SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1177

Klasifikacija emocija konvolucijskim neuronskim mrežama pomoću glasovnih podataka

Nikola Vrebčević

Zagreb, srpanj 2018.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`.

Zahvaljujem prof. dr. sc. Davoru Petrinoviću na mentorstvu tijekom diplomskog studija te mag. ing. Igoru Mijiću na strpljenju i pomoći koju mi je pružio u izradi ovog rada.

SADRŽAJ

1. Uvod

Ljudski glas je jedinstveni oblik signala koji većina ljudi jednostavno generira i obrađuje nizom fizioloških i kognitivnih procesa. Glas je najčešće sredstvo govora u kojem se apstraktne misli pretvaraju u fizikalnu pojavu i obratno. Empirijski i teoretski je poznato da dvije osobe koje verbalno komuniciraju primaju informacije iz značenja riječi te informacije iz neverbalnog dijela komunikacije. Neverbalna komunikacija ostvaruje se kroz izraze lica, gestikulaciju rukama i stav tijela, a pobliže opisuju stavove, razmišljanja i emocionalna stanja pojedinca [1]. Također je bitna paralingvistička informacija, odnosno informacija koja dolazi pakirana s leksičkim sadržajem, a odnosi se na informacijski sadržaj samog govornog signala [2]. Ovaj diplomski rad proučavat će klasifikaciju emocija u govornom signalu na temelju paralingvističke informacije u frekvencijskoj domeni. Proučavanje emocija u glasovnom modalitetu stvara jednostavnu i kvalitetnu podlogu za praktičnu primjenu rezultata istraživanja jer minimalni tehnički uvjeti nisu složeni, a prostor primjena je širok. U vidu klasifikacije emocija, tehnologija se kreće u smjeru transformacije računala u pomoćne strojeve koji će odgovore na naredbe korisnika prilagoditi trenutnom emocionalnom stanju korisnika. Najčešće primjene jesu računalni agenti za pomoć pri učenju, sustavi u pozivnim centrima, sustavi za praćenje stanja vozača u automobilskoj industriji te industrija zabave [3].

Postupak klasifikacije općenito je moguće provesti na dva načina. Prvi način je tradicionalan te se temelji na određivanju značajki iz promatranog skupa podataka. Za svaki skup podataka potrebno je precizno odrediti značajke koje jednoznačno mogu diskriminirati različite klase unutar jednog skupa. Izračunate značajke prosljeđuju se modelu koji vrši klasifikaciju. Najpopularniji modeli korišteni u klasifikaciji nad govornim podacima su Gaussove mješavine, skriveni Markovljevi lanci, logistička regresija i modeli potpornih vektora (engl. *Support Vector Machine – SVM*). Moderniji i trenutačno popularniji način klasifikacije temelji se na metodama dubokog učenja. Kombiniranjem različitog broja perceptrona u različitim konfiguracijama formiraju se različite vrste neuronskih mreža. Neuronske mreže su modeli koje karakterizira mo-

gućnost da samostalno ekstrahiraju relevantne značajke za proces klasifikacije i na taj način reduciraju vjerojatnost izbora irelevantnih značajki koja postoji prilikom korištenja tradicionalnih modela. U ovom radu primjenjuje se drugi pristup klasifikacije korištenjem konvolucijskih, rekurentnih i dubokih neuronskih mreža. Bitna stavka u izradi modela jesu podaci pomoću kojih se model gradi, stoga su za potrebe ovog rada korišteni glasovni podaci iz multimodalne baze *RECOLA* koja objedinjuje izgovore na francuskom jeziku 23 govornika [4].

Ovaj rad će kroz poglavlje 2 Emocije i modeli klasifikacije pokazati značajke govora i građu modela koje se trenutno koriste u istraživanjima za potrebe klasifikacije; u poglavlju 3 Baze emocionalnih izgovora bit će prezentirane baze koje se najčešće koriste za detekciju emocija pomoću govornih podataka te načini predobrade izgovora iz *RECOLA* baze; arhitektura modela i rezultati istraživanja u okviru ovog diplomskog rada prezentirani su kroz poglavlje 4 Klasifikacija emocija.

2. Emocije i modeli klasifikacije

Emocije su odgovor organizma na određeni vanjski ili unutarnji podražaj kojeg osoba smatra značajnim, a manifestiraju se kroz cijeli organizam ili kroz određeni broj organskih sustava [5]. Iz ove definicije slijedi da klasifikacija emocija počiva na pretpostavci preslikavanja psihološkog stanja na izvršne organske sustave poput mišića kroz kompleksan neurološki i endokrini sustav. U vidu govora, emocije se odražavaju na izbor riječi i glasova koje govornik izgovara te na fizikalna svojstva govornog signala. Raspoznavanje emocija kroz određivanje značenja riječi kompleksan je zadatak koji ovisi o jeziku i govorniku. S druge strane, pomoću fizikalnih značajki glasa moguće je određivati emocije neovisno o jeziku i sadržaju govora. Negativna strana korištenja fizikalnih značajki je ta što ih ima puno te trenutno ne postoji skup značajki koji će jednoznačno opisivati emocije. Fizikalne značajke mogu se podijeliti u dvije skupine. Prvoj skupini pripadaju prozodijska obilježja poput osnovne frekvencije titranja glasnica, intenziteta i brzine govora. Druga skupina obuhvaća spektralne značajke govora [3].

Prije opisivanja spektralnih značajki govora, potrebno je poznavati mehanizam nastajanja govornog signala. Govorni signal sačinjavaju zvučni i bezvučni glasovi koji se razlikuju po tipu pobude. Zvučni glasovi nastaju tako da struja zraka iz pluća pobuđuje glasnice koje počnu titrati. Glasnice titranjem periodički propuštaju zrak iz pluća koji potom prolazi kroz ždrijelo, usnu i nosnu šupljinu. Frekvencija otpuštanja zraka iz pluća predstavlja osnovnu frekvenciju titranja glasnica koja definira visinu glasa, a proporcionalno se mijenja u ovisnosti o napetosti glasnica. Bezvučni glasovi nastaju slobodnim prolaskom zraka iz pluća u ždrijelo uz opuštene glasnice što znači da takvi glasovi ne nastaju kao odziv na harmonijsku pobudu, nego kao odziv na pobudu šumom. Preostali dio vokalnog trakta ima svojstvo definiranja vrste glasa tako da mijenjanjem oblika mijenja tok struje zraka te potiče stvaranje perturbacija koje rezultiraju različitim glasovima. Oblik vokalnog trakta određuje harmonijsku strukturu koja se u frekvencijskoj domeni govornog signala pojavljuje u obliku formanta određenog intenziteta, širine i frekvencije, dok će period titranja glasnica određivati harmonike glasa.

Analize govornog signala najčešće se vrše u frekvencijskoj domeni koja prika-

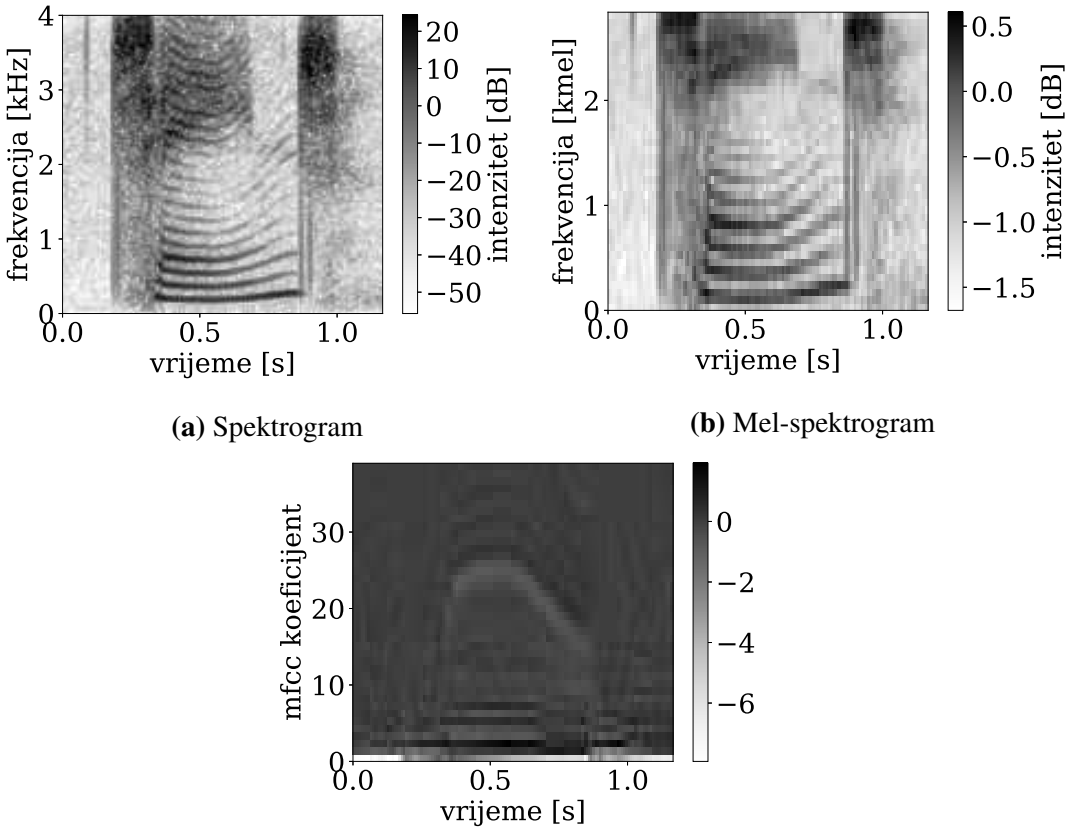
zuje razdiobu energije po frekvencijskim komponentama signala gdje se ovisnost o vremenu u potpunosti zanemaruje. Međutim, vremenska komponenta u govoru također ima svoje značenje te se pojavila potreba za proučavanjem govornog signala u vremenskoj i frekvencijskoj domeni istovremeno. Najpopularniji alat za prikazivanje promjene frekvencijskog sadržaja u vremenu je spektrogram, gdje je moguće pratiti promjene glasova koje odgovaraju promjeni formantne strukture (slika 2.1a). Frekvencijska i vremenska razlučivost spektrograma međusobno je obrnuto proporcionalna, a ovisna o širini otvora analize u vremenskoj domeni. Korištenje užih otvora rezultira slabijom frekvencijskom razlučivosti jer takav otvor u frekvencijskoj domeni obuhvaća više susjednih frekvencijskih komponenti za razliku od širokog otvora u vremenskoj domeni koji gotovo može izolirati susjedne frekvencijske komponente. Pri analizi govornih signala primjenjuje se preklapanje susjednih okvira analize što je sredstvo kontroliranja vremenske razlučivosti neovisno o širini vremenskog otvora. Spektrogram je moguće koristiti kao značajku pri klasifikaciji jer sadrži cjelovitu frekvencijsku informaciju o određenom signalu te je po definiciji prilagođen konvolucijskim neuronskim mrežama jer je dvodimenzionalni signal, a konvolucijske neuronske mreže razvile su se za potrebe klasifikacije slika. Istraživači najčešće koriste spektrograme kao ulazne podatke u izradi modela konvolucijskim mrežama jer vizualno reprezentiraju informaciju koju govorni signal posjeduje. Tako su u [6] korišteni spektrogrami izračunati iz *emoDB* baze, koji su iz nepoznatih originalnih dimenzija transformirani u spektrograme dimenzija 256×256. U [7] korišten je okvir analize širine 256 uzoraka uz 50% preklapanja između susjednih okvira što je rezultiralo spektrogramima dimenzije 128×128 izračunatih nad izgovorima *emoDB* baze. Konačno, istraživači u [8] koristili su Hammingov vremenski prozor dužine 20 ms kako bi obuhvatili barem dva perioda govornog signala najniže frekvencije ljudskog glasa koja iznosi 100 Hz gdje je rezultat bio spektrogram dimenzija 256×96 nad izgovorima u bazama *SUSAS*, *emoDB* i *eNTERFACE'05*.

