16]. Je potrebné zaoberať sa aj dostupnosťou a možnosťami existujúcich

[16] ironpython.net

modulov pre potenciálne vhodné jazyky (napr. Accord.NET[17] pre C#). Ak by sme na extrakciu hlasových charakteristík nenašli vhodnú knižnicu, museli by sme si ju napísať sami. Pochopenie, implementácia a odladenie takýchto algoritmov vyžaduje kombináciu značných matematických a programátorských schopností, a je časovo náročná.

[17] accord-framework.net

ZÁVER

V tejto práci sme experimentálne preskúmali možnosti klasifikácie emócií pomocou neurónových sietí, a ich vhodnosť pre použitie pri klasifikácii v reálnom čase. Ako metóda s veľkým potenciálom v budúcnosti sa ukazuje použitie konvolučných neurónových sietí, ktoré ako svoj vstup používajú časový priebeh mfcc charakteristík. Tento spôsob klasifikácie je veľmi rýchly, a aj napriek nedostatočným údajom, ktoré sme mali k dispozícii na trénovanie modelu, dosiahol dobré výsledky.

Naprogramovali sme vlastné funkčné riešenie, ktoré pomocou pripojeného vstupného zariadenia sníma hlas používateľa, a v reálnom čase vyhodnocuje jeho emocionálny stav. Napriek tomu, že takéto riešenie je veľkým krokom k praktickému použitiu, jeho úspešnosť stále nie je dostatočná, pretože je závislá predovšetkým od kvality použitého modelu. Preto sme identifikovali časti riešenia, ktoré vyžadujú ďalšie experimenty a vylepšenia. Dôležité je, že sme ukázali, že aj takto náročnú úlohu je možné úspešne vyriešiť, a postupným vylepšovaním dospieť až k aplikácii, ktorá bude úspešná aj v praxi.

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV.

ABHANG, P. A.; GAWALI, B. W.; MEHROTRA, S. 2016. *Introduction to EEG- and Speech-Based Emotion Recognition*. San Diego: Academic Press, 2016. 198 s. ISBN 978-0-12-804490-2.

ALBAWI, S.; MOHAMMED, T. A.; AL-AZAWI, S. 2017. Understanding of a Convolutional Neural Network. In *2017 International Conference on Engineering and Technology (ICET)*. Akdeniz University, Antalya. 6 s. DOI 10.1109/ICEngTechnol.2017.8308186.

BACHU, R. G.; KOPPARTHI, S.; ADAPA, B.; BUKET, B. 2010. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. In *Advanced Techniques in Computing Sciences and Software Engineering*. Berlin: Springer, 2010. ISBN 978-90-481-3659-9, p.279-282.

BAHREINI, K.; NADOLSKI, R.; WESTERA, W. 2016. Towards real-time speech emotion recognition for affective e-learning. In *Education and information technologies*. ISSN 1367-1386, vol. 21, no. 5, p.1367-1386.

BATLINER, A.; SCHULLER, B.; SEPPI, D.; STEIDL, S.; DEVILLERS, L.; VIDRASCU, L.; VOGT, T.; & AHARONSON, V.; AMIR, N. 2011. The Automatic Recognition of Emotions in Speech. In *Emotion-Oriented Systems*. Berlin: Springer, 2011. ISBN 978-3-642-15184-2, p.71-99.

BERANEK, L.; MELLOW, T. 2019. *Acoustics: Sound Fields, Transducers and Vibration*. San Diego: Academic Press, 2019. 900 s. ISBN 978-0-12-815227-0.

BOERSMA, P., V. 2001. Praat, a system for doing phonetics by computer. In *Glot International*. ISSN 1381-3439, vol.5, no.9, p.341-345.

BUSSO, C.; LEE, S.; NARAYANAN, S. S. 2007. Using Neutral Speech Models for Emotional Speech Analysis. In *8th Annual Conference Of The International Speech Communication Association (Interspeech 2007)*, vol. 4. Red Hook: Curran Associates Inc, 2008. ISBN 978-1-60560-316-2, p.2304–2308.

