Prepoznavanje govornika korištenjem dubokih neuronskih mreža

Markić, Luka

SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Sveučilišni studij

PREPOZNAVANJE GOVORNIKA KORIŠTENJEM DUBOKIH NEURONSKIH MREŽA

Diplomski rad

Luka Markić

Osijek, 2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK

Obrazac D1: Obrazac za ocjenu diplomskog rada na sveučilišnom diplomskom studiju

Ocjena diplomskog rada na sveučilišnom diplomskom studiju

Ime i prezime pristupnika:	Luka Markić
Studij, smjer:	Sveučilišni diplomski studij Računarstvo
Mat. br. pristupnika, god.	D1307R, 07.10.2022.
JMBAG:	0165084193
Mentor:	izv. prof. dr. sc. Ratko Grbić
Sumentor:	
Sumentor iz tvrtke:	
Predsjednik Povjerenstva:	prof. dr. sc. Robert Cupec
Član Povjerenstva 1:	izv. prof. dr. sc. Ratko Grbić
Član Povjerenstva 2:	doc. dr. sc. Petra Pejić
Naslov diplomskog rada:	Prepoznavanje govornika korištenjem dubokih neuronskih mreža
Znanstvena grana diplomskog rada:	**Umjetna inteligencija (zn. polje računarstvo)**
Zadatak diplomskog rada:	Prepoznavanje osobe putem govora (engl. speaker recognition) ima širok spektar primjena u različitim industrijama i sektorima, npr. za kontrolu pristupa i u biometrijskim aplikacijama, u forenzici i sl. Ovdje postoje dva temeljna zadatka: identifikacija govornika i verifikacija govornika. Osnovni cilj identifikacije govornika je razlikovanje identiteta pojedinca iz skupine poznatih govornika. Verifikacija govornika je pak proces koji uključuje potvrdu identiteta govornika putem njihova govora tj. provjerava je li govornik ono što tvrdi da jest usporedbom njihova glasa s
Datum ocjene pismenog dijela diplomskog rada od strane mentora:	18.09.2024.
Ocjena pismenog dijela diplomskog rada od strane mentora:	Izvrstan (5)
Datum obrane diplomskog rada:	27.09.2024.
Ocjena usmenog dijela diplomskog rada (obrane):	Izvrstan (5)
Ukupna ocjena diplomskog rada:	Izvrstan (5)
Datum potvrde mentora o predaji konačne verzije diplomskog rada čime je pristupnik završio sveučilišni diplomski studij:	01.10.2024.

FERIT
FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA **OSIJEK**

IZJAVA O IZVORNOSTI RADA

Osijek, 01.10.2024.

Ime i prezime Pristupnika:	Luka Markić
Studij:	Sveučilišni diplomski studij Računarstvo
Mat. br. Pristupnika, godina upisa:	D1307R, 07.10.2022.
Turnitin podudaranje [%]:	2

Ovom izjavom izjavljujem da je rad pod nazivom**: Prepoznavanje govornika korištenjem dubokih neuronskih mreža**

izrađen pod vodstvom mentora izv. prof. dr. sc. Ratko Grbić

i sumentora

moj vlastiti rad i prema mom najboljem znanju ne sadrži prethodno objavljene ili neobjavljene pisane materijale drugih osoba, osim onih koji su izričito priznati navođenjem literature i drugih izvora informacija.
Izjavljujem da je intelektualni sadržaj navedenog rada proizvod mog vlastitog rada, osim u onom dijelu za koji mi je bila potrebna pomoć mentora, sumentora i drugih osoba, a što je izričito navedeno u radu.

Potpis pristupnika:

SADRŽAJ

1. UVOD

Razvojem tehnologije broj dostupnih usluga se povećava, no istovremeno se povećava sigurnosni rizik neovlaštenog pristupa tim uslugama i korisničkim podacima. U svrhu povećanja sigurnosti, koristi se autentikacija, čiji je cilj potvrditi identitet osobe i omogućiti ovlašteni pristup uređajima ili uslugama putem mreže, kao što su mobilni uređaji ili internet bankarstvo. Osim za pristup elektroničkim uslugama i uređajima, koristi se i za pristup određenim objektima, kao što je trezor banke ili stambena zgrada. Provjera se provodi na temelju jednog ili više autentikacijskih faktora, a to su: ono što znamo (primjerice, PIN ili zaporka), ono što posjedujemo (primjerice, sigurnosni ključ ili RFID kartica) i ono što nas određuje (primjerice, biometrijske značajke). Dva su temeljna zadatka autentikacije: identifikacija i verifikacija korisnika, gdje se identifikacija odnosi na prepoznavanje identiteta pojedinca iz skupine poznatih osoba, odgovarajući na pitanje „Tko je?“. S druge strane, verifikacija provjerava odgovara li identitet osobe pretpostavljenom identitetu, odgovarajući na pitanje „Je li ovo doista ta osoba?“ [1, 2]. Autentikacija koja koristi biometrijske značajke osobe u novije vrijeme ostvaruje široku primjenu [3].

Biometrija se odnosi na granu znanosti koja na temelju fizičkih i ponašajnih karakteristika vrši prepoznavanje i potvrdu identiteta osobe [4]. Fizičke karakteristike odnose se na karakteristike ljudskog tijela poput: geometrije ruke, otiska prsta, mrežnice, lica, rasporeda krvnih žila i DNK uzorka, dok se ponašajne karakteristike odnose na specifične obrasce ponašanja poput: govora, dinamike tipkanja, putanje dobivene pomicanjem računalnog miša, glasa, specifičnog pokreta i druge [5]. Kao prihvatljiva biometrijska karakteristika odabire se ona karakteristika koju ima svaka osoba, jedinstvena je za svaku osobu, čija promjena nije značajna tijekom vremena, pri čemu postoji mogućnost mjerenja karakteristike (primjerice, snimanje otiska prsta) što je prihvaćeno od strane korisnika [6].

Ako se zadatak identifikacije i verifikacije temelji na biometrijskoj karakteristici glasa, radi se o prepoznavanju govornika (engl. *speaker recognition*). Cilj identifikacije je pronaći govornika iz skupine poznatih govornika na temelju karakteristika glasa. S druge strane, proces verifikacije može biti ostvaren na dva načina. Prvi je onaj u kojemu proces verifikacije ovisi o sadržaju kojeg izgovara govornik, što može biti neki oblik zaporke. U suprotnom, proces ne ovisi o izgovorenom sadržaju, već se samo oslanja na karakteristike glasa. U području prepoznavanja govornika dugo vremena su bile korištene klasične metode poput univerzalnog pozadinskog modela temeljenom na Gaussovom modelu mješavine (engl. *Universal Background Model - Gaussian Mixture Model* - UBM-GMM) i skrivenih Markovljevih modela (engl. *Hidden Markov models* - HMM) [7, 8].

Međutim, danas se primjenjuju moderni algoritmi razvijeni u području dubokog učenja, od kojih se ističu konvolucijska neuronska mreža (engl. *Convolutional Neural Network* - CNN) i sijamska neuronska mreža (engl. *Siamese Neural Network* - SNN). CNN je mreža temeljena na operaciji konvolucije pogodna za obradu ulaznih podataka koji imaju prostornu strukture slike. Zbog načina rada sijamska neuronska mreža koja u osnovnoj izvedbi uči na sličnostima ulaza, prikladna je za proces prepoznavanja govornika usporedbom sličnosti uzoraka govornika. U području prepoznavanja govornika ulaz u mrežu je spektrogram koji daje frekvencijske karakteristike zvučnog signala u vremenu. Cilj spektrograma je vizualno predstaviti zvučni zapis govornika u svrhu izvlačenja karakterističnih informacija tog govornika.

Kroz ovaj diplomski rad razvijeno je i testirano vlastito rješenje za prepoznavanje govornika pristupom verifikacije govornika koja uzima u obzir sadržaj izgovorene fraze. Prvotno se vrši priprema i obrada GRID skupa podataka koji sadrži izgovore različitih fraza od različitih govornika u obliku zvučnih zapisa [9]. Zatim se stvaraju odgovarajući parovi spektrograma koji predstavljaju ulaz sijamske neuronske mreže. Skup podataka se dijeli na dva načina. Prvi se odnosi na podjelu u kojoj se zvučni zapisi istog govornika mogu pojaviti u trening, validacijskom i testnom skupu. Kod druge podjele, svi zvučni zapisi određenog govornika se pojavljuju u samo jednom skupu. S ciljem povećanja robusnosti sijamske neuronske mreže i smanjenja pretjeranog usklađivanja na trening skupu podataka (engl. *overfitting*) umjetno se dodaje Gaussov šum zvučnim zapisima i primjenjuje amplitudna augmentacija na dobivenim spektrogramima. Na temelju tih izmjena stvaraju se nove mreže i uspoređuje se utjecaj augmentacije na rad mreže. Konačan cilj rada je dobiti mrežu koja je sposobna raditi u stvarnom okruženju. Vodeći se time obavljeno je proširenje skupa podataka vlastitim skupom zvučnih zapisa. Na temelju tog skupa podataka procesom prijenosnog učenja i korištenjem mreža dobivenih treniranjem na GRID skupu podataka stvaraju se nove mreže. Naposljetku, izradom aplikacije testira se rad istreniranih mreža u stvarnom okruženju.

U nastavku rada, u drugom poglavlju dan je pregled pojmova vezanih uz prepoznavanje govornika korištenjem dubokih neuronskih mreža, kao i pregled radova vezanih uz područje identifikacije i verifikacije govornika. Nadalje, treće poglavlje odnosi se na treniranje mreže za verifikaciju govornika ovisne o izgovorenom sadržaju korištenjem sijamske neuronske mreže, obradu skupa podataka, proširenje skupa podataka vlastitim skupom i izradu aplikacije. U četvrtom poglavlju evaluiran je rad istrenirane mreže na GRID i vlastitom skupu podataka, te je testiran rad aplikacije u stvarnom okruženju. U konačnici, u petom poglavlju iznesen je zaključak cjelokupnog rada.