Iduća spektralna značajka je melspektrogram (slika 2.1b) koji se izračunava iz spektrograma. Pošto ljudsko osjetilo sluha ima logaritamsku karakteristiku, moguće je frekvencijsku skalu iskazanu u Hertzima pretvoriti u mel skalu prema formuli 2.1 [9]. Pozitivna strana ovog prikaza je u tome što će doći do redukcije informacija na temelju osjetila, odnosno male promjene na višim frekvencijama bit će zanemarene dok će na niskim frekvencijama osjetljivost ostati ista kao pri skali u Hertzima. Melspektrogram izračunava se tako da se spektrogram propusti kroz niz pojasnopropusnih filtara koji obuhvaćaju veći frekvencijski pojas što im je centralna frekvencija viša, a međusobno se preklapaju te se u svakom pojasu izračunava logaritam energije signala.

Rezultat je određeni broj koeficijenta jednak broju upotrebljenih pojasnopropusnih filtara koji nose komprimiranu frekvencijsku informaciju o govornom signalu. Kao što je slučaj sa spektrogramima, i melspektrogrami su popularni oblik podataka za modele kojeg koriste konvolucijske neuronske mreže jer vizualno reprezentiraju govorni signal uz izdvajanje informacijskog sadržaja temeljenog na ljudskoj percepciji. U istraživanjima najčešće se koristi 40 koeficijenta melspektrograma u frekvencijskoj domeni s vremenskim pomakom okvira analize od 10 ms. Znanstveni radovi razlikuju se samo po broju vremenskih okvira analize koje obuhvaća jedan melspektrogram te je u [10] dimenzija podataka 40×26, dok se u [11] koristi jedan vremenski okvir analize s 40 koeficijenta melspektra. Oba istraživanja koriste glasovne podatke iz Googleovih govornih servisa. U [12] koriste se melspektrogrami dimenzija 40×16 izračunati iz *eNTERFACE'05* baze izgovora.

$$f_{mel} = 2595 \cdot \log_{10}\left(1 + \frac{f_{Hz}}{700}\right) \tag{2.1}$$

Treća vrsta spektralnih značajki su melfrekvencijski kepstralni koeficijenti (engl. *mel-frequency cepstral coefficients – MFCC*) koji se izračunavaju iz melspektrograma korištenjem diskretne kosinusne transformacije. Rezultirajući koeficijenti predstavljaju kepstralne koeficijente (slika 2.1c), gdje prvi koeficijent predstavlja energiju signala, a drugi odnos energije na nižim i višim frekvencijama. Ostali koeficijenti nemaju jasnu interpretaciju, ali činjenica je da skup svih kepstralnih koeficijenata nosi informaciju o spektralnom sastavu spektrograma te da može diskriminirati slične zvukove među različitim govornicima [13]. U [14] istraživači su koristili 12 *MFCC* koeficijenata pri klasifikaciji emocija kod 30 rumunjskih govornika dok su *MFCC* kofičijenti u [15] poslužili kao jedan od opisnika niske razine govornog signala (*low-level descriptor – LLD*). Kao značajku za modele s kojima se uspoređuju novi modeli klasifikacije, u [12] su korišteni *MFCC* koeficijenti za model potpornih vektora i rekurentnu duboku neuronsku mrežu (engl. *Long Short-Term Memory Deep Neural Network – LDNN*) gdje se pokazuje da određena količina informacije nestaje kada se melspektar transformira u *MFCC* koeficijente što uzrokuje bolje performanse modela koji kao značajke koriste melspektrograme.



(a) Spektrogram

(b) Mel-spektrogram

(c) MFCC koeficijenti

Slika 2.1: Spektralna i kepstralna reprezentacija riječi *change*

Osim navedenih značajki, pri klasifikaciji emocija pomoću neuronskih mreža moguće je koristiti govorni signal u vremenskoj domeni. Uz pretpostavku minimalnog utjecaja aparature na snimanje određenog izgovora, nad govornim signalom u vremenskoj domeni nije primijenjena transformacija te se očuvala originalna informacija. Potencijalne poteškoće ovog pristupa nalaze se u tome što se slični glasovi mogu pojaviti na različitim faznim pomacima [16]. Radovi koji koriste govorni signal u vremenskoj domeni za klasifikaciju emocija prvenstveno su motivirani ispitivanjem drugačijeg pristupa od tradicionalnog izračunavanja značajki. Tako je u [17] pokazano kako se iz govornog signala pomoću konvolucijske rekurentne neuronske mreže *ConvLSTM-RNN* mogu odrediti značajke koje se konačno koriste za klasifikaciju *SVM* modelom nad bazom emocionalnih izgovora na engleskom jeziku *IEMOCAP*. U [16] uspoređuje se korištenje govornog signala i melspektrograma u procesu prepoznavanja govora koristeći konvolucijsku rekurentnu duboku neuronsku mrežu (engl. *Convolutional Long Short-Term Memory Fully Connected Deep Neural Network – CLDNN*) gdje su rezultati pokazali da model ima jednake performanse u odnosu na dane značajke te da

do poboljšanja dolazi pri korištenju govornog signala i melspektra istovremeno. Prepoznavanje govora vršilo se nad izgovorima na engleskom jeziku prikupljenim kroz Googleove govorne servise. U [18] predstavlja se model koji kombinacijom konvolucijske i rekurentne mreže te korištenjem govornog signala ostvaruje značajno poboljšanje pri klasifikaciji emocija u odnosu na korištenje tradicionalnih značajki i opisnika niske razine (*LLD*) nad *RECOLA* bazom.

U počecima klasifikacije emocija, najčešće su se koristile prozodijske značajke budući da su istraživači bili ograničeni računalnom snagom i poznatim značajkama. Među prvim značajkama koje su bile korištene za proučavanje utjecaja emocija u govoru jesu osnovna frekvencija titranja glasnica, oscilacije osnovne frekvencije te promjene u amplitudi izgovora. Pokazalo se da sve tri značajke imaju značajnu promjenu u ovisnosti o promjeni emocionalnog stanja govornika [19]. Naredna istraživanja počela su koristiti složenije značajke te je napretkom tehnologije skup značajki postao velik i nestandardiziran među istraživačima. Kako bi se pronašao standardizirani skup značajki, 2016. godine znanstvenici su odredili optimalan skup značajki za početak istraživanja i prototipiziranje u području afektivnog računarstva, *GeMAPS* (engl. *The Geneva Minimalistic Acoustic Parameter Set*). Navedeni skup značajki sastoji se od osnovnih i proširenih parametara. Osnovni parametri su frekvencijski (osnovna frekvencija glasa, oscilacije osnovne frekvencije glasa, frekvencije prva tri formanta i spektralna širina prvog formanta), amplitudni (oscilacije amplitude govora, intenzitet govora i odnos energije između zvučnih i bezvučnih komponenata) te spektralni (alfa omjer, Hammarbergov indeks, spektralna ovojnica, energije prva tri formanta, omjer prva dva harmonika osnovne frekvencije glasa i omjer prvog harmonika i najvišeg harmonika). Dodatni parametri su MFCC koeficijenti, spektralni tok te spektralne širine drugog i trećeg formanta [20]. Ovakvi skupovi značajki najčešće se koriste prilikom klasifikacije pomoću određenih statističkih modela ili modela strojnog učenja. Kroz istraživanja najviše se koriste skriveni Markovljevi lanci, Gaussove mješavine, stabla odluke te stroj s potpornim vektorima [3].

Povećavanjem računalne snage u novije vrijeme, metode klasifikacije se počinju temeljiti na dubokom učenju. Modeli dubokog učenja uče rastavljati jedan skup na više podskupova pomoću stvarnih podataka te tako klasificiraju ulazne podatke. Većina modela dubokog učenja temelji se na neuronskim mrežama koje u svojoj osnovnoj formi predstavljaju matematičko pojednostavljenje ljudskog živčanog sustava. Osnovna gradivna jedinica umjetne neuronske mreže je perceptron [21] koji predstavlja model živčane stanice. Sastoji se od produkta n ulaznih vrijednosti $\boldsymbol{x}_i$ i pripadnih težina $\boldsymbol{w}_i$,

funkcije aktivacije g te izlazne vrijednosti $\boldsymbol{y}$ kako je prikazano jednadžbom 2.2.

$$\boldsymbol{y} = g\left[\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{w}_i + \boldsymbol{b}\right] \tag{2.2}$$

Parametar $\boldsymbol{b}$ predstavlja konstantni član perceptrona, a funkcija aktivacije $g[\cdot]$ služi kako bi se konačna vrijednost sume ograničila na određeni raspon vrijednosti definiran svakom funkcijom aktivacije zasebno. Kako bi neuronska mreža mogla izvršavati klasifikaciju, potrebno je težine $\boldsymbol{w}_i$ i konstantni član $\boldsymbol{b}$ postaviti na odgovarajuću vrijednost. Optimalna vrijednost težina pronalazi se optimizacijom modela, odnosno učenjem. Kroz proces optimizacije najčešće se minimizira funkcija razlike između prediktiranih klasa i stvarnih klasa podataka na ulazu modela. Svaki primjerak podataka na ulazu modela ima distribuciju vjerojatnosti po klasama $P(\boldsymbol{x})$, gdje je vjerojatnost nula za klase kojima podatak ne pripada, a jedan za klasu kojoj pripada. Model na svojem izlazu također prikazuje distribuciju vjerojatnosti $Q(\boldsymbol{x})$ gdje po svakoj klasi predviđa koja je vjerojatnost da podatak na ulazu pripada pojedinoj klasi. Kako se na ulazu modela nalazi prava distribucija vjerojatnosti, a na izlazu procijenjena distribucija, razliku između te dvije distribucije moguće je mjeriti unakrsnom entropijom (jednadžba 2.3) čijom minimizacijom možemo optimizirati model [22].

$$H(P, Q) = -E\big[P(\boldsymbol{x})\big] \log\big(Q(\boldsymbol{x})\big) = -\sum_{k} p_k \log(q_k) \tag{2.3}$$

U postupku optimizacije najčešće se koristi operator gradijenta zato što nosi informaciju o promjenama u funkciji razlike $H(P, Q)$. Cilj procesa optimizacije je pronaći globalni minimum funkcije razlike, jer je tamo pogreška između stvarne distribucije i procijenjene distribucije najmanja. Kako se radi o funkciji više varijabli, prati se usmjerena derivacija koja osim iznosa promjene funkcije, može pratiti kako se funkcija mijenja u određenom smjeru gdje je potrebno odrediti smjer promjene koji je orijentiran prema globalnom minimumu. Općenito, osnovni problem ovakvih algoritama optimizacije je u tome što je složenost proporcionalna broju podataka nad kojima se model optimizira (jednadžbe 2.4 i 2.5). Nadogradnja ovakvog pristupa leži u algoritmu optimizacije pomoću stohastičkog gradijentnog spusta koji unosi statističku dimenziju u proceduru određivanja gradijenta [22]. Ako se model optimizira nad skupom podataka $\boldsymbol{x}$ uz pripadajuću distribuciju $\hat{p}$ te ako su $\boldsymbol{\theta}$ skup parametara modela, može se definirati ukupna funkcija razlike koja se sastoji od pojedinačnih funkcija razlika $L(\boldsymbol{x}_i, y_i, \boldsymbol{\theta})$ za svaki primjer podatka iz skupa podataka te pripadajući gradijent:

$$J(\boldsymbol{\theta}) = E\big[L(\boldsymbol{x}, y, \boldsymbol{\theta})\big] = \frac{1}{m} \sum_{i=1}^{m} L(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) \tag{2.4}$$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) \tag{2.5}$$