CEN, L.; WU, F.; YU, Z. L.; HU, F. 2015. A real-time speech emotion recognition system and its application in online learning. In *Emotions, technology, design, and learning*. San Diego: Academic Press, 2015. ISBN 978-0-12-801856-9, p.27-46.

CHERNYKH, V.; PRIKHODKO, P. 2018. *Emotion recognition from speech with recurrent neural networks*. 14 s. arXiv preprint arXiv:1701.08071v2.

CHEUNG, S.; LIM, J. S. 1992. Combined multiresolution (wide-band/narrow-band) spectrogram. In *IEEE Transactions on Signal Processing*. ISSN 1053-587X, vol.40, no.4, p.975–977.

EL AYADI, M.; KAMEL, M. S.; KARRAY, F. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. In *Pattern Recognition*. ISSN 0031-3203, vol.44, no.3, p.572–587.

EYBEN, F.; WENINGER, F.; WÖLLMER, M.; SCHULLER, B. 2016. OpenSMILE:) Open-Source Media Interpretation by Large Feature-Space Extraction Version 2.3. [online]. 65 s. [citované: 19.3. 2020]. Dostupné na internete: <audeering.com/download/opensmile-book-latest/?wpdmdl=4841&refresh=5eaca5ec421921588372972>

ETIENNE, C.; FIDANZA, G.; PETROVSKII, A.; DEVILLERS, L.; SCHMAUCH, B. 2018. CNN+LST architecture for speech emotion recognition with data augmentation. In *Proc. Workshop on Speech, Music and Mind 2018*. DOI 10.21437/SMM.2018-5. p.21-25.

BURKHARDT F.; PAESCHKE A.; ROLFES, M.; SENDLMEIER W. F.; WEISS B. 2005. A Database of German Emotional Speech. In *6th Interspeech 2005 and 9th European Conference on Speech Communication and Technology (EUROSPEECH)*, vol.3. Red Hook: Curran Associates Inc, 2008. ISBN 978-1-60423-448-0, p.1516-1520.

FENG, T.; YANG, S. 2018. Speech Emotion Recognition Based on LSTM and Mel Scale Wavelet Packet Decomposition. In *ACAI 2018: Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*. article no.38. 7 s. DOI 10.1145/3302425.3302444.

GANGAMOHAN, P.; KADIRI, S. R.; YEGNANARAYANA, B. 2016. Analysis of Emotional Speech—A Review. In *Intelligent Systems Reference Library*, ISSN 1868-4394, vol.105, p.205–238.

GAO, Y.; LI, B.; WANG, N.; ZHU, T. 2017. Speech Emotion Recognition Using Local and Global Features. In *Lecture Notes in Computer Science*. ISSN 0302-9743, vol.10654, p.3–13.

GIANNAKOPOULOS T. 2015. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. In *PLoS ONE*. 17s. ISSN 1932-6203, vol.10, no.12, PMCID PMC4676707.

HAJAROLASVADI, N.; DEMIREL, H. 2019. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. In *Entropy*. 17 s. ISSN 1099-4300, vol.21, no.5, DOI 10.3390/e21050479.

HAN, K.; YU, D.; TASHEV, I. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, vol.1. Red Hook: Curran Associates Inc, 2015. ISBN 978-1-63439-435-2. p.223-227.

HAYKIN, S. 2010. *Neural Networks and Learning Machines, 3rd Edition*. Upper Saddle River: Pearson Prentice Hall, 2010. 936 s. ISBN 978-81-203-4000-8.

HILLENBRAND, J. 2011. Acoustic Analysis of Voice: A Tutorial. In *Perspectives on Speech Science and Orofacial Disorders*. ISSN 1940-7572, vol.21, no.2, p.31-43.

HUANG, Y.; HU, M.; YU, X.; WANG, T.; YANG, C. 2016. Transfer Learning of Deep Neural Network for Speech Emotion Recognition. In *Pattern Recognition: 7th Chinese Conference, CCPR 2016, Chengdu, China, November 5-7, 2016, Proceedings, Part II*. ISSN 1865-0929, p.721–729.

JADOUL, Y.; THOMPSON, B.; DE BOER, B. 2018. Introducing parselmouth: A python interface to praat. In *Journal of Phonetics*. ISSN 0095-4470, vol.71, p.1-15.