2. PREGLED PODRUČJA PREPOZNAVANJA GOVORNIKA KORIŠTENJEM DUBOKIH NEURONSKIH MREŽA

Ovim poglavljem dan je pregled osnovnih pojmova vezanih uz područje prepoznavanja govornika korištenjem dubokih neuronskih mreža. Osim pregleda pojmova konvolucijske neuronske mreže i sijamske neuronske mreže, dan je kratak opis i razlog primjene spektrograma kao vizualne reprezentacije zvučnog signala koji je u području prepoznavanja govornika često ulaz u navedene mreže. Međutim, glavni fokus je analiza ključnih zadataka prepoznavanja govornika, a to su identifikacija i verifikacija govornika kroz pregled radova u području strojnog učenja [7].

2.1. Pregled pojmova u području prepoznavanja govornika korištenjem dubokih neuronskih mreža

U nastavku se razmatraju pojmovi strojnog učenja vezani uz prepoznavanje govornika, uključujući konvolucijsku i sijamsku neuronsku mrežu. Osim tih mreža, dan je uvid u pojam spektrograma, koji predstavlja vizualnu reprezentaciju zvučnog zapisa i koristi se kao ulaz navedenih mreža.

2.1.1. Konvolucijska neuronska mreža

Konvolucijska neuronska mreža pripada kategoriji neuronskih mreža koja obrađuje podatke koji imaju prostornu strukturu s rasporedom podataka u stupcima i retcima, što odgovara strukturi matrice ili slike. Tip je duboke neuronske mreže kod koje se podaci obrađuju sekvencijalno, drugim riječima, smjer podataka ide od ulaza prema izlazu bez povratne veze. Mreža je inspirirana radom vizualnog korteksa koji sadrži mali broj ćelija iznimno osjetljivih na određene vizualne podražaje [10]. Navedeno je spoznato eksperimentom izvedenim 1962. godine od strane neurofiziologa David H. Hubela i Torsten Wiesela, koji su proveli eksperiment na primarnom vidnom korteksu anesteziranе mačke [11]. Mački su na zaslonu bile prikazane linije te je uočena aktivacija određenih neurona prikazom linija pod različitim kutovima. Tim je zaključeno kako vizualni sustav mačke gradi sliku iz jednostavnijih podražaja u složenu reprezentaciju. Stoga se duboke neuronske mreže mogu promatrati kao kaskadni modeli stanica inspirirani opažanjima Hubela i Wiesela, što možemo predstaviti sustavom slojeva neurona mreže [10, 11]. Standardna konvolucijska neuronska mreža sastoji se od konvolucijskog sloja, sloja sažimanja i potpuno povezanog sloja.

Konvolucijski sloj predstavlja važnu ulogu u radu mreže. Rad konvolucijskog sloja zasniva se na matematičkoj operaciji konvolucije koja se unutar konvolucijskog sloja izvodi nad dvjema matricama. Jedna matrica naziva se ulazna matrica ili ulaz, a druga filtar ili jezgra, a operacijom

konvolucije dobiva se izlazna matrica [12]. Filtar je obično malih dimenzija, ali procesom konvolucije prolazi kroz sve ulazne podatke za određeni iznos koraka (engl. *stride*). Često se koristi veći broja filtara. Razlog primjene većeg broja filtara je izvlačenje karakterističnih informacija, drugim riječima značajki, gdje svaki pojedini filtar služi izvlačenju određene značajke ulaza [13]. Dimenzije izlaza ovise o dimenzijama ulaza, veličini filtra, iznosu koraka i drugim parametrima. Broj kanala izlaza jednak je broju filtara, dok je iznos visine i širine izlaza moguće dobiti iz sljedećeg izraza:

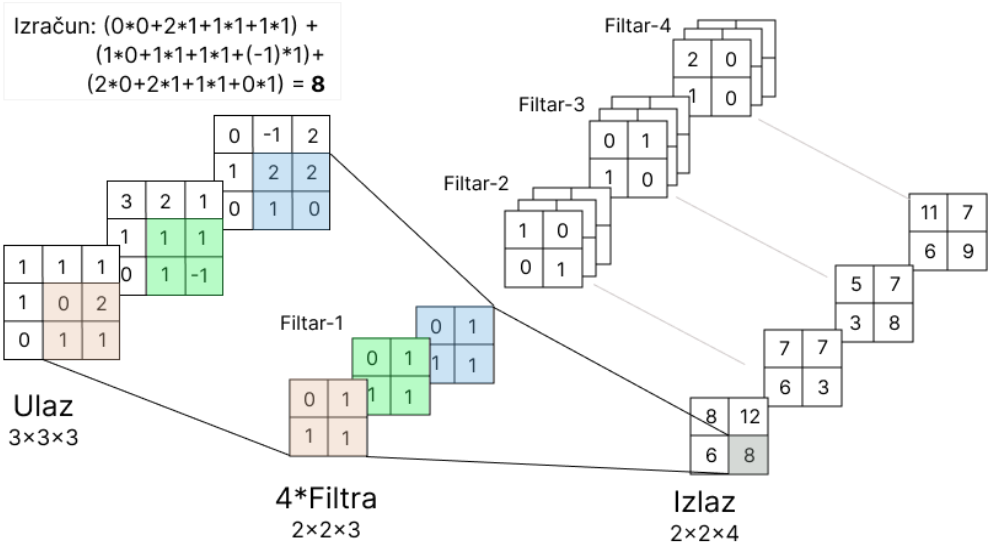
$$n_{izlaza} = \left|\frac{n_{ulaza} - k - 2 * p}{s}\right| + 1. \tag{2-1}$$

Prema izrazu (2-1), parametar n_{izlaza} odnosi se na visinu ili širinu izlaza, a parametar n_{ulaza} na visinu ili širinu ulaza. Nadalje, parametar k predstavlja iznos visine ili širine filtra (jezgre), parametar s iznos koraka za koji se filtar pomiče kroz ulaz i parametar p iznos ispunjenja slike (engl. *padding*), a ako je vrijednost parametra p jednaka nuli, onda ispunjenje slike ne postoji [14].

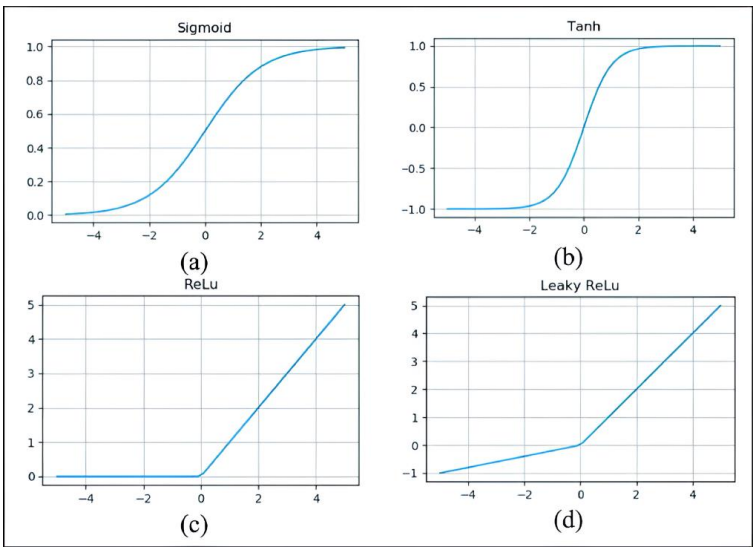
Vizualizacija procesa konvolucije u trodimenzionalnom prostoru, odnosno konvolucije čiji ulaz ima više kanala i veći broj filtara prikazana je slikom 2.1. Dimenzija ulaza je 3x3x3, prikazana su 4 filtra s dimenzijama 2x2x3 koji djeluju po cijeloj dubini ulaznog volumena i pomiču se za jedan korak. U konačnici, to rezultira izlazom s dimenzijom 2x2x4. Za razumijevanje rada konvolucijskog sloja u trodimenzionalnom prostoru, prvotno je potrebno razmotriti postupak u dvodimenzionalnom prostoru. U dvodimenzionalnom prostoru, postupak se izvodi na način da se filtar pomiče po ulaznoj matrici za odgovarajući iznos koraka. Pri svakom pomicanju prozora odabire se manja matrica iz ulazne matrice koja ima iste dimenzije kao filtar, a čiji elementi se preklapaju s elementima filtra. Takva matrica naziva se prozor ulaza. Elementi na jednakim indeksima u prozoru ulaza i filtra se množe, a zbroj tih umnožaka predstavlja rezultat konvolucije za dani prozor. Upravo ta vrijednost je vrijednost pripadajućeg prozora izlazne matrice. U trodimenzionalnom prostoru, postupak je proširen na način da svaki filtar vrši konvoluciju po cijeloj dubini ulaznog volumena. Drugim riječima, određeni filtar vrši konvoluciju po cijeloj dubini ulaza na način da kanali ulaza i filtra s istim indeksima vrše konvoluciju. Primjerice, prvi kanal filtra s prvim kanalom ulaza. U konačnici se matrice dobivene procesom konvolucije svakog kanala ulaza i filtra zbroje u jednu matricu. Zbroj matrica odrađen je na način da se zbroje elementi na istim mjestima u matrici. Time se dobiva matrica koja predstavlja jedan od kanala izlaza, odnosno kanal izlaza koji je vezan uz filtar. Drugim riječima, ako je konvolucija ulaza izvedena s prvim filtrom onda se radi o prvom kanalu izlaza. Slikom 2.1. prikazan je izračun postupka konvolucije prvog filtra po dubini ulaznog volumena za donji lijevi prozor. Prvo se po cijeloj

dubini ulaza primjeni konvolucija s prvim filtrom, nakon čega se vrijednosti rezultata konvolucije donjeg lijevog prozora pojedinih kanala zbroje i ta se vrijednost upisuje u donji lijevi prozor prvog kanala izlaza [13, 15].

Uz konvolucijski sloj vežu se aktivacijske funkcije. Aktivacijske funkcije igraju ključnu ulogu u uspjehu učenja dubokih neuronskih mreža. Primjenjuju se na izlaz svakog neurona prethodnog sloja u mreži (najčešće konvolucijski ili potpuno povezani sloj), uzimajući ponderirani zbroj ulaza i daje izlaz koji se zatim prosljeđuje sljedećem sloju [16]. Ove funkcije mogu biti linearne ili nelinearne, ovisno o potrebama mreže i podacima za učenje. Nelinearne funkcije dodaju nelinearnost mreži, a neke od poznatijih nelinearnih funkcija: Rectified Linear Unit (ReLU), Sigmoid, Tanh, Leaky ReLU. Navedene aktivacijske funkcije prikazane su slikom 2.2. [17, 18].