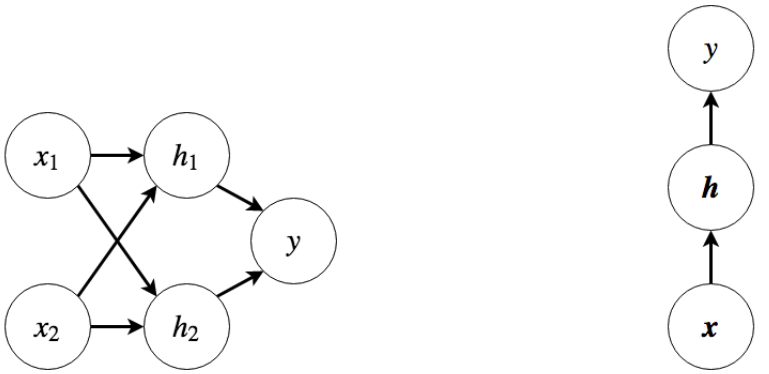
Iz jednadžbe 2.5 vidi se da je ukupni gradijent očekivanje gradijenta za svaki primjer podatka. Očekivanje je moguće procijeniti iz manjeg skupa podataka $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ koji su odabrani iz osnovnog skupa ($n \ll m$), što rezultira time da algoritam učenja modela nije više ovisan ukupnom broju podataka te je gradijent jednak:

$$\boldsymbol{g} = \frac{1}{n} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{n} L(\boldsymbol{x}_i, y_i, \boldsymbol{\theta}) \tag{2.6}$$

Parametri se tada mijenjaju suprotno od smjera gradijenta uz određeni koeficijent proporcionalnosti ε, popularnije poznat kao stupanj učenja modela [22]:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \varepsilon \boldsymbol{g} \tag{2.7}$$

Konkretni modeli klasifikacije koji su korišteni u istraživanju za potrebe ovog rada jesu konvolucijske, rekurentne i duboke neuronske mreže. Potrebno je napomenuti da su navedene vrste neuronskih mreža predstavljale slojeve konačne neuronske mreže koji su se vezali jedan na drugi. Osnovna vrsta neuronskih mreža jest višeslojni perceptron (engl. *Multilayer Perceptron – MLP*), a sastoji od ulaznog sloja perceptrona koji prima podatke, izlaznog sloja perceptrona koji vrši klasifikaciju te jednog sloja između ulaznog i izlaznog sloja koji se naziva skriveni sloj. Dodavanjem dva ili više skrivena sloja, *MLP* postaje duboka neuronska mreža. Na slici 2.2 prikazan je jednostavan primjer višeslojnog perceptrona s ulaznim slojem $\boldsymbol{x}$, jednim skrivenim slojem $\boldsymbol{h}$ i izlaznom vrijednošću y.

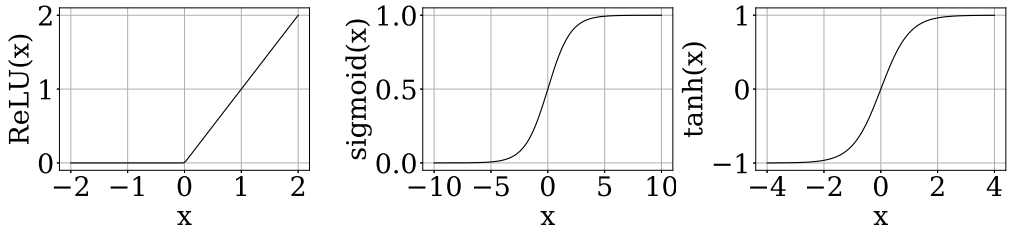


(a) Eksplicitni prikaz perceptrona **(b)** Vektorski prikaz perceptrona

Slika 2.2: Višeslojni perceptron

Duboke neuronske mreže u odnosu na linearne modele (npr. linearna regresija) imaju mogućnost aproksimiranja nelinearnih funkcija tako što prijelazi između neurona nisu čvrsto definirani, nego neuronska mreža samostalno uči kakvi prijelazi

moraju postojati između neurona. Neuroni se organiziraju u slojeve gdje svaki sloj aproksimira jednu funkciju iz kompozicije funkcija pri čemu je cilj mrežom modelirati cijelu kompoziciju koja definira funkcionalnost modela. Iz jednadžbe 2.2 može se vidjeti da neuron sadrži linearan član koji sam za sebe ne može aproksimirati nelinearne funkcije. Svojstvo neuronske mreže za aproksimaciju nelinearnih funkcija nalazi se u funkciji aktivacije $g[\cdot]$ koja linearnu funkciju transformira u nelinearnu. Postoji veliki broj funkcija aktivacije koje se odabiru ovisno o podacima nad kojima se gradi model i namjeni modela. Najpoznatije funkcije su ispravljena linearna funkcija (engl. *Rectified Linear Unit – ReLU*) koja je prikazana na slici 2.3a, tangens hiperbolni (slika 2.3b) te sigmoidalna funkcija (slika 2.3c) [22].



(a) Ispravljena linearna funkcija **(b)** Tangens hiperbolni **(c)** Sigmoidalna funkcija

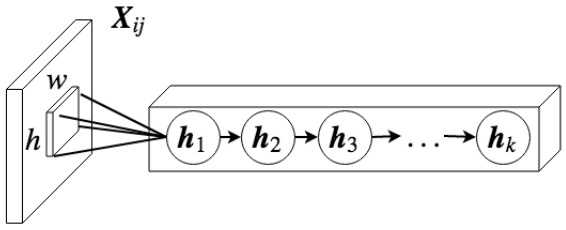
Slika 2.3: Funkcije aktivacije

Konvolucijske neuronske mreže varijanta su neuronskih mreža gdje se barem u jednom sloju umjesto linearnih operatora koristi operacija konvolucije te imaju svojstva interakcije s podacima pomoću rijetkih matrica, dijeljenja parametara i ekvivarijance pri pomaku [22]. Kod dubokih neuronskih mreža dimenzije matrice parametara koje dolaze u interakciju s podacima istih su dimenzija kao podaci. Interakcija pomoću rijetkih matrica ostvaruje se tako što konvolucijske mreže koriste filtar za određivanje značajki iz podataka čije dimenzije mogu biti manje od ulaznih podataka (npr. slike). Na ovaj način smanjuju se potrebe za računalnim resursima koji se koriste pri modeliranju. Dijeljenje parametara omogućuje da se isti parametri ponovno upotrebljavaju u više operacija, odnosno da jedan sloj konvolucijske mreže dijeli isti filtar. Na ovaj način modelu je potrebno manje memorije za pohranjivanje parametra, te se u procesu učenja modela parametri za cijeli sloj mreže mijenjaju samo jednom, a ne za svaki neuron sloja. Konačno, svojstvo ekvivarijance na pomak sugerira mogućnost mreže da bude neosjetljiva na pomak u podacima, to jest da za klasifikaciju nije bitno gdje se informacija nalazi u podacima (primjerice objekt na slici). Ulazni podaci konvolucijskih mreža su najčešće jednodimenzionalni i dvodimenzionalni te se razlika u is-

korištavanju ove dvije vrste podataka nalazi isključivo u dimenzijama konvolucijskog filtra mreže [22]. Konačni produkt konvolucijske funkcije su značajke te se za svaku konvolucijsku funkciju definira K značajki. Tada je funkcija konvolucije $\boldsymbol{F}_{conv}$ nad podacima $\boldsymbol{X} \in \mathbb{R}^{C \times H_0 \times W_0}$ preslikavanje $\boldsymbol{F}_{conv} : \mathbb{R}^{C \times H_0 \times W_0} \longmapsto \mathbb{R}^{K \times H_1 \times W_1}$. Svaka značajka ima dimenzije filtra $h \times w$ te se konačna funkcija konvolucije može zapisati kao:

$$\boldsymbol{F}_{conv,kij} = \boldsymbol{h}_k * \boldsymbol{X}_{ij}^{hw} + \boldsymbol{b}_{kij} = \sum_{q=0}^{C-1} \sum_{r=0}^{h-1} \sum_{p=0}^{w-1} \boldsymbol{X}_{ij}[q, r, p] \boldsymbol{h}_k[q, r, p] + \boldsymbol{b}_{kij} \tag{2.8}$$

U jednadžbi 2.8, $\boldsymbol{X}_{ij}[q, r, p]$ predstavlja podskup ulaznog podatka koji se odabire kao $\boldsymbol{X}[q, i \cdot s_h + r, j \cdot s_w + p]$ pri čemu su s_h, s_w pomaci filtra u vertikalnom i horizontalnom smjeru. Filtar predstavlja član $\boldsymbol{h}_k[q, r, p]$ te konstantni član $\boldsymbol{b}_{kij}$. Jednadžba 2.8 grafički je prikazana na slici 2.4. U okviru klasifikacije emocija konvolucijskim mrežama na temelju spektralnih podataka izgovora postoje četiri vrste filtara $\boldsymbol{h}$ koji se međusobno razlikuju u svojim dimenzijama te na taj način pogoduju određivanju značajki vezanih uz frekvencijsku ili vremensku dimenziju. Prva vrsta je filtar koji pokriva cijelu širinu frekvencijskog područja (engl. *full-spectrum temporally*) te je pogodan za određivanje globalne spektralne informacije kroz nekoliko okvira analize izgovora. Druga vrsta je filtar koji filtrira određeni dio frekvencijske i vremenske dimenzije (engl. *spectral-temporally*) i na taj način određuje lokalne spektralne značajke. Treća vrsta filtra je filtar koji izračunava značajke u vremenskoj dimenziji kroz jednu frekvencijsku komponentu (engl. *temporally-only*). Zadnja vrsta filtra je filtar komplementaran trećoj vrsti, odnosno filtar koji određuje značajke u određenoj širini frekvencijskog pojasa za jedan korak analize (engl. *spectral-only*) [12].



Slika 2.4: Neuron konvolucijske neuronske mreže

Zadnja vrsta neuronskih mreža koje se koriste u istraživanju emocija jesu rekurentne mreže, konkretno implementacija rekurentne mreže s kratkotrajnom memorijom (engl. *Long Short-Term Memory – LSTM*). Specifičnost rekurentnih mreža je ta što neuroni imaju unutarnje stanje koje služi za propagaciju informacija koje su ovisne o vremenu. Unutarnje stanje može se smatrati vrstom memorije pomoću koje neuron rekurentne mreže ima mogućnost pamtiti informacije iz prošlosti te birati koje će informacije zadržati iz sadašnjosti. Kako samo ime nalaže, rekurentni neuron nema jedan ulaz i jedan izlaz, kako je to bio slučaj s do sada opisanim neuronskim mrežama, nego na ulaz dovodi trenutne podatke $\boldsymbol{x}_t$, prethodno stanje neurona $\boldsymbol{c}_{t-1}$ te izlazne značajke izračunate u prethodnom koraku $\boldsymbol{y}_{t-1}$. Ažuriranje stanja i određivanje izlaza neurona vrši se pomoću vrata (engl. *gate*) koja posjeduju matrice težina $\boldsymbol{W}$ za sve pripadne ulaze. Kroz jednadžbe 2.9a - 2.9e prikazane su funkcije pojedinih vrata. Ulazna vrata $\boldsymbol{i}$ u trenutku t ovise o ulaznim podacima $\boldsymbol{x}$, izlaznom stanju $\boldsymbol{c}$ i značajkama $\boldsymbol{y}$ iz trenutka $t-1$ (jednadžba 2.9a). Vrata $\boldsymbol{f}$ koja reguliraju količinu informacije koja propagira iz trenutka $t-1$ u trenutak t služe kao faktor zaboravljanja (engl. *forget gate*) te ovise o istim ulaznim vrijednostima kao i ulazna vrata uz drugačije parametre (jednadžba 2.9b). Stanje neurona ažurira se pomoću $\boldsymbol{c}$ vrata gdje je novo stanje superpozicija prethodnog stanja uz određeni faktor zaboravljanja te trenutnim ulaznim podacima i značajkama iz prethodnog koraka (jednadžba 2.9c). Izlaz rekurentnog neurona $\boldsymbol{y}$ moduliran je izlaznim vratima $\boldsymbol{o}$ te trenutnim stanjem $\boldsymbol{c}$ (jednadžbe 2.9d i 2.9e) [12], [11].

$$\boldsymbol{i}_t = \sigma\Big[\boldsymbol{W}_{\boldsymbol{ix}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{iy}}\boldsymbol{y}_{t-1} + \boldsymbol{W}_{\boldsymbol{ic}}\boldsymbol{c}_{t-1} + \boldsymbol{b}_{\boldsymbol{i}}\Big] \tag{2.9a}$$

$$\boldsymbol{f}_t = \sigma\Big[\boldsymbol{W}_{\boldsymbol{fx}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{fy}}\boldsymbol{y}_{t-1} + \boldsymbol{W}_{\boldsymbol{fc}}\boldsymbol{c}_{t-1} + \boldsymbol{b}_{\boldsymbol{f}}\Big] \tag{2.9b}$$