KELLER, E. 2004. The Analysis of Voice Quality in Speech Processing. In *Nonlinear Speech Modeling and Applications*. Berlin: Springer, 2005. ISBN 978-3-540-27441-4, p.54-73.

KIM, S.; GEORGIOU, P. G.; LEE, S.; NARAYANAN, S. 2007. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *2007 IEEE 9th Workshop on Multimedia Signal Processing*. Montreal: IEEE, 2007. ISBN 978-1-4244-1274-7, p. 48-51.

LATIF, S.; RANA, R.; KHALIFA, S.; JURDAK, R.; EPPS, J. 2019. Direct modelling of speech emotion from raw speech. In *Proc. Interspeech 2019*. ISSN 1990-9772, p. 3920-3924.

LIVINGSTONE, S. R.; RUSSO, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. In *PLoS ONE*. 35 s. ISSN 1932-6203, vol.13, no.5, PMCID PMC5955500.

LOIZOU, P. C. 2013. *Speech enhancement: theory and practice 2nd edition*. 711 s. Boca Raton: CRC press, 2017. ISBN 978-1-138-07557-3.

JACKSON, P. J. B.; Haq, S. 2011. Surrey Audio-Visual Expressed Emotion (SAVEE) database.

MAHALAKSHMI, P. 2016. A Review On Voice Activity Detection And Mel-Frequency Cepstral Coefficients For Speaker Recognition (Trend Analysis). In *Asian Journal of Pharmaceutical and Clinical Research*. ISSN 0974-2441, vol.9, no.9, p.360-363.

MCFEE, B.; RAFFEL, C.; LIANG, D.; ELLIS, D. P.; MCVICAR, M.; BATTENBERG, E.; NIETO, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*. ISSN 2575-9752, vol.8, p.18-25.

MENON, A. G.; ANJUSHA, V. K. 2017. Analysis of Feature Extraction Methods for Speech Recognition. In *IJISET-International Journal of Innovative Science, Engineering & Technology*. ISSN 2348-7968 , vol.4, no.4, p.328-333.

MIRSAMADI, S.; BARSOUM, E.; ZHANG, C. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention, In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Montreal: IEEE, 2017. ISBN 978-1-5090-4117-6, p. 2227-2231.

NIELSEN, M. A. 2015. *Neural networks and deep learning*. [online] 216 s. [citované 3.3. 2020]. Dostupné na internete: <https://s3.amazonaws.com/academia.edu.documents/58251403/neural_networks_and_deep_learning.pdf>

NIU, Y.; ZOU, D.; NIU, Y.; HE, Z.; TAN, H. 2017. *A breakthrough in speech emotion recognition using deep retinal convolution neural networks*. 7 s. arXiv preprint arXiv:1707.09917.

QAYYUM, A. B. A.; AREFEEN, A.; SHAHNAZ, C. 2019. Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. In *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*. Montreal: IEEE, 2019. ISBN 978-1-7281-4294-4. p.122-125.

RAO, K. S.; REDDY, V. R.; MAITY, S. 2015. *Language Identification Using Spectral and Prosodic Features (SpringerBriefs in Electrical and Computer Engineering)*. 98 s. Cham: Springer, 2015. ISBN 978-3-319-17162-3.

RAO, K. S.; KOOLAGUDI, S. G.; VEMPADA, R. R. 2012. Emotion recognition from speech using global and local prosodic features. In *International Journal of Speech Technology*. ISSN 1381-2416, vol.16, no.2, p.143–160.

RENJITH, S.; MANJU, K. G. 2017. Speech based emotion recognition in Tamil and Telugu using LPCC and hurst parameters—a comparitive study using KNN and ANN classifiers. In *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT 2017)*. Montreal: IEEE, 2017. ISBN 978-1-5090-4967-7, p.311-317.

SATO, N.; OBUCHI, Y. 2007. Emotion Recognition using Mel-Frequency Cepstral Coefficients. In *Journal of Natural Language Processing*. ISSN 1340-7619, vol.14, no.4, p.83–96.