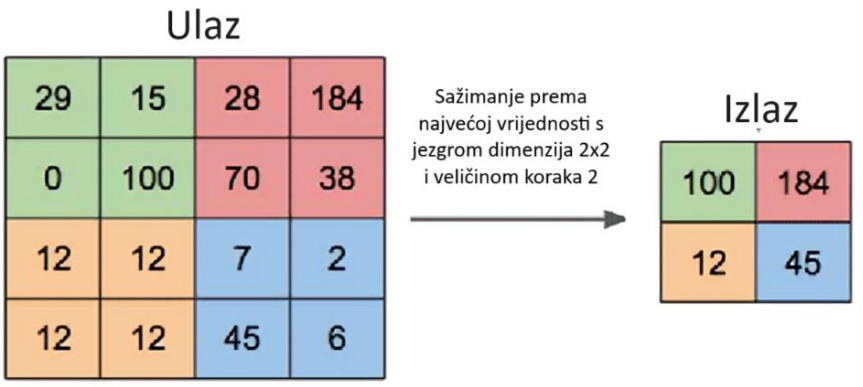


Slika 2.1. *Demonstracija rada konvolucije nad višekanalnim ulazom s većim brojem višekanalnih filtara*



Slika 2.2. *Prikaz nelinearnih aktivacijskih funkcija: a) Sigmoid, b) Tanh, c) ReLu i d) Leaky ReLu* [19]

Sloj sažimanja (engl. *pooling layer*) je sloj koji nastoji smanjiti dimenzije istovremenim očuvanjem važnih podataka ulaza. Pomaže kontroliranju kompleksnosti mreže, smanjenju pretjeranog usklađivanja na trening podatke i ubrzava treniranje mreže smanjenjem broja parametra. Neke od metoda sažimanja su: izračunavanje prosječne (engl. *Average Pooling*), najveće (engl. *Max Pooling*) i ponderirane prosječne vrijednosti (engl. *Weighted Average Pooling*). Jedna od popularnijih metoda je sažimanje prema najvećoj vrijednosti. Demonstracija rada sažimanja prema najvećoj vrijednosti za matricu dimenzija 4x4 prikazana je slikom 2.3., gdje se ulaz dijeli na prozore pomicanjem jezgre sloja sažimanja za iznos koraka 2. Dimenzija prozora jednaka je dimenziji jezgre sloja sažimanja, a ona iznosi 2x2. Nadalje, u svakom prozoru odabire se najveća vrijednost koja će predstavljati vrijednost tog prozora u novoj slici. U konvolucijskoj mreži ulaz je većinom višekanalni. Kod višekanalnog ulaza sloj sažimanja se primjenjuje na svaki kanal ulaznog volumena zasebno. Primjerice, za ulaz dimenzija 80x100x3 primjenom sloja sažimanja na svakom kanalu zasebno s pomicanjem jezgre sloja sažimanja dimenzija 2x2 za iznos koraka 2 rezultira izlazom dimenzija 40x50x3 [20].



Slika 2.3. *Demonstracija rada sloja sažimanja prema najvećoj vrijednosti* [20]

Potpuno povezani sloj je sloj čija je svrha prepoznati globalne uzorke i odnose u ulaznim podacima povezivanjem svakog neurona iz prethodnog sloja sa svakim neuronom u potpuno povezanom sloju. Obično se postavlja na kraju konvolucijske neuronske mreže s primarnom funkcijom donošenja odluka na temelju značajki izvučenih u prethodnim slojevima. Ovo se postiže učenjem složenih odnosa i zavisnosti između ulaznih i izlaznih podataka. Osim za izvlačenje značajki, u kombinaciji s nekom od tehnika kao što je nasumično izbacivanje neurona, omogućuje poboljšani rad mreže u smislu smanjenja usklađivanja na trening skupu podataka [21].

Sloj s nasumičnim izbacivanjem neurona (engl. *droput layer*) koristi se u dubokim neuronskim mrežama kako bi se spriječilo pretjerano usklađivanje na trening podatke. Navedeno može dovesti do loše generalizacije na novim podacima. Rad se zasniva na nasumičnom postavljanju ulaznih

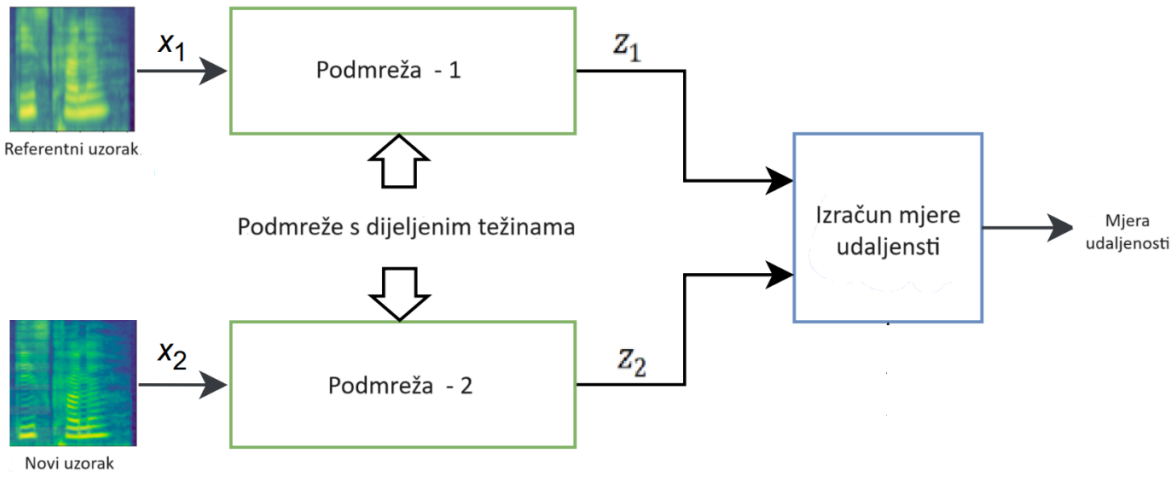
neurona na 0, odnosno isključivanju neurona. Broj isključenih neurona ovisi o stopi isključenja, koja se kreće od 0 do 1. Vrijednost stope isključenja odgovara udjelu isključenih neurona prethodnog sloja. Ako je ta vrijednost 0, ne dolazi do isključenja neurona, a ako je vrijednost 1 isključuju se svi neuroni prethodnog sloja [22].

U zadacima klasifikacije moguće je koristiti Softmax aktivacijsku funkciju u izlaznom sloju mreže. Ona transformira skup vrijednosti iz posljednjeg skrivenog sloja ili izlaznog sloja u distribuciju vjerojatnosti za pojedine klase. Softmax transformacija osigurava da su dobivene vjerojatnosti nenegativne i da se je njihov zbroj jednak 1. Dakle, Softmax sloj proizvodi distribuciju vjerojatnosti za *k* različitih klasa, omogućujući mreži donošenje odluke kojoj klasi ulaz pripada [15, 23].

2.1.2. Sijamska neuronska mreža

Sijamska neuronska mreža je tip neuronske mreže čiji rad je zasnovan na mjerenju sličnosti, odnosno različitosti parova uzorka koji se predaju kao ulaz mreže. Primjenu ostvaruju upravo u područjima gdje je važna usporedba sličnosti uzoraka, a to su: prepoznavanje lica osobe, izgovora, potpisa i slično [24]. Kao prvi primjer uporabe navodi se sijamska neuronska mreža vezana uz rješavanje problema prepoznavanja potpisa kojeg su 1990. godine predstavili Bromley i LeCun [25]. Arhitektura tipične sijamske neuronske mreže prikazana je na slici 2.4. Mreža se sastoji od dvije identične podmreže, koje dijele iste težine. Ulaz u sijamsku neuronsku mrežu su parovi uzoraka, na slici 2.4. označenih s x_1 i x_2, gdje prvi uzorak predstavlja referentni uzorak s kojim se uspoređuje novi uzorak. U većini literature referentni uzorak naziva se uzorak sidra, a u području prepoznavanja govornika koristi se i naziv evaluacija. Drugi uzorak para odnosi se na novi uzorak za kojeg se određuje mjera sličnosti s referentnim uzorkom. U području prepoznavanja govornika takav uzorak često se naziva uzorak upisa ili validacije. Nadalje, svaka podmreža obrađuje podatke nekog od uzoraka s ciljem dohvaćanja karakterističnih podataka o uzorku. Nakon obrade podataka ulaza svaka podmreža kao izlaz daje vektor značajki pojedinog uzorka koji su na slici 2.4. označeni sa z_1 i z_2. Na temelju tih vektora računa se mjera udaljenosti, odnosno mjera koja označava različitost vektora. Ako je udaljenost mala, odnosno bliža nuli, može se reći da se radi o sličnim uzorcima i mreža uzorke klasificira kao istu klasu, odnosno kao pozitivan par. Ako se pretpostavi da mreža vrši predikciju sličnosti dva uzorka, u navedenom slučaju vjerojatnost predikcije mreže treba biti što bliža iznosu 100%. U suprotnom, ako je mjera udaljenost velika, mreža uzorke klasificira kao različite klase, tj. kao negativan par. U tom slučaju vjerojatnost predikcije mreže treba biti što bliža nuli. Za izračun mjere udaljenosti najčešće se koriste Manhattan udaljenost (L1) i Euklidska udaljenost (L2) [8, 24, 26].

U području prepoznavanja govornika sijamska neuronska mreža zahtijeva manje uzoraka za treniranje mreže kako bi se zadržala visoka točnost, za razliku od pristupa konvolucijske neuronske mreže za klasifikaciju govornika, koja za postizanje visoke točnosti mora biti trenirana na dovoljno velikom skupu podataka. Prednost sijamske neuronske mreže iskazana je i u smislu skalabilnosti skupa podataka jer omogućuje dodavanje uzorka novog govornika ili promjenu uzorka postojećeg govornika bez potrebe za ponovnim treniranjem mreže. Za navedeni slučaj konvolucijska neuronska mreža koja služi za klasifikaciju govornika zahtjeva ponovno treniranje mreže što je računalno i vremenski zahtjevno [8]. Treniranje sijamske neuronske mreže na malom skupu podataka može biti ostvareno korištenjem *one-shot* mehanizma. To je način treniranja mreže koji se primjenjuje u klasifikacijskim zadacima u kojima se mreža trenira na skupu gdje je dovoljan samo jedan uzorak iste klase. Primjenjuje se u područjima gdje nije moguće ili je jako teško ostvariti veliki skup podataka. Sustavi kojih ih primjenjuju ostvaruju dobre rezultate u usporedbi sličnosti uzoraka, ali nisu prikladni za primjenu na druge probleme [25, 27].