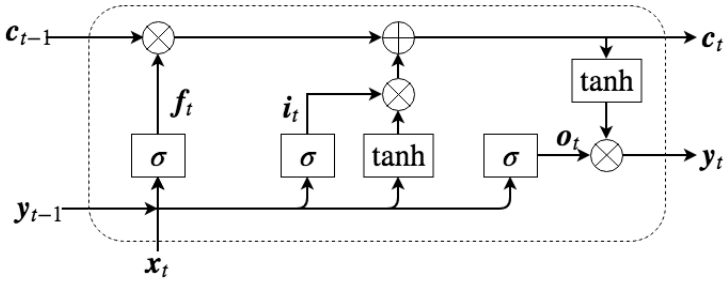
$$\boldsymbol{c}_t = \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot \tanh\Big[\boldsymbol{W}_{\boldsymbol{cx}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{cy}}\boldsymbol{y}_{t-1} + \boldsymbol{b}_{\boldsymbol{c}}\Big] \tag{2.9c}$$

$$\boldsymbol{o}_t = \sigma\Big[\boldsymbol{W}_{\boldsymbol{ox}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{oy}}\boldsymbol{y}_{t-1} + \boldsymbol{W}_{\boldsymbol{oc}}\boldsymbol{c}_t + \boldsymbol{b}_{\boldsymbol{o}}\Big] \tag{2.9d}$$

$$\boldsymbol{y}_t = \boldsymbol{o}_t \cdot \tanh[\boldsymbol{c}_t] \tag{2.9e}$$

Grafički prikaz jednadžbi 2.9a – 2.9e prikazan je na slici 2.5 gdje radi jednostavnosti prikaza nisu istaknute težine $\boldsymbol{W}$ i konstantni članovi $\boldsymbol{b}$.

U vidu klasifikacije emocija, modeliranje pomoću neuronskih mreža mlado je područje istraživanja. Kako bi istraživači postigli što bolje modele, istraživanja su vršili nad modelima koji objedinjuju sve navedene neuronske mreže u jednu mrežu. Tako je u [6] korištena složena neuronska mreža za klasifikaciju sedam osnovnih emocija koja se sastoji od tri konvolucijska sloja i tri sloja perceptrona koji čine duboku neuronsku podmrežu. Autori su za aktivaciju neurona koristili ispravljenu linearnu funkciju. U [10] je predloženo unaprijeđenje modela iz [6] gdje su se između dva konvolucijska



Slika 2.5: Neuron *LSTM* rekurentne mreže

sloja i dva sloja perceptrona dodala dva rekurentna *LSTM* sloja. Takav pristup iskorišten je u [7] uz jedan rekurentni sloj te u [18] s dva rekurentna sloja koji se nastavljaju jedan iza drugoga.

3. Baze emocionalnih izgovora

Kvaliteta modela uvelike ovisi o kvaliteti podataka pomoću kojih je model izgrađen. Kvalitetni podaci dobiveni su obradom izvornih podataka koji odgovaraju potrebama modela, a metode obrade moraju biti takve da ne narušavaju informacijski sadržaj izvornih podataka. Izvorni podaci organiziraju se u baze koje su istraživači dizajnirali za potrebe istraživanja. Korištenje istih podataka među istraživačima olakšava međusobnu usporedbu kvalitete modela, ali i načina strukturiranja podataka za izgradnju modela. Modeli izrađeni na temelju neuronskih mreža zahtijevaju velike količine podatka [22] te je izgradnja jedne baze koja pruža podatke za takve modele kompleksan posao. Ovisno o podacima koji se prikupljaju, veličina baze i trajanje prikupljanja podataka mogu varirati. Ono što sve baze moraju ostvariti je kvalitetan opis podataka pomoću oznaka. Pridjeljivanje oznaka podacima vrši se kroz proces anotiranja podataka čija je složenost također ovisna o vrsti podataka koje se anotira. U domeni govornih podataka anotiranje teži vremenskoj sinkronizaciji oznaka s pojavama u govornom signalu. Za potrebe prepoznavanja riječi, proces anotiranja identičan je transkripciji određenog govora, međutim anotiranje emocija nije trivijalan zadatak jer prvenstveno ovisi o subjektivnom mišljenju anotatora ili predviđenoj emociji koju je pojedini dio eksperimenta u kojem se snimao govor trebao prouzročiti. Ukoliko se označavaju emocije u spontanom govoru, tada je običaj da više anotatora preslušava i kontinuirano anotira svaki govorni signal. Konačna emocija određuje se funkcijom izglednosti gdje se utjecaj anotatora na različite načine uzima u obzir. U slučajevima kada se oznake određuju prema emociji koju je eksperiment trebao isprovocirati, postoji mogućnost da do željene emocije nije došlo, a ako i jest, onda postoji sumnja u to je li određena emocija bila prisutna kroz cijelo vrijeme trajanja izgovora kako to oznaka određuje [15]. Zbog razlika u karakteristikama govora između žena i muškaraca, baze govornih podataka trebaju sadržavati uravnotežen broj ženskih i muških govornika kako bi modeli mogli klasificirati određenu pojavu u glasu neovisno o spolu.

3.1. Poznate baze

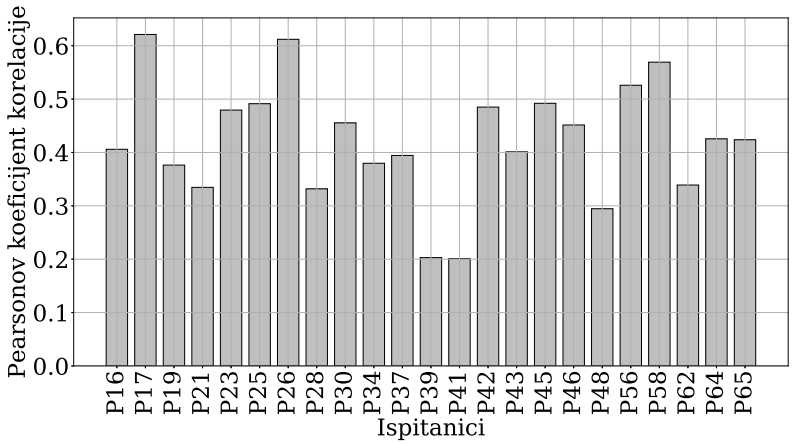
U području emocionalnih izgovora nekoliko je baza koje su najčešće korištene u istraživanjima. Baza izgovora pod stvarnim i induciranim stresom (engl. *A Speech Under Simulated and Actual Stress – SUSAS*) [23] koja objedinjuje govorne podatke pet različitih vrsta stresa. Najzanimljiviji izgovori su oni u kojima je stres spontano induciran naglim promjenama smjera kretanja u zrakoplovu i na vožnjama u zabavnom parku. Baza emocionalnih izgovora na njemačkom jeziku (engl. *A Database of German Emotional Speech – emoDB*) [24] sastoji se od govornih podataka pet ženskih i pet muških glumaca koji su koji su simulirali emocije ljutnje, straha, sreće, tuge, gađenja, dosade i neutralnog stanja. Simuliranje emocija potencijalno može uzrokovati lažnu reprezentaciju govornog signala za pojedine emocije, međutim autori *emoDB* baze problematiku simuliranja emocija opravdavaju time što se emocije inače ne pojavljuju često u svojoj punini te da ih je teško inducirati u kontroliranom eksperimentu. Modernije emocionalne baze kombiniraju govorne podatke i video sekvence. Jedna takva audiovizualna baza podataka je *eNTERFACE'05* [25] koju čine snimke 42 govornika. Govornici su slušali šest priča od kojih je svaka trebala inducirati određenu emociju. Nakon eksperimenta, dva procjenitelja ocjenjivala su uspješnost induciranja emocije pojedinom pričom te ako je predviđeni emocionalni odgovor nedostajao, tada se snimka nije dodala u bazu. Objedinjavanjem audiovizualnih snimki govornika proširenih podacima elektrokardiograma i elektrodermalne aktivnosti, izgrađena je multimodalna baza spontane afektivne interakcije na francuskom jeziku (engl. *Remote Collaborative and Affective interactions – RECOLA*) [4]. Ova baza detaljnije je opisana u sljedećem poglavlju.

3.2. RECOLA

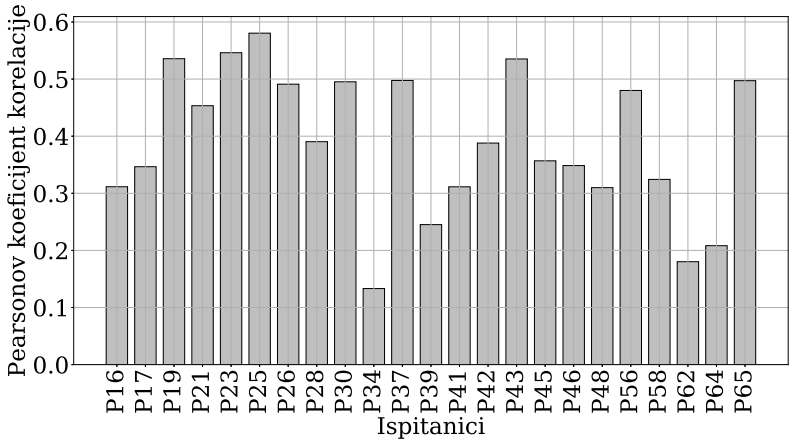
Za potrebe istraživanja u okviru izrade ovog diplomskog rada korišteni su izgovori iz multimodalne baze *RECOLA* (engl. *Remote Coleborative and Affective interactions*) [4]. Baza se sastoji od video i audio snimki te elektrokardiograma i elektrodermalne aktivnosti 23 govornika (12 žena i 11 muškaraca). Govornici su za vrijeme snimanja govorili francuskim jezikom. Emocije su bilježene u kontinuiranom prostoru kojeg čine ugoda (engl. *valence*) i pobuđenost (engl. *arousal*). Emocije se u takvom prostoru nalaze unutar jedinične kružnice postavljene oko ishodišta. Ugoda i pobuđenost poprimaju vrijednosti u intervalu $[-1, 1]$ gdje -1 označava negativnu ugodu i pasivnost, a 1 označava pozitivnu ugodu i aktivnost. Ovakav način određivanja emocija temelji se

na pretpostavci da emocije nisu diskretna psihološka stanja, nego složeni kontinuirani procesi [26].

Označavanje emocija provodilo je šest anotatora koji govore francuski jezik (tri žene i tri muškarca). Svaki anotator zasebno, za svakog govornika, označavao je iznose pobuđenosti i ugode pomoću pomičnih pokazivača frekvencijom 25 Hz koristeći alat ANNEMO. Kao mjera suglasja između anotatora na slici 3.1 i 3.2 prikazani su koeficijenti Pearsonove korelacije koji su usrednjeni po anotatorima za svakog govornika. Može se vidjeti da korelacija nije visoka te najviše iznosi 0.6211 za pobuđenost i 0.5804 za ugodu.