SATT, A.; ROZENGERG, S.; HOORY, R. 2017. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In *Proc. Interspeech 2017*. ISSN 1990-9772, p.1089-1093.

SHAW A.; VARDHAN, R. K.; SAXENA, S. 2016. Emotion recognition and classification in speech using Artificial neural networks. In *International Journal of Computer Applications*. ISSN 0975 – 8887, vol.145, no.8, p.5-9.

SHIH, P. Y.; CHEN, C. P.; WANG, H. M. 2017. Speech emotion recognition with skew-robust neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE: Montreal, 2017. ISBN 978-1-5090-4117-6, p.2751-2755.

SULKA, T. 2008. *Detekcia a klasifikácia emocionálneho stavu používateľa na základe hlasu*. 43 s. Nitra: Univerzita Konštantína Filozofa.

SUNNY, S.; PETER, D.; JACOB, K. P. 2012. Recognition of speech signals: an experimental comparison of linear predictive coding and discrete wavelet transforms. *International Journal of Engineering Science*. ISSN 0975-5462, vol.4, no.4, p.1594-1601.

STEIDL, S. 2008. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Erlangen: University of Erlangen-Nuremberg, 2009. 260 s. ISBN 978-383251455.

STOLAR, M. N.; LECH, M.; BOLIA, R. S.; SKINNER, M. 2017. Real time speech emotion recognition using RGB image classification and transfer learning. In *2017 11th International Conference on Signal Processing and Communication Systems (ICSPCS)*. 8 s. Montreal: IEEE, 2017. ISBN 978-1-5386-2887-4. DOI 10.1109/ICSPCS.2017.8270472

THAKARE, C.; CHAURASIA, N. K.; RATHOD, D.; JOSHI, G.; GUDADHE, S. 2019. Comparative Analysis of Emotion Recognition System. In *International Research Journal of Engineering and Technology (IRJET)*. ISSN 2395-0072, vol.6, no.2, p.380-384.

TZINIS, E.; POTAMIANOS, A. 2017. Segment-based speech emotion recognition using recurrent neural networks. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. Montreal: IEEE, 2017. ISBN 978-1-5386-0564-6, p.190-195.

VASILEV, I.; SLATER, D.; SPACAGNA, G.; ROELANTS, P.; ZOCCA, V. 2019. *Python Deep Learning: Exploring deep learning techniques and neural network architectures with Pytorch, Keras, and TensorFlow*. 386 s. Birmingham: Packt Publishing Ltd, 2019. ISBN 978-1-78934-846-0.

VOGT, T.; ANDRÉ, E.; BEE, N. 2008. EmoVoice—A framework for online recognition of emotions from voice. In *4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Berlin: Springer, 2008. ISBN 978-3-540-69368-0, p.188-199.

WAGNER, J.; LINGENFELSER, F.; BAUR, T.; DAMIAN, I.; KISTLER, F.; ANDRÉ, E. 2013. The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *Proceedings of the 21st ACM international conference on Multimedia*. New York: Association for Computing Machinery, 2013. ISBN 978-1-4503-2404-5, p.831-834.

WILDER, L. 1975. Articulatory and Acoustic Characteristics of Speech Sounds. In *Understanding Language*. New York: Academic Press, 1975. ISBN 978-0-12-478350-8, p.31–76.

ZEWOUDIE, A. W.; LUQUE, J.; HERNANDO, J. 2016. Short- and Long-Term Speech Features for Hybrid HMM-i-Vector based Speaker Diarization System. In *Proc. Odyssey 2016*. ISSN 2312-2846, p.400-406.

ZOU, J.; HAN, Y.; SO, S. S. 2008. Overview of Artificial Neural Networks. In *Artificial Neural Networks*. New Jersey: Humana Press, 2009. ISBN 978-1-58829-718-1, p.14–22.

ZOZNAM PRÍLOH

Príloha A – zdrojový kód programu EmoRec 2

Príloha B – skript *PreprocessData.py* vytvorený na zautomatizovanie predspracovania dát

Príloha C – súbory súvisiace s experimentom 1

Príloha D – súbory súvisiace s experimentom 2

Príloha E – súbory súvisiace s experimentom 3

Príloha F – kompletné výsledky experimentov