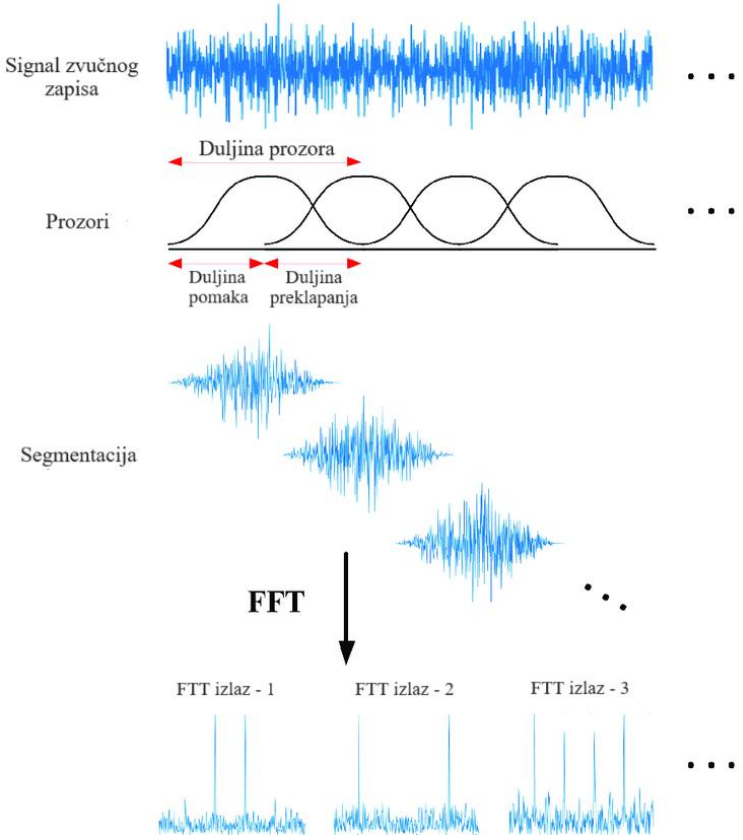


Slika 2.4. *Prikaz arhitekture tipične sijamske neuronske mreže* [8]

2.1.3. Spektrogram

Spektrogram predstavlja vizualizaciju zvučnog zapisa s prikazom frekvencijskih karakteristika u vremenu. Spektrogram se stvara primjenom kratkotrajne Fourierove transformacije (engl. *short-time Fourier transform* - STFT) na zvučnom zapisu [28]. STFT radi razdvajanjem ulaznog signala na više dijelova na kojima se naknadno primjenjuje brza Fourierova transformacija (engl. *fast Fourier transform* - FFT). Razdvajanje se vrši na način da se slijedno izdvajaju određeni segmenti ulaznog signala koji odgovaraju određenom vremenskom intervalu signala, pri čemu se ti segmenti nazivaju prozorima. Svaki prozor jednake je duljine, a odabir vremenskog intervala iz kojeg se izdvaja signal radi se mehanizmom pomicanja prozora. Prozor se pomiče kroz ulazni signal u pozitivnom smjeru za iznos vrijednosti pomaka (engl. *hop length*), a pri svakom pomaku odabire

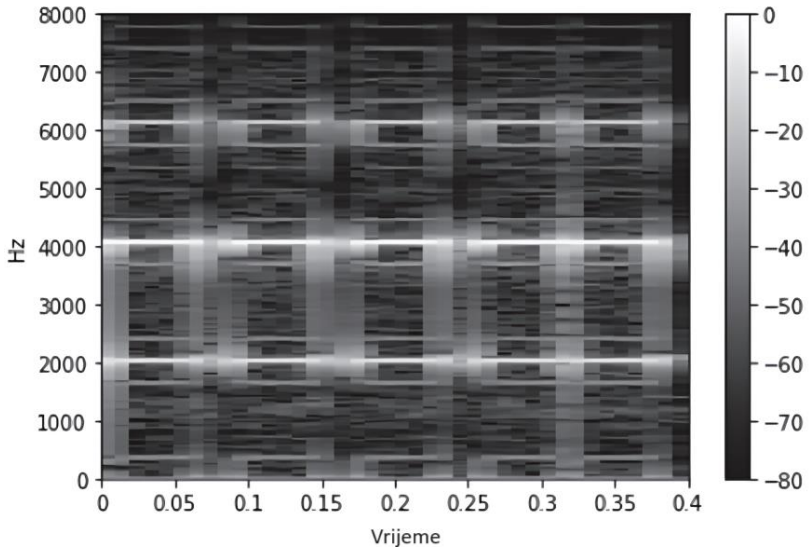
se novi vremenski interval, tj. izdvaja se novi segment ulaznog signala. Ako je duljina pomaka manja od duljine prozora dolazi do preklapanja prozora, odnosno dva susjedna prozora dijele isti dio signala, a čija je duljina određena duljinom preklapanja (engl. *overlap length*). Na izdvojene dijelove signala primjenjuje se funkcija prozora (engl. *window function*) u svrhu smanjenja nekonzistentnih promjena signala koje nastaju na granicama svakog prozora prilikom segmentacije. Nadalje, na te prozore primjenjuje se FFT u svrhu dobivanja frekvencijske karakteristike pojedinog prozora. Budući da se STFT može izvesti tijekom pomicanja prozora dobiva se frekvencijska karakteristika u vremenu [29]. Na taj način je moguće sačuvati informacije o vremenu i amplitudi za odgovarajuću frekvenciju. U konačnici, kombiniranjem svih segmenata na koje je primijenjen FFT dobiva se spektrogram zvučnog zapisa. Kada govorimo o vizualnoj reprezentaciji spektrograma, x-os spektrograma predstavlja vremensku os, dok je y-os frekvencijska, a boje ili nijanse predstavljaju razinu intenziteta (amplitudu) svake frekvencije u određenom vremenskom trenutku. Važno je istaknuti kada se govori o amplitudi spektrograma, radi se o intenzitetu frekvencije, a ne jačini zvučnog signala [28]. Slikom 2.5. prikazan je princip rada STFT na signalu zvučnog zapisa.



Slika 2.5. *Prikaz principa rada STFT na signalu zvučnog zapisa* [29]

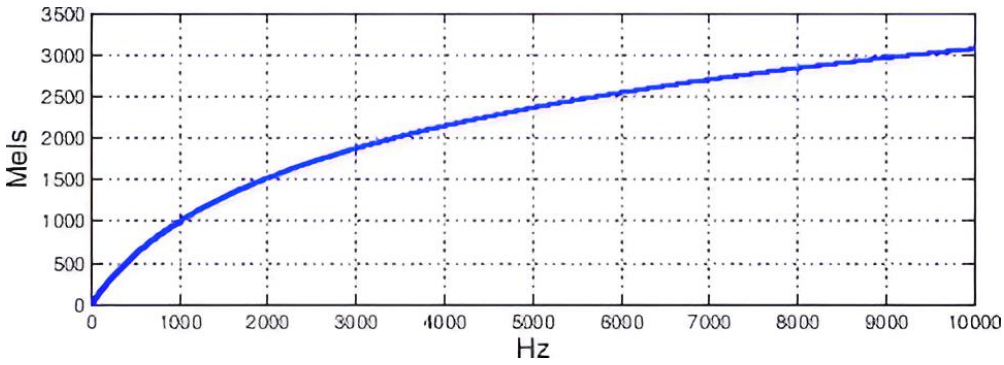
Slikom 2.6. prikazan je primjer spektrograma dobiven primjenom STFT. Iz slike je moguće vidjeti nijanse amplituda frekvencije u danom vremenu. Amplitude su iskazane u decibelima, gdje one

koje imaju vrijednost bližu 0 predstavljaju veće amplitude i one su obojene svjetlijom nijansom, dok one koje se imaju nižu amplitudu poprimaju tamnije nijanse [28].



Slika 2.6. *Prikaz izgleda spektrograma dobivenog primjenom STFT* [28]

Spektrogram koji je prethodno objašnjen predstavlja najjednostavniju izvedbu spektrograma koja se naziva STFT spektrogram. Takav spektrogram izveden je iz ravnomjerno raspoređenih frekvencija, zbog čega nije prikladan za rad vezan uz govor. Prikladnija je uporaba Mel spektrograma [15]. Mel spektrogram je vrsta spektrograma gdje je frekvencijska os pretvorena u Mel ljestvicu. Mel ljestvica je nelinearna ljestvica koja odgovara percepciji ljudskog uha na visinu tona pri različitim frekvencijama. Nelinearnost proizlazi iz činjenice da ljudsko uho ima različitu percepciju tona pri različitim razinama frekvencije [30]. Slikom 2.7. prikazana je Mel krivulja [31]. U području obrade zvuka povezanog s govorom, prisutan je još jedan tip spektrograma - logaritamski Mel spektrogram. Razlog uporabe je smanjenje raspona amplitudnih vrijednosti spektrograma uz očuvanje važnih informacija. Ovo pomaže boljem prepoznavanju karakteristika zvuka u području niskih i visokih frekvencija [28, 32].