Slika 3.1: Pearsonov koeficijent korelacije anotacija pobuđenosti usrednjen po svim anotatorima, prikazan za svakog govornika



Slika 3.2: Pearsonov koeficijent korelacije anotacija ugode usrednjen po svim anotatorima, prikazan za svakog govornika

Očito je iz prethodnih slika 3.1 i 3.2 da je potrebno postići veći konsenzus između anotatora kako bi procjena emocija bila što točnija. Jedna moguća metoda temelji se na

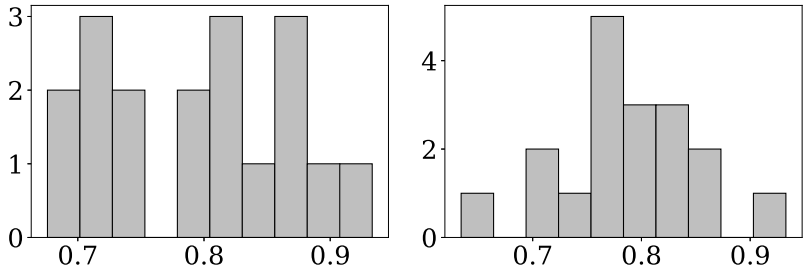
estimaciji korištenjem težina anotatora (engl. *Evaluator Weighted Estimator – EWE*) [27]. Određivanjem prosječnog koeficijenta Pearsonove korelacije anotatora $\bar{r}_i$ dobiva se težina, odnosno postotak koji označava koliko se oznake jednog anotatora slažu s oznakama ostalih anotatora [28]. Tako je za svaki uzorak oznaka x_n moguće odrediti nove oznake x_n^{EWE} temeljene na *EWE* estimaciji za K anotatora [28]:

$$x_n^{EWE} = \frac{1}{\sum_{i=1}^{K} \bar{r}_i} \sum_{i=1}^{K} \bar{r}_i x_n \tag{3.1}$$

Jednadžbom 3.1 može se estimirati jedna vrijednost ugode i pobuđenosti iz K pripadnih vrijednosti koje su dali anotatori uz maksimiziranje suglasja među anotatorima pomoću pripadnih težina. Kako bi se pokazalo da je ovaj postupak točan i opravdan, potrebno je usporediti vrijednosti ugode i pobuđenosti s najboljom i verificiranom reprezentacijom najbolje estimacije (engl. *gold standard*). Za verifikaciju estimacije u ovom radu koristila se najbolja estimacija ponuđena na AVEC 2018 natjecanju u detekciji emocija gdje se koriste govorni podaci iz *RECOLA* baze. Uz najbolju estimaciju dane su originalne vrijednosti ugode i pobuđenosti te je moguće usporediti najbolju estimaciju s *EWE* estimacijom nad originalnim vrijednostima. Usporedba se izvodi pomoću koeficijenta konkordancije korelacije (engl. *concordance correlation coefficient*) koji mjeri suglasnost između dva anotatora x i y te ponovljivost rezultata (jednadžba 3.2) na temelju Pearsonovog korelacijskog koeficijenta ρ, standardnih devijacija σ i srednjih vrijednosti μ oba anotatora [29].

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x + \mu_y)^2} \tag{3.2}$$

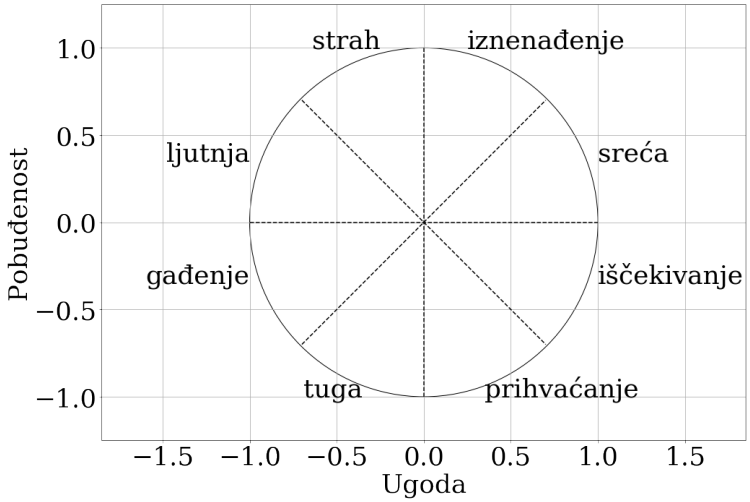
Na slici 3.3 prikazani su histogrami koeficijenata konkordancije korelacije za ugodu i pobuđenost. Može se vidjeti da većina koeficijenata poprima vrijednosti između 0.7 i 0.9 što znači da estimacija *EWE* estimatorom dobro korelira s najboljom estimacijom.



(a) Pobuđenost **(b)** Intrinzična aktivacija

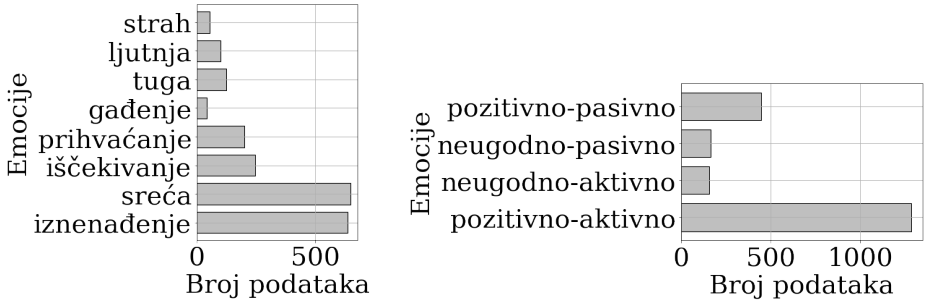
Slika 3.3: Histogram koeficijenta konkordancije korelacije anotacija za pobuđenost i ugodu između anotatora nakon estimacije pobuđenosti i ugode *EWE* estimatorom

Kako bi model mogao klasificirati diskretne emocije potrebno je vremenske signale ugode i pobuđenosti transformirati u diskretne emocije. Prema [26], prostor ugode i pobuđenosti može se podijeliti na osam emocija kako je prikazano slikom 3.4. Neutralno stanje osobe nalazi se na koordinatama $(0, 0)$, međutim u literaturi općenito nije eksplicitno navedeno kolike površine zauzimaju emocije u prostoru ugode i pobuđenosti pa tako nije precizno određeno koliki radijus neutralno stanje zauzima oko ishodišta.



Slika 3.4: Emocije u prostoru intrinzične aktivacije i pobuđenosti

Pomoću tangensa kuta u ishodištu pravokutnog trokuta čije su katete ugoda i pobuđenost moguće je odrediti osam diskretnih emocija sa slike 3.4. Rezultat su oznake emocija za pojedinačne izgovore koje opisuju emocionalno stanje govornika svakih 40 ms. Kada se oznake grupiraju po emocijama u trajanju od 1.2 s, što će se kasnije pokazati da je vremenska dimenzija obrađenih izvornih podataka, primjećuje se da broj podataka s emocijama sreće i iznenađenja dominiraju kod govornika, dok je jako mali broj pojavljivanja emocija poput gađenja i straha (slika 3.5a). Kako bi broj podataka po emocijama bio prikladniji za raspodjelu cijelog skupa podataka na podskupove za učenje modela, validaciju i testiranje, razmatrana je druga podjela prostora ugode i pobuđenosti po kvadrantima gdje emocije iznenađenja i sreće čine pozitivno-aktivnu emociju (1. kvadrant), strah i ljutnja čine negativno-aktivnu emociju (2. kvadrant), gađenje i tuga čine negativno-pasivnu emociju (3. kvadrant), a prihvaćanje i iščekivanje čine pozitivno-pasivnu emociju (4. kvadrant). Raspodjela emocija po kvadrantima (slika 3.5b) s jedne strane onemogućava diskriminaciju emocija unutar jednog kvadranta, ali s druge strane, zbog spajanja srodnih emocija u istu klasu unosi se minimalan šum koji ne narušava proces klasifikacije.



(a) Osam emocija **(b)** Četiri emocije

Slika 3.5: Distribucija podataka po emocijama u prostoru ugode i pobuđenosti

Zadnji korak u obradi izvornih podataka je određivanje značajki izgovora i formiranje konačnog skupa podataka. Govorni signali u bazi otipkani su frekvencijom 44.1 kHz. U svrhu reduciranja veličine podataka u memoriji uz očuvanje govornog informacijskog sadržaja, snimke su pretipkane na frekvenciju uzorkovanja 16 kHz. Značajka koja također smanjuje memorijsku veličinu podataka, a ne narušava informaciju je melspektrogram. Svaki melspektrogram izračunat je na temelju 40 koeficijenata za uskopojasni spektrogram, srednjepojasni spektrogram i širokopojasni spektrogram čije su širine Hammingovog vremenskog otvora analize 64 ms, 32 ms i 16 ms. Svaki spektrogram određen je brzom Fourierovom transformacijom u 4096 točaka, dok je razmak između susjednih okvira analize 10 ms što za sve širine vremenskog otvora daje dovoljan postotak preklapanja susjednih okvira.

Budući da je govor sniman zajedno s videom, postoje određeni dijelovi govorne snimke na kojima govornik nije govorio. Kako bi se olakšalo istraživačima, baza ima anotirane vremenske intervale u kojima je govornik govorio. Dijelovi u kojima su govornici govorili ekstrahirani su iz melspektrograma cijele snimke kako bi se izbjegli spektralni artefakti koji bi nastali Fourierovom transformacijom govornog signala s uklonjenim dijelovima snimke u kojima govornici nisu govorili. Konačan skup podataka sadrži melspektrograme podijeljene na fragmente po emocijama duljine 1.2 s. Vremenska duljina melspektrograma određena je eksperimentalno tako da obuhvaća što veći vremenski raspon uz što veću iskoristivost podataka iz baze s obzirom na to da postoje sekvence u kojima emocije traju kraće od 1.2 s. Konačna iskorištenost podataka iz baze, uz zanemarivanje odbačenih dijelova gdje govornici nisu govorili, iznosi 94%. Budući da je melspektrogram dvodimenzionalan podatak, te da su dimenzije širokopojasnog, srednjepojasnog i uskopojasnog melspektrograma iste, oni su kombinirani u kanale slike što je rezultiralo trodimenzionalnim vektorom $40 \times 120 \times 3$ koji nosi informaciju o globalnim i lokalnim frekvencijskim fluktuacijama izgovora.

4. Klasifikacija emocija

U nastavku ukratko će biti prezentirani rezultati najnovijih istraživanja u ovom području. Točnost klasifikacije emocija modelima dubokog učenja varira u rasponu od 26% do 94%. Širok raspon preciznosti modela rezultat je uporabe različitih arhitektura neuronskih mreža, značajki i metoda testiranja modela.

Korištenjem tri konvolucijska sloja na koje se nadovezuje duboka neuronska mreža s tri sloja perceptrona u [6] pri klasifikaciji straha, ljutnje, sreće, tuge, gađenja, dosade i neutralnog stanja postignuta je prosječna točnost od 56.19%. Najveću individualnu točnost klasifikacije navedenih emocija imali su ljutnja i tuga s 84.21% i 83.08%, slijede ih gađenje i dosada s 68.35% i 53.91%. Neutralno stanje, sreća i strah klasificirani su s 42.11%, 36.36% i 25.33% gdje je model neutralno stanje najviše zamjenjivao s dosadom, a sreću sa strahom i obrnuto. U [17] korištena je kombinacija konvolucijskih i *LSTM* rekurentnih slojeva za određivanje značajki koje su korištene za klasifikaciju ljutnje, sreće, tuge i neutralnog stanja potpornim vektorima i diskriminantnom analizom. Najveća točnost modela postiže se korištenjem *SVM* klasifikatora s polinomnom jezgrenom funkcijom za povećavanje dimenzionalnosti prostora parametara, a iznosi 65.13%. Korištenjem isključivo neuronske mreže, autori su u [7] pomoću modela koji se sastoji od dva konvolucijska sloja i dva *LSTM* rekurentnih slojeva postigli točnost 88.01% klasificirajući sreću, dosadu, tugu, gađenje, strah, ljutnju i neutralno stanje. Još jedno istraživanje koristilo je sličnu arhitekturu kao u prethodnom slučaju. U [18] povezana su dva konvolucijska sloja i dva *LSTM* rekurentna sloja koja čine neuronsku mrežu za predviđanje kontinuiranih signala pobuđenosti i ugode. Autori koriste dvije različite funkcije koje se optimiziraju učenjem modela. Točnost modela iskazana je koeficijentom konkordancije korelacije ρ_c između izvornih i prediktiranih vrijednosti pobuđenosti i ugode. Točnost predviđanja pobuđenosti korištenjem srednje kvadratne pogreške iznosi $0.684\rho_c$, a ugode $0.249\rho_c$ dok je točnost modela minimizacijom konkordancije korelacije stvarne vrijednosti i predviđene vrijednosti pobuđenosti i ugode $0.688\rho_c$ i $0.261\rho_c$.