Slika 2.7. *Prikaz promjene ljudske percepcije tona s obzirom na promjenu frekvencije* [31]

2.2. Identifikacija govornika korištenjem dubokih neuronskih mreža

Identifikacija govornika je zadatak određivanja govornika iz skupine poznatih govornika na temelju karakteristika govornika kao što su: izgovor, naglasak, ritam, visina glasa i slično. Uz pretpostavku da je x^t testni uzorak, $\{x_k^e | k = 1,2, \dots, K\}$ skup podataka spremljenih uzoraka i K broj osoba čiji je uzorak unesen u skup podataka. Identifikacija se vrši određivanjem uzorka iz skupa koji ima najveću sličnost danom testnom uzorku funkcijom sličnosti f prema parametrima w:

$$k^* = argmax \{f(x_1^e, x^t; w), f(x_2^e, x^t; w), \dots, f(x_K^e, x^t; w)\} \quad (2\text{-}2)$$

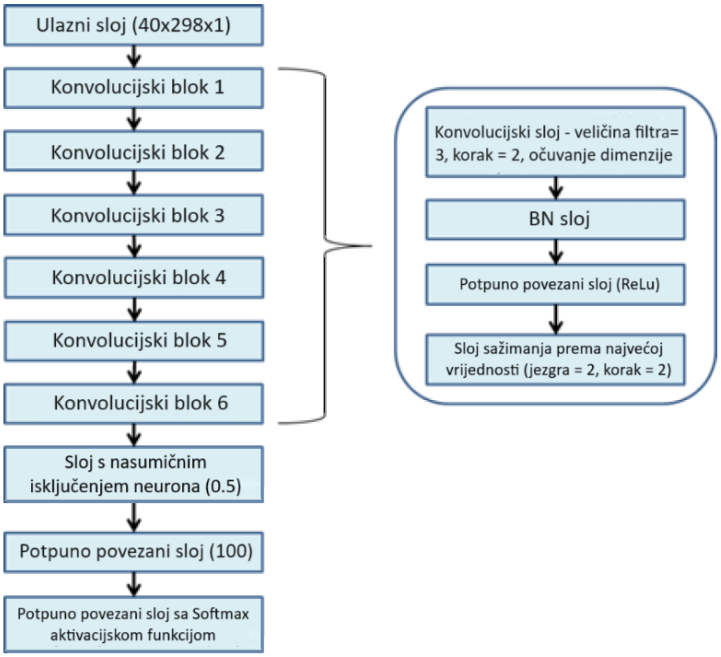
[7]. Iz izraza (2-2), vidljiva je jasna povezanost rada sijamske neuronske mreže i procesa identifikacije, a to je prepoznavanje sličnosti uzoraka. Primjenom sijamske neuronske mreže za zadatak identifikacije govornika, proces se svodi na usporedbu sličnosti testnog uzorka sa svakim uzorkom poznatih govornika spremljenih u bazi podataka, a odabire se onaj koji ima najveću sličnost. Time se ogleda jedna od prednosti uporabe sijamske mreže u području prepoznavanja govornika, a to je jednostavno proširenje baze postojećih govornika. Međutim, ako je mreža unaprijed trenirana na skupu podataka poznatih govornika koji su predstavljeni kao pojedine klase, kao što je identifikacija govornika primjenom CNN, određivanje identiteta govornika testnog uzorka vrši se na način da mreža klasificira testni uzorak kao određenu klasu koja predstavlja identitet govornika na kojem je mreža trenirana. Drugim riječima, mreža je u stanju samo prepoznati one govornike na kojima je mreža trenirana, bez mogućnosti jednostavnog proširenja baze podataka poznatih govornika kao kod sijamske neuronske mreže [8, 15]. U nastavku su navedeni radovi koji ostvaruju identifikaciju govornika na jedan od ova dva načina.

2.2.1. Konvolucijska neuronska mreža za identifikaciju korisnika u uvjetima sa i bez šuma

Radom [15] predlaže se arhitektura konvolucijske neuronske mreže prikazane slikom 2.8. Predložena arhitektura konvolucijske neuronske mreže sastoji se od: ulaznog sloja, 6 konvolucijskih blokova, sloja s nasumičnim isključenjem neurona s faktorom isključenja vrijednosti 0.5 te potpuno povezanog sloja sa Softmax aktivacijskom funkcijom. Svaki konvolucijski blok sastoji se od 4 sloja:

- konvolucijski sloj s veličinom filtra 3x3 i korakom 2, sa zadržavanjem veličine ulaza,
- *Batch* normalizacijski sloj,
- potpuno povezanog sloja s ReLU aktivacijskom funkcijom i

- sloja sažimanja prema najvećoj vrijednosti s jezgrom veličine 3x3 i korakom 2.



Slika 2.8. *Prikaz predložene arhitekture konvolucijske neuronske mreže za identifikaciju govornika iz rada* [15]

Za potrebe treniranja i testiranja mreže korišteni su zvučni zapisi od 100 različitih govornika iz TIMIT skupa podataka [33]. Skup podataka je podijeljen na način da se 80% uzoraka koristi za trening, 10% za validaciju i 10% za testiranje rada mreže. Svaki govornik ima 480 rečenica na koje je umjetno dodan šum (8 rečenica s 15 vrsta šuma i četiri razine šuma) i dodatnih 8 rečenica na koje nije umjetno dodan šum, što ukupno čini 488 rečenica za svakog govornika koje se koriste za treniranje mreže. Svaka rečenica je duljine 3 sekunde što je prosječna duljina izgovora fraze unutar TIMIT skupa podataka, gdje su rečenice koje su kraće od 3 sekunde dopunjene nulama, a rečenice koje su dulje od 3 sekunde izrezane. Umjetno dodavanje šuma ostvareno je u omjeru signala u odnosu na šum (engl. *Signal-to-Noise Ratio* - SNR) u iznosima od: 0 dB, 5 dB, 10 dB i 15 dB. Manjim iznosom SNR vrijednosti ostvaruju se veće deformacije izvornog signala. Korišteno je 15 različitih vrsta šumova dobivenih iz NOISEX-92 skupa podataka [34]. Za sprječavanje pretjeranog usklađivanja mreže na trening podatke, prilikom treniranja mreže korištena je augmentacija u dva koraka. Prvi korak augmentacije temelji se na modifikaciji zvučnog zapisa, a drugi na modifikaciji spektrograma. Modifikacija zvučnog zapisa uključuje augmentaciju koja se izvodi na način da se izreže uzorak rečenice nasumične duljine i taj uzorak zamjeni s nasumičnim uzorkom iste rečenice obrnutog redoslijeda izvođenja. Nakon prvog koraka augmentacije izračunava se logaritamski Mel spektrogram iz zvučnog zapisa. Izračun rezultira slikom s jednim kanalom, visine 40 piksela i širine 298 piksela. U drugom koraku augmentacije, dobivena slika se pomiče u vodoravnom smjeru, s nasumičnom iznosom pomaka u rasponu od

-30 do 30 piksela. Osim pomaka, također se vrši množenje svakog piksela slike s nasumičnom iznosom raspona od 0.8 do 1.2. Dobivena slika je ulaz konvolucijske neuronske mreže. Treniranje mreže provedeno je u 8 epoha stohastičkom metodom gradijentnog spusta (engl. *stochastic gradient descent* - SGD) s momentom iznosa 0.9, *batch* veličinom 256, stopom učenja od 0.0005. Smanjenje stope učenja primjenjuje se svake četvrte epohe, množenjem trenutne stope učenja s faktorom smanjenja stope učenja čiji je iznos 0.1.

Kako bi odredila performansa rada mreže predložene radom [15] u području identifikacije govornika izvršena je usporedba s klasičnom tehnikom u području identifikacije govornika koja se naziva univerzalni pozadinski model temeljen na Gaussovom modelu mješavine. Model predložen radom [15] je pokazao značajno poboljšanje u točnosti identifikacije, posebice pri niskim razinama SNR vrijednosti. Međutim, rezultati ukazuju na nižu točnost za čiste zvučne zapise u usporedbi s UBM-GMM tehnikom koja postiže gotovo savršen rezultat [15].

2.2.2. Sijamske neuronske mreže za identifikaciju govornika

Proces identifikacije govornika korištenjem sijamske neuronske mreže svodi na sljedeće korake. Prvi korak je izdvajanje značajki iz zapisa, nakon čega slijedi usporedba tih značajki sa značajkama zapisa svih poznatih govornika u bazi podataka te u konačnici dolazi do donošenja odluke na temelju usporedbe. Proces identifikacije govornika uporabom sijamske mreže temelji se na usporedbi uzorka nepoznatog govornika sa svakim uzorkom poznatih govornika. Ako je vjerojatnost predikcije sličnosti govornika iznad postavljenog praga odluke, taj govornik predstavlja identitet nepoznatog govornika. U suprotnom, ako usporedba s nijednim uzorkom nije rezultirala vjerojatnosti predikcije koja je veća od praga odluke, govornik nije prepoznat. Međutim, ispravnost rada identifikacije govornika ovisi o odabiru vrijednosti praga odluke. Niska vrijednost praga može dovesti do neovlaštenog pristupa prepoznavanjem lažnog govornika. S druge strane, visoka razina praga odluke može otežati identifikaciju povećanjem broja lažnih odbijanja. U osnovni tri su ključna koraka u identifikaciji govornika primjenom sijamske mreže. Prednost korištenja sijamske neuronske mreže u području identifikacije govornika iskazana je smanjenim brojem uzoraka potrebnih za treniranje i lakšom prilagodbom proširenja skupa podataka u odnosu na klasične metode GMM, HMM, metode potpornih vektora, ali i CNN [8].

Treniranje neuronske mreže temelji se na funkciji gubitka (engl. *loss function*) koja kvantificira razliku između stvarnih vrijednosti i vrijednosti predviđenih mrežom. Na temelju iznosa funkcije gubitka parametri neuronske mreže se prilagođavaju kako bi se postigao niži iznos gubitka. Tijekom treniranja sijamske neuronske mreže može se koristiti nekoliko različitih funkcija

gubitka, uključujući binarni unakrsni entropijski gubitak (engl. *binary cross-entropy loss*), kontrastni gubitak (engl. *contrastive loss*), gubitak tripleta (engl. *triplet loss*) i gubitak konstelacije (engl. *constellation loss*). Binarna unakrsna entropija je funkcija gubitka koja se može prilagoditi u svrhu identifikacije govornika jer uspoređuje svaku predviđenu vrijednost sa stvarnom vrijednosti. Međutim, kontrastni gubitak nadmašuje funkciju binarnog unakrsnog entropijskog gubitka za treniranje sijamske neuronske mreže koja se koristi za zadatak identifikacije govornika, budući da je odličan u razlikovanju dvaju uzoraka. Prema tome minimiziranje funkcije kontrastnog gubitaka odabire se kao temelj optimiziranja parametara mreže u radu [8]. Princip rada funkcije kontrastnog gubitaka se temelji na minimiziranju mjere udaljeni sličnih vektora i maksimiziranju mjere udaljenosti različitih vektora. Izračun kontrastnog gubitka prikazan je sljedećim izrazom:

$$L = (1 - Y) \times \left||z_1 - z_2|\right|^2 + Y \times max\left(0, m - \left||z_1 - z_2|\right|^2\right). \quad (2\text{-}3)$$

Prema izrazu (2-3), parametar L predstavlja gubitka, Y stvarnu vrijednost, z_1 vektor značajki referentnog uzorka, z_2 vektor značajki novog uzorka, $\left||z_1 - z_2|\right|^2$ mjera udaljenosti vektora i m predstavlja hiperparametar koji navodi donju granicu udaljenosti između različitih uzoraka [8]. Vrijednost parametra Y postavlja se na 0, ako se pretpostavi da su uzorci slični. U suprotnom se postavlja na vrijednost 1. Rezultat gubitka se nastoji smanjiti minimiziranjem člana $||z_1 - z_2||^2$, ako se pretpostavi da se radi o sličnim uzorcima. U suprotnom se maksimizira izraz $max\left(0, m - \left||z_1 - z_2|\right|^2\right)$.

Radom [8] implementirana je identifikacija govornika pomoću sijamskih neuronskih mreža na ugradbenom Raspberry Pi 4 uređaju. Zbog ograničenja računalne snage Raspberry Pi uređaja, kao prihvatljiva podmreža za izvlačenje značajki odabire se manje zahtjevna mreža, poput MobileNetv2 ili SqueezeNet mreže. Nedostatak manje zahtjevnih mreža je ograničenost pri učinkovitosti izvlačenja značajki što utječe na točnost rada mreže. Podmreže poput VGG16 i ResNet50 ostvaruju bolje rezultate točnosti modela, no zahtijevaju veću računalnu snagu i memoriju zbog čega nisu prikladne za implementaciju na ugradbenim sustavima. Za potrebe treniranja i testiranja korišten je VoxCeleb2 skup podataka koji ima više od milijun zvučnih zapisa od 6112 različitih govornika dobivenih iz video uradaka objavljenih na YouTube platformi [35]. Podaci su podijeljeni na način da se 80% podataka koristi za treniranje mreže, a preostalih 20% podataka za testiranje rada mreže. Podmreži se kao ulaz predaje Mel spektrogram te se mreža trenira u dvije faze s različitom funkcijom gubitka. U prvoj fazi treniranja koristi se MobileNetv2 podmreža, funkcija gubitka je binarna unakrsna entropija, mreža kositi Adam optimizator i stopu

učenja iznosa 0.0003. *Batch* veličina je 32, a mreža se trenira 40 epoha s 1000 koraka po epohi. U drugoj fazi rada kao funkcija gubitka korišten je kontrastni gubitak, a odabrane podmreže su: MobileNetv2, SqueezeNet i MCUNet256kb. *Batch* veličina je povećana na 64 kako bi se iskoristila dostupna memorija, dok svi ostali parametri ostaju isti.

Zaključno, mreža identifikacije govornika koja koristi podmrežu MCUNet256kb postigla je najveću točnost, dok mreža koja koristi SqueezeNet podmrežu ostvaruje najbrže vrijeme donošenja zaključka. Problem se u konačnici sveo na kompromis između razine točnosti identifikacije govornika i brzine kojom mreža dolazi do zaključka.

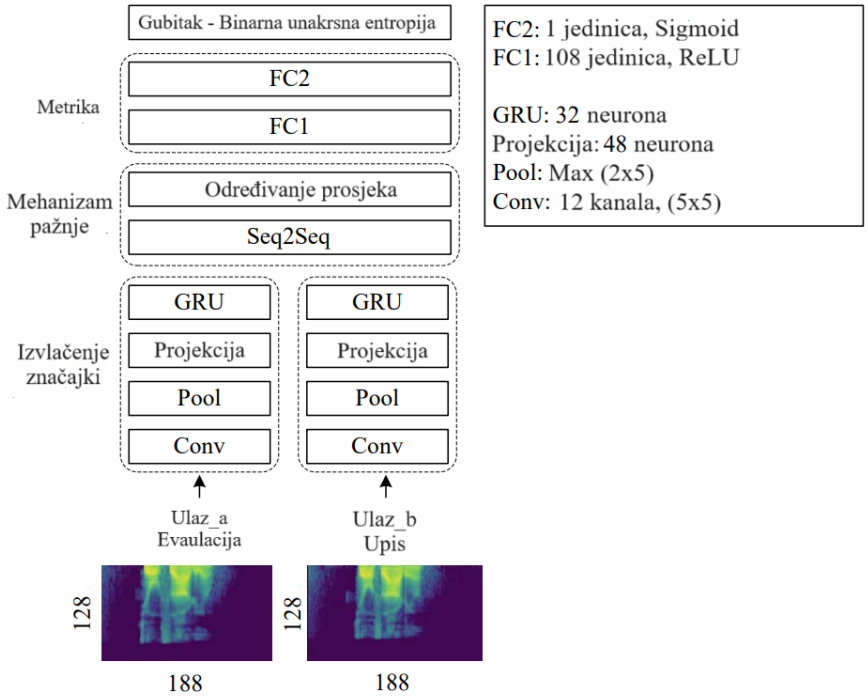
2.3. Verifikacija govornika korištenjem dubokih neuronskih mreža

Pojam verifikacije govornika odnosi se na proces potvrde pretpostavljenog identiteta govornika na temelju zvučnog zapisa govornika. Drugim riječima, ako pretpostavimo da se radi o određenom govorniku, cilj verifikacije je potvrditi radi li se doista o tom govorniku na temelju zvučnog zapisa, tj. uzorka govornika [36, 37]. Nadalje, zadatak verifikacije govornika kategorizira se u one ovisne o izgovorenom sadržaju (engl. *text-dependent*) i one neovisne o izgovorenom sadržaju (engl. *text-independent*). Kod pristupa verifikacije koja je ovisna o izgovorenom sadržaju, potvrda identiteta govornika ovisit će o izgovorenom sadržaju zvučnog zapisa govornika, što može biti fraza, zaporka i slično. Međutim, kod verifikacije koja je neovisna o izgovorenom sadržaju to nije slučaj. Kod pristupa verifikacije neovisne o izgovorenom sadržaju, treniranje mreže zahtjeva uzorke veće duljine kako bi se smanjio utjecaj fonetske varijabilnosti koja nastaje zbog različitog sadržaja svakog zvučnog zapisa. S druge strane, kod pristupa verifikacije ovisne o izgovorenom sadržaju postoji mala fonetska varijabilnost zbog jednakog sadržaja zvučnog zapisa i moguće je ostvariti treniranje mreže na kratkim uzorcima uz postizanje visoke točnosti. Ovaj pristup postiže bolje rezultate verifikacije, ali zahtjeva treniranje mreže na velikom skupu podataka. Međutim, prikupljanje velikog broja zvučnih zapisa koji imaju isti izgovoreni sadržaj i izgovara ih isti govornik, u praksi je vrlo zahtjevno i teško ostvarivo [26, 37]. U nastavku su navedeni radovi koji ostvaruju verifikaciju govornika na jedan od ova dva načina.

2.3.1. Sijamske neuronske mreže za verifikaciju govornika ovisne o izgovorenom sadržaju upotrebom mehanizam pažnje od sekvence do sekvence

Radom [26] predlaže se Seq2Seq-ASNN (engl. *Sequence-to-Sequence Attentional Siamese Neural Network*) mreža za verifikaciju govornika ovisna o izgovorenom sadržaju. Navedena mreža koristi sijamsku neuronsku mrežu u kombinaciji s mrežom pažnje koja nastoji riješiti problem neusklađenosti u fonetskom kontekstu i razlike duljine zapisa između evaluacije (engl. *evaluation*)

i upisa (engl. *enrollment*). Upis predstavlja referentni uzorak na temelju kojeg se vrši provjera podudaranja. Pojam evaluacije odnosi se na ulaz u mrežu kojeg korisnik unosi kada započinje s postupkom verifikacije. Neusklađenost u fonetskom kontekstu se može dogoditi kada se predani i referentni uzorak razlikuju u izgovoru riječi, naglasku i slično. Slikom 2.9. prikazana je arhitektura mreže predložene radom [26]. Sijamska neuronska mreža sastoji se od dvije podmreže s identičnom strukturom koje služe za izvlačenje značajki iz spektrograma koji su ulaz mreže. Svaka se podmreža sastoji od konvolucijskog sloja nakon kojeg slijedi povratni sloj (engl. *Gated Recurrent Units* – GRU) koji služi za izdvajanje vremensko-frekvencijskih značajki. Izlazi iz podmreža predaju se mreži pažnje koja vremenski usklađuje značajke evaluacije i upisa. Na kraju, dva potpuno povezana sloja donose odluku radi li se o istom govorniku. Optimizacija parametara mreže temelji se na minimizaciji funkcije gubitka od kraja do kraja (engl. *end-to-end loss*).

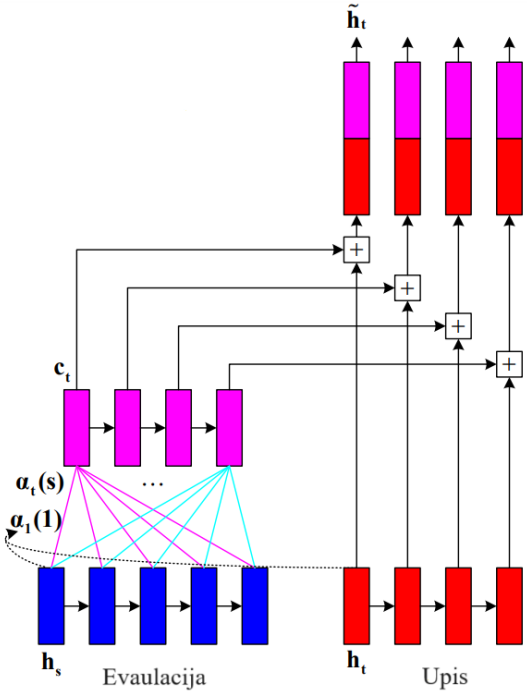


Slika 2.9. *Prikaz arhitekture Seq2Seq-ASNN mreže za verifikaciju govornika predložene radom* [26]

U radu [26] mehanizam pažnje od sekvence do sekvence koristi se kako bi se vremenski okviri značajki upisa uskladili s vremenskim okvirima značajki evaluacije, pri čemu se navedene značajke dobivaju kao izlazi iz podmreža. Svaki okvir značajki upisa usklađuje se s težinskim prosjekom okvira evaluacije. Razlog uporabe težinskog prosjeka okvira značajki evaluacije je određivanje okvira koji ima veći značaj od drugih. Nadalje, takvi težinski prosječni okviri značajki evaluacije spajaju se s izvornim okvirima značajki upisa, što rezultira vektorom koji sadrži informacije o značajkama iz evaluacije i upisa. Za taj vektor također je izračunat prosjek što je

izlaz navedenog mehanizma. Struktura mehanizama pažnje od sekvence do sekvence predložene radom [26] prikazana je na slici 2.10., gdje su:

- h_s i h_t značajke govornika na razini okvira za zapise evaluacije i upisa,
- c_t – najznačajnije informacije o zapisu evaluacije i
- $h_t\sim$ - vektor značajki upisa i evolucije.