Jedan od trenutno najboljih rezultata prezentiran je u [12] gdje se neuronska mreža sastoji od dva konvolucijska sloja praćena s jednim *LSTM* rekurentnim slojem na koji se nadovezuju četiri sloja perceptrona, a klasificirala je emocije ljutnje, gađenja, straha, sreće, tuge i iznenađenja. Autori su proučavali kako različite dimenzije filtara konvolucijskih slojeva mreže, opisanih u poglavlju 2, utječu na točnost klasifikacije. Korištenjem filtra koji pokriva cijelu širinu frekvencijskog područja kroz određeni broj koraka analize, točnost modela iznosi 94.58%. Filtrom koji filtrira dio frekvencijske i određeni dio vremenske domene, točnost modela iznosi 93.75%. Model koji filtrima filtrira jednu frekvencijsku komponentu kroz veći broj vremenskih koraka ima točnost 91.67%, a ukoliko se filtrira određeni frekvencijski pojas u jednom vremenskom trenutku, točnost modela iznosi 92.92%.

4.1. Model klasifikacije

U ovom diplomskom radu koristi se model izveden neuronskom mrežom koju sačinjavaju tri konvolucijska sloja, dva *LSTM* rekurentna sloja te tri sloja perceptrona za klasifikaciju četiri kvadranta u prostoru pobuđenosti i ugode (pozitivno-pasivno, neugodno-pasivno, neugodno-aktivno i pozitivno-aktivno) opisana u poglavlju 3.2. Podaci za klasifikaciju imaju dimenzije $40 \times 120 \times 3$ jer jedan podatak čine tri melspektrograma koji su složeni u trećoj dimenziji, a izračunati su nad širokopojasnim, srednjepojasnim i uskopojasnim spektrogramima. Svaki melspektrogram sastoji se od 40 koeficijenata i 120 vremenskih okvira analize. Tri vrste spektrograma korištene su kako bi model imao veću količinu informacija koju obrađuje tijekom učenja i klasifikacije.

Pri modeliranju, podaci se dijele u tri podskupa: podskup za učenje, podskup za validaciju, podskup za testiranje. Podskup podataka za učenje modela sadrži sve podatke nad kojima model optimizira klasifikaciju i namješta parametre te je neophodno da u volumenu ovaj podskup bude najveći. Podaci za validaciju su podaci koji služe za prilagođavanje hiperparametara modela kao što su stupanj učenja, funkcija optimizacije, veličina filtera i broj značajki konvolucijskih slojeva te svi ostali parametri koji definiraju građu modela, a ne funkcionalnost. Podaci za testiranje su podskup svih podataka koje model nikada nije iskoristio za optimizaciju ili validaciju te služe kako bi se potvrdilo da je model naučio klasificirati određene pojave u podacima, a ne podatke kao takve.

Kao što je poželjno da baze podataka imaju izjednačen broj ženskih i muških govornika, tako je za podatke za učenje, verifikaciju i testiranje potrebna ravnoteža između ženskih i muških izgovora zbog toga što je frekvencijski sadržaj govora između

spolova značajno različit. Podjela podataka na opisane podskupove u ovom radu uvjetovana je metodama testiranja modela.

Prva metoda testiranja je u pripadnim podskupovima za učenje, validaciju i testiranje modela sadržavala određeni udio svih izgovora koji čine *RECOLA* bazu tako da su u svakom podskupu zastupljeni izgovori svakog govornika. Druga metoda temeljena je na izostavljanju izgovora jednog govornika iz podskupa za učenje i validaciju (engl. *leave one subject out – LOSO*) gdje izgovori izostavljenog govornika čine podskup za testiranje, a podskupovi za učenje i validaciju formirani su kao u prvoj metodi testiranja. Prva metoda podijele podataka rezultira modelom koji je istovremeno neovisan o govorniku zbog velikog broja govornika nad kojima je model izgrađen, ali je ovisan o toj skupini govornika. Model naučen nad podacima podijeljenim drugom metodom spada u skupinu modela koji su neovisni o govorniku. Konkretno, prvom metodom skup izgovora svih govornika podijeljen je na tri dijela slučajnim uzorkovanjem gdje prvi dio čini podskup za učenje volumena 60% broja svih izgovora, a podskup za validaciju i testiranje pojedinačno imaju 20% broja svih izgovora. Drugom metodom iz skupa svih izgovora izdvojeni su govorni podaci jednog govornika koji čine podskup za testiranje, dok podskup za učenje čini 75% preostalih izgovora, a podskup za validaciju čini preostalih 25% izgovora.

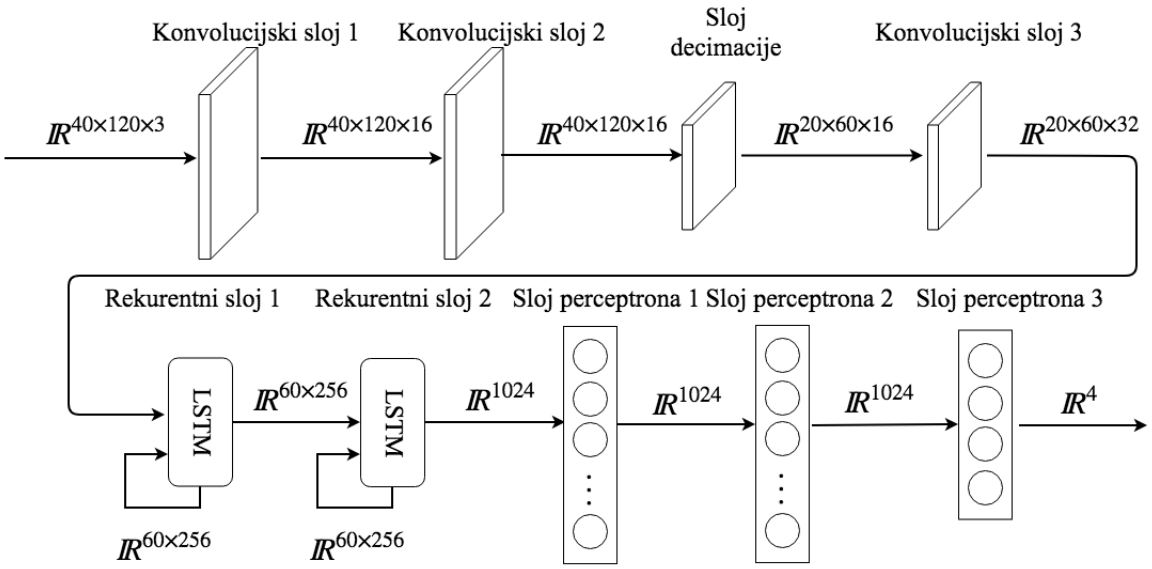
4.1.1. Arhitektura modela

Ulazni podaci prvo se filtriraju konvolucijskim slojevima. Prvi konvolucijski sloj ima filtar dimenzije 7×7 i 16 značajki. Takav filtar ima svojstvo ekstrahiranja značajki kroz 7 koeficijenta melspektra u vremenskom rasponu od 70 ms. Filtar se pomiče za jedan uzorak u svim dimenzijama te se filtracija ponavlja dok cijeli podatak nije filtriran. U drugom konvolucijskom sloju filtri imaju dimenzije 5×5 i 16 značajki te, u odnosu na prvi konvolucijski sloj, ovi filtri ekstrahiraju sitnije detalje i manje promjene. Nakon drugog konvolucijskog filtra slijedi sloj sažimanja (engl. *pooling*) u kojem se podaci decimiraju za faktor 2 u dimenzijama frekvencije i vremena. Decimacija se provodi tako da se za svako područje melspektrograma veličine 2×2 bira maksimalna vrijednost koja reprezentira pojedino područje (engl. *max polling*). Podaci se decimiraju zbog redukcije broja parametara modela kako bi učenje bilo brže. Nakon decimacije slijedi zadnji konvolucijski sloj s filtrom dimenzija 3×3 i 32 značajke koji filtrira usrednjenu informaciju globalnog i lokalnog frekvencijskog sadržaja. Izlazi neurona iz konvolucijskih slojeva propuštaju se kroz ispravljenu linearnu funkciju (*ReLU*) koja pridonosi modeliranju nelinearnih pojava u podacima.

Izračunate značajke prosljeđuju se *LSTM* rekurentnim slojevima koji se nadovezuju jedan na drugi. Budući da melspektrogam nosi vremensku informaciju, rekurentni slojevi u jednom trenutku analiziraju jedan vremenski okvir Fourierove analize dok ostali čine vremenski kontekst. U ovom slučaju će rekurentni slojevi analizirati značajke iz konvolucijske podmreže u 60 vremenskih koraka zbog toga što je početna vremenska dimenzija veličine 120 vremenskih koraka decimirana faktorom 2. Svaki *LSTM* rekurentni sloj složen je od određenog broja ćelija koje su osnovna gradivna jedinica rekurentnih slojeva (slika 2.5), a u ovom modelu rekurentni slojevi posjeduju 256 ćelija gdje je svaka ćelija jedna memorijska jedinica [30]. Dimenzije podataka nakon analize rekurentnim slojevima jesu 60×256 koji sadrže veliku količinu nelinearnosti s obzirom na to da je za izlaznu vrijednost jedne *LSTM* ćelije potrebno izvesti 5 nelinearnih operacija nad ulaznim podacima.

U posljednjem dijelu modela nalazi se višeslojni perceptron koji u prva dva sloja pojedinačno posjeduje 1024 perceptrona, a u zadnjem sloju sadrži 4 perceptrona. Između prvog i drugog sloja perceptrona nalazi se regularizacijski sloj koji sprječava pojavu prenaučenosti modela (engl. *overfitting*). Prenaučenost je pojava kada model podatke za učenje klasificira s točnošću 100% dok podatke koji nisu dio podskupa za učenje, a odabrani su iz iste populacije kao i podaci za učenje, klasificira sa slabijom točnosti (uobičajeno ispod 50%). Do pojave prenaučenosti modela dolazi kada je obujam podataka za učenje premali u odnosu na broj parametara modela te se gubi generalizacija, stoga regularizacijski sloj odbacuje određeni postotak parametara kako bi se izbjeglo prenaučavanje modela [31]. Perceptroni zadnjeg dijela neuronske mreže provode klasifikaciju nad značajkama koje su izračunate u konvolucijskim i rekurentnim slojevima. Zbog uvođenja nelinearnosti u prethodnim slojevima, perceptroni u zadnjem dijelu neuronske mreže nemaju funkciju aktivacije. Drugim riječima, izlazne vrijednosti nisu rezultat nelinearne, nego linearne operacije. Izlazne vrijednosti zadnjeg sloja višeslojnog perceptrona predstavljaju rezultat klasifikacije, odnosno ulazni podaci modela pripadaju klasi koju klasificira perceptron s najvećom izlaznom vrijednosti. Opisana struktura modela prikazana je na slici 4.1 gdje je na prijelazima između slojeva iskazana dimenzionalnost podataka.

Optimizacija modela vršila se *Adam* optimizacijskim algoritmom koji optimizaciju bazira na metodi gradijentnog spusta gdje se prati usrednjeni gradnijent i usrednjeni kvadrat gradijenta koji određuju iznos i smjer promjene parametara modela [32]. Ova metoda superiorna je u odnosu na optimizaciju metodom čistog gradijentnog spusta jer uzima u obzir informaciju o momentu gradijenta, odnosno informaciju o srednjoj trajektorija gradijenta. Optimizatorom se minimizira funkcija razlike između stvar-



Slika 4.1: Struktura implementirane neuronske mreže za klasifikaciju emocija nad govornim podacima

nih klasa podataka i prediktiranih klasa, a u ovom slučaju koristi se funkcija unakrsne entropije. Kako unakrsna entropija zahtjeva dvije razdiobe vjerojatnosti, potrebno je izlazne vrijednosti modela preslikati u prostor vjerojatnosti dok su ulazne vrijednosti klasa prikazane kao distribucija vjerojatnosti. Za preslikavanje koristi se *Softmax* funkcija koja je slična sigmoidalnoj funkciji po tome što preslikava ulazne vrijednosti funkcije u interval $[0, 1]$, a razlikuje se od sigmoidalne funkcije po tome što za višedimenzionalne ulazne vrijednosti, ukupna suma vjerojatnosti za svaku dimenziju zbraja u 1 dok za sigmoidalnu funkciju to ne mora biti slučaj.