Slika 2.10. *Struktura mehanizma pažnje od sekvence do sekvence predložene radom* [26]

Mreža je trenirana i evaluirana na Tencent skupu podataka koji uključuje 3324 govornika s ravnopravnom zastupljenosti spola. Svaki govornik ima zabilježenih 30 izgovora ključne riječi „9420“ izgovorenih na kineskom jeziku. Cijeli skup podataka podijeljen je na 2570 govornika za treniranje, 635 za validaciju i 119 za testiranje. U postupku obrade zvučnih zapisa, zapisi se uzorkuju na 16 kHz i sijeku na duljine 3 sekunde te nadopunjuju nulama ako je zapis kraći od 3 sekunde. Zvučni zapisi se zatim pretvaraju u logaritamski Mel spektrogram s duljinom prozora od 32 ms i preklapanjem od 16 ms što rezultira dimenzijom spektrograma visine 128 i širine 188. Tijekom treniranja sijamskoj se neuronskoj mreži predaju pozitivni i negativni parovi spektrograma. Budući da se kroz rad [26] implementira verifikacija govornika ovisna o izgovorenom sadržaju uzorka, pozitivni parovi odgovaraju onim zvučnim zapisima koji pripadaju istom govorniku i izgovara se isti sadržaj. U suprotnom, ako uzorci ne pripadaju istom govorniku ili se ne izgovara isti sadržaj radi se o negativnim parovima. Nadalje, za treniranje mreže korištena je stopa učenja stohastičkog gradijentnog spusta iznosa 0.1, uz stopu opadanja (engl. *decay rate*) od 0.001 i konstantu momenta (engl. *momentum constant*) od 0.9, a *batch* veličina iznosi 256.

Treniranje mreže prekida se ako u 3 slijedne epohe ne dođe do značajnije promjene vrijednosti gubitaka na validacijskom skupu.

Za potrebe ocjenjivanja rada mreže koristi se stopa jednake pogreške (engl. *Equal Error Rate* - EER). U usporedbi rada mreže, odnosno modela verifikacije govornika, korištena je i informacija o složenosti mreže u smislu broja parametara koji se koriste u procesu treniranja mreže. Usporedba je izvršena s nekoliko različitih modela od kojih su neki: d-Vector, i-Vector/PLDA, Google-End2End-1 i Google-End2End-2 modeli, a rezultati su prikazani tablicom 2.1. Iz rezultata prikazanih tablicom 2.1., vidljivo je kako model Seq2Seq-ASNN predložen radom [26], iako ima manju složenost modela, nadmašuje složenije modele poput Google-End2End i d-Vector modela. Model Seq2Seq-ASNN predložen radom [26] nadmašuje model Siamese-CNN-GRU koji je zapravo dio izvornog modela koji ne koristi mehanizam pažnje s vremenskim usklađivanjem signala, ukazujući na učinkovitost mehanizma pažnje za vremensko usklađivanje u području verifikacije govornika smanjenjem iznosa EER s 1.87% na 0.36% [26].

Tablica 2.1. *Prikaz usporedbe modela korištenjem EER metrike i veličine modela različitih konfiguracija* [26]

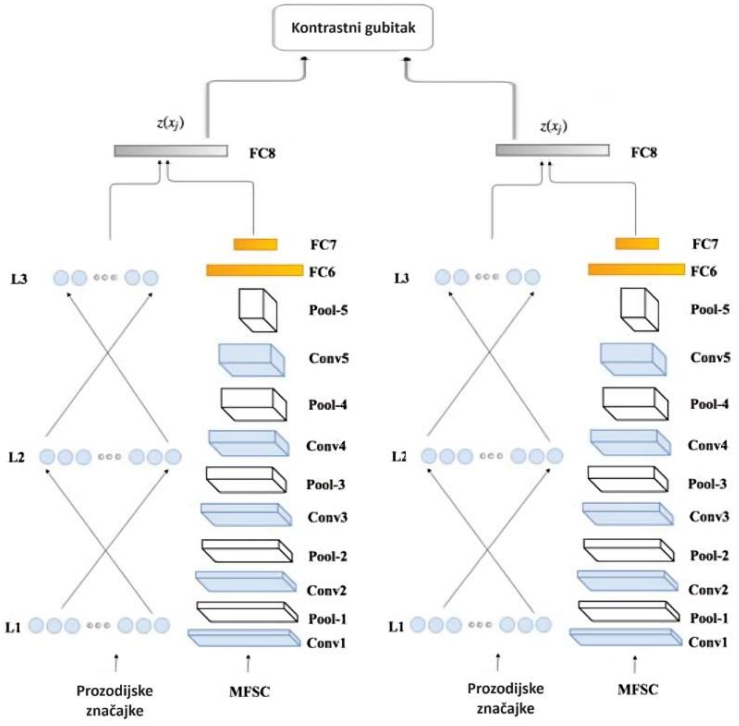
Model	**Veličina modela (broj parametara)**	**EER**
i-Vector/PLDA	-	0.56%
Google-End2End-1 (40 × 80)	1.1 milijuna	4.56%
Google-End2End-2 (128 × 188)	1.1 milijuna	4.28%
d-Vector (w/o Cosface)	0.3 milijuna	8.00%
d-Vector (w/ Cosface)	0.3 milijuna	1.50%
Self-ASNN	149.7 tisuća	1.73%
Siamese-CNN	146.7 tisuća	3.40%
Siamese-CNN-GRU	146.7 tisuća	1.87%
Seq2Seq-ASNN	149.7 tisuća	**0.36%**

2.3.2. Prozodijski poboljšane sijamske neuronske mreže za verifikaciju govornika neovisne o izgovorenom sadržaju s implementacijom na različitim uređajima

Radom [38] predlaže se sijamska neuronska mreža koja koristi pristup verifikacije govornika neovisne o izgovorenom sadržaju. Mreža, osim što koristi mehanizam naziva koeficijenti Mel frekvencijskog spektrograma (engl. *Mel-frequency spectrogram coefficients* - MFSC) za izvlačenje značajki, upotrebljava i višeslojnu mrežu perceptrona (engl. *Multilayer Perceptron* - MLP). Razlog leži u nedostatku MFSC tehnike kojom se predstavljaju samo frekvencijske

karakteristike govornika. Međutim, ljudski glas sastoji se od jezičnih karakteristika kao što su akustika, leksikon, prozodija i fonetika koje se mogu koristiti u zadacima prepoznavanja govornika. Višeslojnom mrežom perceptrona obuhvaćene su prozodijske značajke (engl. *prosodic features*), podrhtavanje (engl. *jitter*) i svjetlucanje (engl. *shimmer features*) [38]. Prozodijske značajke su aspekti govora koji uključuju: naglasak, stres, ritam, tonalitet, visinu tona i intonaciju. Ovisno o jeziku i geografskom položaju govornika, navedene značajke mogu varirati što se iskazuje razlikom u frekvenciji, duljini trajanja i u obrisu zapisa [39, 40]. Podrhtavanje se definira kao varijacija osnovne frekvencije od ciklusa do ciklusa i mjera je stabilnosti glasa. S druge strane, svjetlucanje je promjena amplitude glasa. Ukupno 18 prozodijskih značajki izdvaja se iz zvučnih zapisa, uključujući značajke povezane s trajanjem riječi, frekvencijom te podrhtavanjem i svjetlucanjem.

Predložena arhitektura sijamske neuronske mreže u radu [38] dana je slikom 2.11. i sastoji se od dvije podmreže koje dijele iste težine. Svaka podmreža uključuje višeslojnu mrežu perceptrona i konvolucijsku neuronsku mrežu te sloj zajedničkog prikaza (engl. *joint representation layer*). Pojedinačne mreže se spajaju kako bi se dobio vektor značajki na temelju kojeg se vrši usporedba. Mel frekvencijski spektrogrami stvaraju se iz zvučnih zapisa govornika i predaju se konvolucijskoj neuronskoj mreži koja služi za izvlačenje značajki iz tog spektrograma. Višeslojna mreža perceptrona sadrži dva skrivena sloja, pri čemu svaki skriveni sloj sadrži 64 neurona, dok izlazni sloj uključuje 32 neurona.



Slika 2.11. *Prikaz sijamske neuronske mreže za verifikaciju govornika predložene radom* [38]

Treniranje i testiranje mreže provedeno je na FBI Voice Collection 2016 skupu podataka, koji uključuje izjave 411 različitih govornika, odnosno 205 muških i 206 ženskih govornika. Snimci su zabilježeni putem tri različita uređaja: visokokvalitetni mikrofon, digitalni video snimač (engl. *Digital video recorder* - DVR) i snimač na mobilnom uređaju. Svi izgovori inicijalno su uzorkovani s 48 kHz. Prozodijske značajke za svaki izgovor izdvojene su pomoću Praat [41] programske podrške za akustičnu analizu koji se predaju MLP mreži. Za otkrivanje glasovnih segmenata, odnosno dijelova zvučnog zapisa gdje se pojavljuje izgovoreni sadržaj govornika, koristi se alat Voicebox toolbox [42] kroz MATLAB programsku podršku. Mel frekvencijski spektrogram dobiven je na temelju parametara duljine prozora iznosa 25 ms i s 60 postotnim preklapanjem prozora. U konačnici, konvolucijskoj neuronskoj mreži predaje se spektrogram iznosa visine 40, širine 300 piksela i s tri kanala. Tijekom treniranja konvolucijske neuronske mreže primijenjeno je nasumično izbacivanje neurona s vjerojatnošću od 50%, *batch* veličina je 32. Početna stopa učenja postavljena je na 0.1, a smanjenje stope učenja ostvareno je množenjem stope učenja s 0.1 svake druge epohe. Optimizacija parametara mreže temelji se na minimizaciji funkcije kontrastnog gubitka.