Bitan segment modeliranja neuronskim mrežama jesu načini inicijalizacije parametara u pojedinim slojevima koji mogu imati značajan utjecaj na konvergenciju modela optimalnoj klasifikaciji. Ako se težine u slojevima inicijaliziraju istom vrijednosti, tada će optimizacijski algoritam jednako ažurirati težine perceptrona koji imaju isti ulazni parametar čime se gubi mogućnost generalizacije koju kvalitetni modeli posjeduju. Zbog toga, parametri se inicijaliziraju slučajnim vrijednostima uniformne ili Gaussove distribucije [22]. Međutim, u [33] se pokazalo da uniformna distribucija korištena pri inicijalizaciji ima negativne učinke na varijancu gradijenta koji se koristi u proceduri ažuriranja parametra, a negativni učinci su ti da je varijanca ovisna o veličini pojedinih slojeva neuronske mreže. Takvo ponašanje gradijenta može usporiti učenje ili dovesti do toga da slojevi s velikim brojem perceptrona nemaju mogućnost učenja. Kao rješenje autori predlažu normaliziranu inicijalizaciju koja se naziva *Xavier* inicijalizacija

prema jednom od autora. Parametri modela ovog rada inicijalizirani su *Xavier* inicijalizacija uniformnom inicijalizacijom.

Opisani model klasifikacije ostvaren je kroz programski jezik Python 3.6.3 [34] pomoću TensorFlow 1.4.0 [35] biblioteke koja posjeduje širok spektar implementiranih algoritama u domeni strojnog i dubokog učenja. Sklopovlje koje je korišteno za učenje, validaciju i testiranje modela sastoji se od Intel Core i7-6800K procesora s radnim taktom 3.4 GHz, 32 GB radne memorije i nVidia GeForce GTX 1070 grafičke kartice kapaciteta grafičke memorije od 8 GB s taktom procesora 1.683 GHz.

4.2. Rezultati klasifikacije

Kako je najavljeno u prethodnom potpoglavlju, analiza točnosti modela u ovom diplomskom radu vršila se na dva načina. Prvi način obuhvaća modeliranje pomoću cijelog skupa govornih podataka gdje se u podskupovima za učenje, validaciju i testiranje nalaze izgovori svih govornika, dok se drugi način modeliranja temelji na izostavljanju izgovora jednog govornika iz podskupova učenja i validacije, dok podskup testiranja obuhvaća izostavljene izgovore.

4.2.1. Modeliranje preko cijelog skupa govornih podataka

Tijekom procesa učenja model je bio definiran pripadnim parametrima i hiperparametrima od kojih su neki bili promjenjivi, a neki nepromjenjivi. Parametri modela su zbog svoje naravi promjenjivi i oni su se adaptirali kroz proces optimizacije. U domeni hiperparametara, jedini promjenjivi parametar je bio stupanj učenja koji se smanjuje po zakonu polinoma trećeg reda od početne vrijednosti 10^{-3}. Uobičajeno je stupanj učenja postaviti kao promjenjivu vrijednost zbog pretpostavke da model tijekom učenja sve više konvergira globalnom minimumu funkcije razlike ulaznih i prediktiranih klasa, stoga je poželjno da promjene u ažuriranju parametara modela nakon svakog koraka budu manje (jednadžba 2.7). Proces učenja modela trajao je 200 epoha gdje jedna epoha označava period u kojem je cijeli podskup podataka za učenje modela bio obrađen. Na kraju svake epohe, obrađuje se validacijski podskup podataka koji, kako je već navedeno, služi kao pobuda modelu na čiji se odziv ažuriraju hiperparametri. Kako je jedini promjenjivi hiperparametar ovog modela stupanj učenja, koji se mijenja neovisno o točnosti klasifikacije validacijskog podskupa, svrha validacijskog podskupa je praćenje točnosti modela kroz epohe učenja.

Konačni rezultati klasifikacije prikazani su konfuzijskom matricom u tablici 4.1

gdje se u redovima nalazi distribucija klasifikacije izvornih klasa po prediktiranim klasama u stupcima. Odmah je uočljivo da se na dijagonali nalaze najveći postoci što ukazuje da je model u većini slučajeva izvorne podatke klasificirao ispravno. Druga pojava koja je vidljiva jest klasificiranje izvorne klase kao susjedne klase u prostoru pobuđenosti i ugode. Primjerice kod pozitivno-aktivne emocije određeni podaci klasificirani su kao negativno-aktivne emocije, a neki kao pozitivno-pasivne emocije. Ova pojava je očekivana s obzirom na to da model nema apsolutnu točnost te je frekvencijski sadržaj susjednih emocija generalno sličan. Treća pojava je klasificiranje izvornih podataka kao dijametralno suprotne klase u prostoru pobuđenosti i ugode gdje je najbolji primjer negativno-pasivna emocija koja je u 10% slučajeva klasificirana kao pozitivno-aktivna emocija. Uzrok zadnjoj pojavi može biti nedovoljna uravnoteženost skupa izgovora po diskretnim klasama emocija te model zbog toga ne može kvalitetno generalizirati emocije koje opisuje puno manje podataka u odnosu na druge emocije u skupu (slika 3.5b). Konačna točnost ovog modela određena kao omjer ispravno klasificiranih podataka i broja svih podataka iznosi 94.75%

Tablica 4.1: Konfuzijska matrica prosječne točnosti po klasama klasifikacije

Prediktirano / Izvorno	negativno-pasivno	negativno-aktivno	pozitivni-aktivno	pozitivni-pasivno
negativno-pasivno	80.00%	5.00%	10.00%	5.00%
negativno-aktivno	0	83.87%	6.45%	9.68%
pozitivni-aktivno	1.18%	1.96%	94.90%	1.96%
pozitivni-pasivno	4.26%	6.38%	6.38%	82.98%

4.2.2. Modeliranje izostavljanjem govornika

Metodom modeliranja izostavljanjem govornika (*LOSO*) cilj je utvrditi performanse modela u slučajevima kada se model testira svim izgovorima jednog govornika koji su u potpunosti izostavljeni iz podataka za izgradnju modela. Za svakog izostavljenog govornika izgrađuje se zaseban model, a svi modeli temeljeni su na istoj arhitekturi s istim hiperparametrima. Testiranje točnosti modela u ovom procesu vrši se unakrsnim testiranjem u N koraka (engl. N*-fold cross validation*) tako da se za svaki izgrađeni model utvrđuje njegova točnost gdje je N broj nezavisnih modela koji se testiraju. U ovom radu provedeno je unakrsno testiranje u 23 koraka pri čemu je u svakom koraku izgrađen novi model koji u procesu učenja i validacije nije analizirao izgovore izostav-

ljenog govornika. Svi modeli izgrađeni su na isti način kako je to opisano u poglavlju 4.1.1 gdje je početni stupanj učenja postavljen na vrijednost 10^{-3} koja se postupno smanjivala učenjem modela kao što je u slučaju modeliranja nad cijelim skupom izgovora (potpoglavlje 4.2.1).

Zastupljenost emocija u podskupovima za testiranje varira između govornika. Kod pojedinih govornika pojavljuju se emocije iz dva kvadranta u prostoru pobuđenosti i ugode, dok je kod govornika koji su pokazali emocije iz sva četiri kvadranta udio emocija iz drugog ili trećeg kvadranta bio zanemarivo malen u odnosu na broj emocija iz prvog i četvrtog kvadranta. Zbog slabe balansiranosti podskupa za testiranje, u svim koracima unakrsnog testiranja uspješnost klasifikacije mjerila se dodatnim veličinama: osjetljivost (engl. *recall*), preciznost (engl. *precision*) i F1 mjera (engl. *F1 score*) koja se izračunava kao harmonijska sredina osjetljivosti i preciznosti.

U svakom koraku unakrsnog testiranja prikazana je točnost, osjetljivost, preciznost i F1 mjera. Točnost modela izračunata je kao omjer ispravno klasificiranih podataka u odnosu na ukupan broj podataka. Za određivanje osjetljivosti i preciznosti koriste se iznosi podataka koji su klasificirani kao ispravno pozitivni, ispravno negativni, lažno pozitivni i lažno negativni. U grupu ispravno pozitivnih i ispravno negativnih podataka spadaju svi podaci koje model klasificira kao klasu kojoj originalno podaci pripadaju. Podaci jedne klase koji su klasificirani kao da pripadaju drugim klasama čine grupu lažno negativnih, a podaci drugih klasa koji su klasificirani kao jedna određena klasa spadaju u grupu lažno pozitivnih podataka. Osjetljivost se definira kao omjer ispravno pozitivnih podataka i sume ispravno pozitivnih i lažno negativnih podataka, dok je preciznost definirana omjerom ispravno pozitivnih podataka i sume ispravno pozitivnih i lažno pozitivnih podatka. Osjetljivost, preciznost i F1 mjera određuju se za svaku klasu posebno, stoga je u svrhu praćenja metrike modela korištena usrednjena vrijednost po klasama za sve tri veličine.

U tablici 4.5 prikazani su rezultati unakrsnog testiranja u 23 koraka. Kao što je vidljivo iz rezultata, može se reći da je složenost modela prevelika u odnosu na količinu podataka pomoću kojih se model izgrađuje što rezultira slabom generalizacijom zbog čega model uz karakteristike govora u određenom emocionalnom stanju govornika, uči specifičnosti glasa pojedinih govornika. Dodatno, postoji mogućnost da su distribucije glasovnih podataka pojedinih govornika za pojedine emocije drugačije od ostalih govornika iz uzorka što dodatno otežava prepoznavanje emocija za te govornike. Navedenim pretpostavkama ide u prilog činjenica da izostavljanjem nekih govornika, pojedini modeli u fazi testiranja imaju visoku točnost kao što je slučaj u jedanaestom koraku (78.33%), ali su srednja osjetljivost i preciznost neproporcionalno

niske što ukazuje da model ispravno klasificira jednu klasu visokom točnošću, dok je udio ispravne klasifikacije ostalih klasa manji. Prethodno navedeno vidljivo je u tablici 4.2 gdje je prikazana konfuzijska matrica 11. koraka unakrsnog testiranja u kojoj je vidljiva raspodjela klasifikacije izvornih klasa koje se nalaze u redovima. Uspješan primjer klasifikacije vidi se u šestom koraku gdje točnost iznosi 96.67%, a osjetljivost, preciznost i F1 mjera ne zaostaju (tablica 4.3). Dodatno, iz navedenih primjera može se vidjeti da model najbolje nauči klasificirati najbrojnije klase: pozitivno-aktivna i pozitivno-pasivna. Tu činjenicu potvrđuje visoka točnost pri testiranju izgovora 6. govornika u kojima se isključivo pojavljuju emocije iz najbrojnijih klasa i činjenica da su emocije 11. govornika većinom klasificirane kao najbrojnije emocije.

Između izgrađenih modela u ovom potpoglavlju vidljive su razlike između klasifikacije emocija kod ženskih i muških ispitanika (tablica 4.4). Uzrok ovoj pojavi može biti način provođenja Fourierove analize nad muškim i ženskim izgovorima. Naime, najmanji okvir analize u vremenskoj domeni obuhvaća vremensku širinu od 16 ms. Kod ženskih glasova, prosječnih frekvencija titranja glasnica 200 Hz, takav vremenski okvir obuhvaća tri puna perioda govornog signala te su zbog toga u frekvencijskoj domeni jasno vidljive promjene osnovne frekvencije glasa. U slučaju muških glasova, isti vremenski otvor obuhvaća jednu ili manje od jedne periode govornog signala što u frekvencijskoj domeni uzrokuje nedostatak informacije o osnovnoj frekvenciji titranja glasnica te je izgled formanata u spektrogramu i melspektrogramu drugačiji u odnosu na ženske izgovore. Točnije, melspektrogrami ženskih izgovora sadrže vidljive harmonike dok su harmonici u melspektrogramima muških glasova distorzirani ili nedostaju. Ovisnost točnosti modela o pojavljivanju harmoničke strukture u melspektrogramima može se utvrditi izračunavanjem osnovne frekvencije titranja glasnica svih govornika te ako točnost opada proprocionalno s frekvencijom titranja glasnica, tada je utvrđena ovisnost točnosti i harmoničke strukture u melspektrogramima.

Tablica 4.2: Konfuzijska matrica prosječne točnosti po klasama klasifikacije za 11 izostavljenog govornika

Prediktirano / Izvorno	negativno-pasivno	negativno-aktivno	pozitivni-aktivno	pozitivni-pasivno
negativno-pasivno	0	0	50.00%	50.00%
negativno-aktivno	0	33.33%	33.33%	33.33%
pozitivni-aktivno	42.86%	0	57.14%	0
pozitivni-pasivno	2.08%	2.08%	8.33%	87.50%

Tablica 4.3: Konfuzijska matrica prosječne točnosti po klasama klasifikacije za šestog izostavljenog govornika

Prediktirano / Izvorno	negativno-pasivno	negativno-aktivno	pozitivni-aktivno	pozitivni-pasivno
negativno-pasivno	0	0	0	0
negativno-aktivno	0	0	0	0
pozitivni-aktivno	0	0	100.00%	0
pozitivni-pasivno	0	0	17.65%	82.35%

Tablica 4.4: Prosječna točnost, osjetljivost, preciznost i F1 mjera svih modela izgrađenih modeliranjem izostavljanjem ispitanika

Spol	Točnost	Osjetljivost	Preciznost	F1 mjera
muški i ženski	37.55%	31.04%	33.63%	26.86%
muški	26.48%	22.60%	25.99%	18.56%
ženski	47.70%	38.78%	40.63%	34.47%

Modele izgrađene modeliranjem izostavljanjem govornika moguće je objediniti u ansambl u kojem svaki od modela zasebno vrši klasifikaciju emocija te se na temelju rezultata svakog modela odabire klasa u koju se ulazni podatak svrstava [36]. Na ovaj način formira se jedan model iz više njih za koji se može pretpostaviti da će klasifikaciju izgovora iz baze ili klasifikaciju novih izgovora vršiti s većim postotkom točnosti od pojedinačnih modela zbog različitih vrijednosti parametara pojedinih modela koji rezultiraju različitim načinima generalizacije, čijom kombinacijom se postiže točnija klasifikacija emocija.

Tablica 4.5: Performanse modela u procesu unakrsnog testiranja kroz 23 koraka (M - muški, Ž - ženski)

Izostavljeni govornik	Točnost	Osjetljivost	Preciznost	F1 mjera	Spol
1	36.67%	28.88%	39.66%	32.27%	M
2	43.33%	53.88%	34.17%	33.95%	M
3	72.22%	39.11%	33.70%	34.99%	Ž
4	31.11%	41.45%	32.20%	20.46%	Ž
5	32.50%	22.84%	22.83%	15.43%	Ž
6	96.67%	91.18%	98.03%	94.15%	Ž
7	10.00%	28.57%	5.61%	9.38%	M
8	25.00%	31.25%	11.90%	17.24%	Ž
9	41.67%	33.55%	57.05%	36.99%	Ž
10	30.00%	19.27%	16.40%	16.42%	M
11	78.33%	44.49%	46.36%	44.59%	Ž
12	25.00%	35.07%	35.78%	26.21%	M
13	61.33%	21.75%	32.62%	26.10%	M
14	53.33%	29.53%	28.65%	29.00%	Ž
15	28.33%	22.50%	27.18%	24.62%	Ž
16	50.00%	50.00%	54.83%	46.38%	Ž
17	44.44%	47.09%	61.69%	38.22%	Ž
18	18.89%	12.31%	12.16%	11.44%	Ž
19	13.33%	18.27%	45.71%	14.82%	M
20	12.22%	8.70%	12.19%	8.26%	M
21	5.00%	7.50%	27.00%	7.40%	M
22	28.89%	14.72%	15.22%	13.98%	M
23	25.56%	12.05%	21.54%	15.45%	M

5. Zaključak

Primjenjivanje neuronskih mreža za klasifikaciju emocija u govoru mlado je istraživačko područje koje proučava modele koji samostalno određuju optimalne značajke za proces klasifikacije. Modeliranje neuronskim mrežama unaprijedila je upotreba slojeva konvolucijskih i rekurentnih neuronskih mreža. Konvolucijski slojevi imaju svojstvo filtriranja informacija i ekstrahiranja značajki koje nisu ili su vremenski kratko ovisne gdje se karakteristike filtra određuju tijekom učenja modela dok rekurentni slojevi ekstrahiraju informaciju koja je vremenski ovisna, kao što je slučaj kod govornog signala.

Emocije se mogu određivati kroz diskretna stanja poput sedam osnovnih emocija ili kao kontinuirane vrijednosti pobuđenosti i ugode. Ova dva načina nisu međusobno isključiva, već se diskretne emocije nalaze u prostoru definiranim pobuđenosti i ugodom. Određivanje emocija u pojedinom izgovoru subjektivan je zadatak nekoliko nezavisnih procjenitelja. Kako bi se osigurala nepristrana procjena emocije u određenom trenutku, potrebno je ostvariti suglasje između procjenitelja pomoću estimatora. Pokazalo se da je za estimaciju emocija na temelju podataka više procjenitelja pogodan težinski estimator koji pomoću težina određuje doprinose pojedinih procjenitelja u konačnoj estimaciji.

U ovom radu ispitane su performanse modela kojeg čini neuronska mreža sastavljena od tri konvolucijska sloja, dva *LSTM* rekurentna sloja i tri sloja perceptrona. Prikazani su rezultati testiranja modela izgrađenog nad cijelim skupom izgovora i rezultati unakrsnog testiranja izostavljanjem izgovora jednog govornika iz podskupa za učenje modela. Vidljivo je da u oba slučaja modeli imaju prostora za napredak koji bi se trebao ostvariti kroz smanjivanje broja parametara modela, povećanjem broja podataka za izgradnju modela kako bi model imao mogućnost naučiti glasovne značajke emocija neovisne o jednom ili više govornika te dodatnom predobradom podataka za učenje, validaciju i testiranje modela. Također, poboljšanje performansi modela može se postići preciznijim definiranjem neutralnog stanja govornika u prostoru pobuđenosti i ugode koje je trenutno u literaturi definirano kao ishodište tog prostora, a ne kao određeni dio tog prostora.

LITERATURA

[1] M. Argyle, "Non-verbal Communication and Language," *Royal Institute of Philosophy Lectures*, vol. 10, no. 1976, pp. 63–78, 1976. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0080443600011079

[2] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - State-of-the-art and the challenge," *Computer Speech and Language*, vol. 27, no. 1, pp. 4–39, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.02.005

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013.

[5] K. R. Scherer and M. R. Zentner, "Emotional Effects of Music: Production Eules." *Music and emotion: Theory and research*, pp. 361–392, 2001.

[6] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," *International Conference on Platform Technology and Service*, pp. 1–5, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/7883728/

[7] W. Lim, D. Jang, and T. Lee, "Speech Emotion Recognition using Convolutional and Recurrent Neural Networks," *Asia-Pacific Signal and Information Processing Association*, pp. 1–4, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7820699/

[8] N. Weiskirchen, R. Bock, and A. Wendemuth, "Recognition of Emotional Speech with Convolutional Neural Networks by Means of Spectral Estimates," *Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, no. October, pp. 50–55, 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8272585/

[9] S. S. Stevens, J. Volkmann, and E. B. Newman, "WO different concepts of pitch have," *The Journal of the Acoustical Society of America*, vol. 8, pp. 14–19, 1937.

[10] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, Long Short-Term Memory, Fully Connected Deep Neural Networks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4580–4584, 2015.

[11] F. Beaufays, H. Sak, and A. Senior, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling Has," *Interspeech*, no. September, pp. 338–342, 2014.

[12] C.-W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," *IEEE Transactions*, pp. 1–20, 2017. [Online]. Available: http://arxiv.org/abs/1706.02901

[13] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008.

[14] E. Fran I, I. Ispas, V. Dragomir, M. Dasc, E. Alu, Zoltan, and I. C. Stoica, "Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots," *Romanian Journal of Information Science and Technology*, vol. 20, no. 3, pp. 222–240, 2017. [Online]. Available: http://www.romjist.ro/full-texts/paper562.pdf

[15] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," *Interspeech*, pp. 1537–1540, 2015.

[16] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the Speech Front-end With Raw Waveform CLDNNs," *International Speech Communication Association*, vol. 2015-Janua, pp. 1–5, 2015.

[17] N. Kurpukdee, T. Koriyama, and T. Kobayashi, "Speech Emotion Recognition using Convolutional Long Short-Term Memory Neural Network and Support

Vector Machines," *Asia-Pacific Signal and Information Processing Association*, pp. 1744–1749, 2017.

[18] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, S. Zafeiriou, and T. U. München, "Adieu Features ? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5200–5204, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7472669/

[19] J. A. Bachorowski and M. J. Owren, "Vocal Expression of Emotion: Acoustic Properties of Speech Are Associated With Emotional Intensity and Context," *Psychological Science*, vol. 6, no. 4, pp. 219–224, 1995.

[20] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[21] F. Rosenblatt, "The Perceptron : A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[22] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[23] J. H. Hansen and S. E. Bou-Ghazale, "Getting Started With A Speech Under Simula ted and A ctual Stress D t a abase John," *Eurospeech*, pp. 1–4, 1997.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Interspeech*, pp. 1517—-1520, 2005. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.8506{\&}rep=rep1{\&}type=pdf

[25] O. Martin, I. Kotsia, B. Macq, I. Pitas, and P. Levant, "The eNTERFACE ' 05 Audio-Visual Emotion Database," *Proceedings of the 22nd International Conference on Data Engineering Workshops*, no. 1, pp. 2–9, 2006.

[26] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[27] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," *IEEE Automatic Speech Recognition and Understanding Workshop*, vol. 2005, pp. 361–365, 2005.

[28] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-p. Thiran, T. Ebrahimi, and D. Lalanne, "Prediction of asynchronous dimensional emotion ratings from," *Pattern Recognition Letters*, pp. 1–10, 2014.

[29] L. I.-k. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility Author (s): Lawrence I-Kuei Lin Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2532051 REFERENCES Linked references are available on JSTOR for thi," *Biomatrics*, vol. 45, no. 1, pp. 255–268, 1989.

[30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/neco.1997.9.8.1735

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[32] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, pp. 1–15, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proceedings of 13th International Conference on Artificial Inteligence and Statistics*, vol. 9, pp. 249–256, 2010.

[34] "Python 3.6.3," https://www.python.org/downloads/release/python-363/, accessed: 2018-03-15.

[35] "Tensorflow 1.4.0," https://www.tensorflow.org/, accessed: 2018-03-15.

[36] L. K. Hansen and P. Salamon, "Neural Network Ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

Klasifikacija emocija konvolucijskim neuronskim mrežama pomoću glasovnih podataka

Sažetak

Računala su postala neizostavan dio svakodnevice, međutim komunikacija između korisnika i računala nije prilagođena uobičajenom načinu na koji ljudi komuniciraju. Afektivno računarstvo ima cilj osposobiti računala kako bi mogla prepoznavati emocije te prilagođavati odgovor na korisničke naredbe ovisno o emocionalnom stanju korisnika. U tom procesu potrebno je odrediti emocije na neinvazivan način te se kao jedna moguća metoda pokazala detekcija emocija iz govora. U ovom radu proučava se klasifikacija emocija pomoću neuronskih mreža koje se sastoje od konvolucijskih i rekurentnih slojeva nad glasovnim podacima *RECOLA* baze izgovora.

Ključne riječi: glas, govor, emocije, klasifikacija emocija, konvolucijske neuronske mreže, rekurentne neuronske mreže, neuronske mreže

Convolutional Neural Networks for Emotion Recognition Tasks Using Voice/Speech Data

Abstract

In recent times computers can be found everywhere. People use them to execute certain commands but interaction between human–computer is nothing like communication between people. The goal of affective computing is to develop methods and algorithms for computers in order to estimate user's emotional state and adjust the response accordingly. One of the noninvasive paths of achieving this goal is an emotion classification from speech which is described in this master's thesis. This was achieved by neural network consisted of convolutional and recurrent layers for classification task using *RECOLA* speech database.

Keywords: speech, emotions, emotion classification, convolutional neural network, recurrent neural network, neural network