Tablica 2.2. prikazuje rezultate evaluacije mreže. Osim rezultata mreže predloženog radom [38], dodatno su prikazani rezultati pojedinačnih CNN i MLP mreža kao i spoj tih mreža na razini rezultata (engl. *Score-level fusion*). Performanse su predloženog modela uspoređene s performansama i-vektor/PLDA algoritama koji predstavljaju tradicionalni pristup verifikacije govornika. Iz tablice je moguće vidjeti kako model predložen radom [38] daje najbolje rezultate i time je moguće zaključiti kako je kombinacija CNN i MLP mreže bila uspješna.

Tablica 2.2. *Prikaz rezultata usporedbe modela* [38]

Algoritam/Model	Točnost	EER
i-vektor/PLDA (MFSC)	0.9153	0.1579
i-vektor/PLDA (MFCC)	0.9185	0.1526
CNN	0.9218	0.1421
MLP	0.9011	0.1673
Spoj na razini rezultata	0.9148	0.1578
Predložena mreža	0.9358	0.1311

Drugi dio rada vezan je uz implementaciju mreže na različitim uređajima u smislu različitih izvora prikupljanja zvučnih zapisa govornika. Korišteni uređaji su: visoko kvalitetni mikrofon, digitalni video snimač i mobilni uređaj. Tablicom 2.3. i tablicom 2.4. prikazani su rezultati evaluacije na

temelju metrike stope jednake pogreške i točnosti mreže na različitim uređajima. Iz rezultata je moguće uočiti kako su najbolji rezultati dobiveni korištenjem visoko kvalitetnog mikrofona [38].

Tablica 2.3. *Prikaz rezultata stope jednake pogreške mreže ovisno o uređaju kojim je snimljen zvučni zapis* [38]

Stopa jednake pogreške - EER			
Uređaj	Mikrofon	DVR	Mobilni uređaj
Mikrofon	0.0712	0.1132	0.1247
DVR		0.0827	0.2069
Mobilni uređaj			0.1316

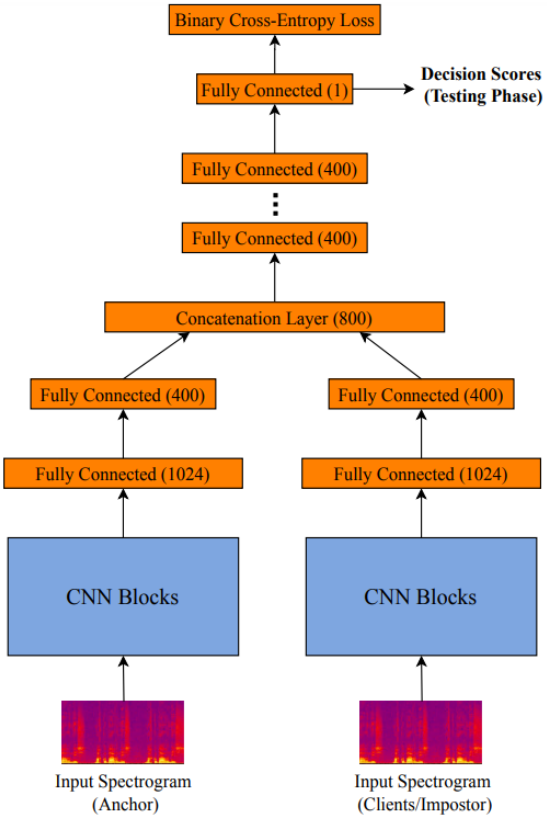
Tablica 2.4. *Prikaz točnosti mreže ovisno o uređaju kojim je snimljen zvučni zapis* [38]

Točnost mreže			
Uređaj	Mikrofon	DVR	Mobilni uređaj
Mikrofon	0.9785	0.9537	0.9512
DVR		0.9717	0.8467
Mobilni uređaj			0.9547

2.3.3. Sijamske mreže za verifikaciju govornika primjenom nenadziranog učenja

U radu [43] su predložene dvije sijamske neuronske mreže koje primjenjuje pristup nenadziranog učenja. Nenadzirano učenje je vrsta strojnog učenja kod kojeg se mreža trenira na podacima koji nemaju odgovarajuće izlazne veličine. Preciznije, mreža trenira na neoznačenim podacima, odnosno podacima za koje nije poznato kojoj klasi pripadaju, s ciljem izvlačenja zajedničkih značajki iz ulaznih podataka [44]. Razlog uporabe nenadziranog učenja u radu [43] je ograničenost pristupu velikom broju podataka s poznatim identitetom govornika. Drugim riječima, nije moguće odrediti koji su to govornici u skupu podataka isti, odnosno koji uzorci pripadaju istoj klasi. Budući da unaprijed nisu poznati identiteti govornika, nije moguće jednostavno izraditi pozitivne i negativne parove kao kod nadziranog učenja. Radom [43] implementira se verifikacija govornika neovisna o sadržaju uzorka, prema tome pozitivni par predstavlja par uzoraka istog govornika, a negativni kombinaciju uzoraka različitih govornika. Pozitivni parovi su odabrani na način da je zvučni zapis referentnog uzoraka sidra (engl. *anchor utternace*) uspoređen s podacima iz prvog skupa podataka te se odabire onaj zapis koji ima najveći kosinusni rezultat u usporedbi sa zapisom referentnog uzorka sidra u i-vektorskom prostoru. Spektrogram takvog zvučnog zapisa koji u kombinaciji sa spektrogramom uzorka sidra stvara pozitivni par naziva se uzorak klijenta (engl. *client utternace*). Nadalje, negativni parovi odabrani su iz drugog skupa u kojem je sa sigurnošću poznato da nema zvučnih zapisa govornika iz prvog skupa podataka. Spektrogram koji s uzorkom

sida stvara negativan par naziva se uzorak uljeza (engl. *imposter utternace*). Arhitektura sijamskih neuronskih mreža prikazana je slikama 2.12. i 2.13. Slikom 2.12. prikazana je arhitektura sijamske mreže koja koristi dvije grane za izvlačenje značajki. Mreža kao ulaz prima parove Mel spektrograma. U svakoj od grana nalazi se konvolucijska neuronska mreža i dva potpuno povezana sloja na temelju kojih se izvlače značajke pojedinih spektrograma. Optimizacija parametara mreže temelji se na minimizaciji funkcije binarne unakrsne entropije.



Slika 2.12. *Prikaz arhitekture sijamske neuronske mreže koja koristi dvije grane predložene radom* [43]

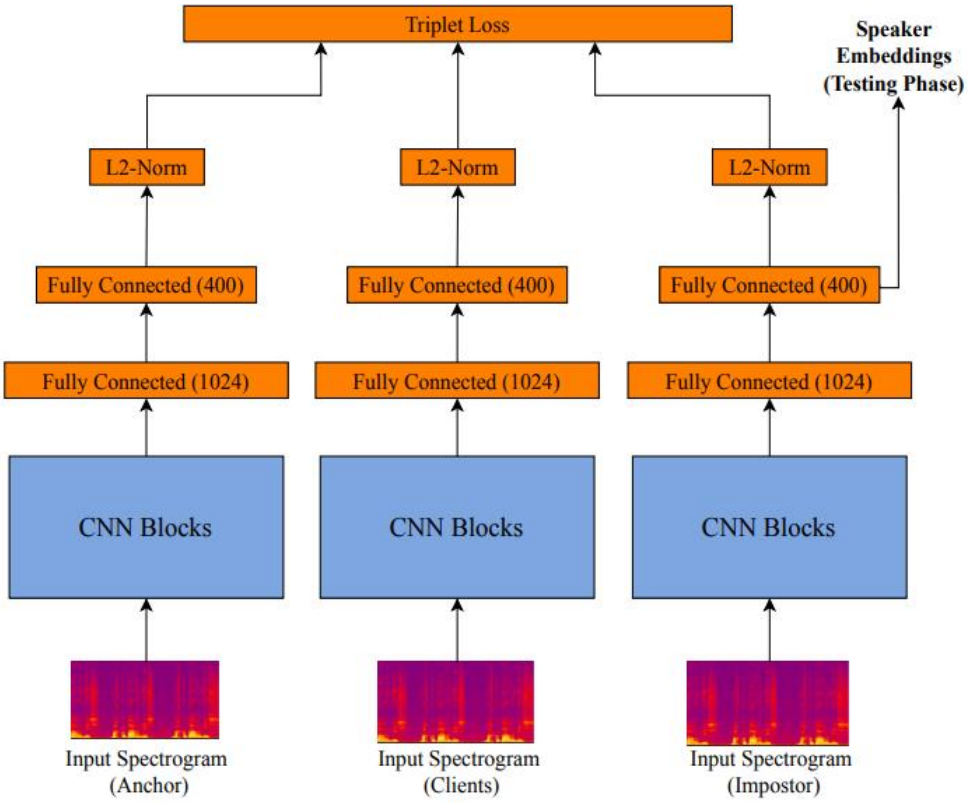
Slikom 2.13. prikazana je arhitektura sijamske neuronske mreže koja koristi 3 grane za izvlačenje značajki. U usporedbi s mrežom koja koristi dvije grane, kao ulaz se predaju tri spektrograma. Prvi spektrogram odgovara uzorku sidra, drugi uzorku klijenta i treći spektrogram uzorku uljeza. Arhitektura pojedinih grana je slična onoj prikazane slikom 2.12. Izlaz svake podmreže je L2 normalizacijski sloj koji se primjenjuje u svrhu smanjenja utjecaja razlike u amplitudi izgovara unutar zvučni zapisa s ciljem fokusa na stvarnu razliku između značajki govornika. Optimizacija parametara mreže temelji se na minimizaciji gubitka tripleta koja je jednaka:

$$L = max(d(a,c) - d(a,i) + m, 0), \tag{2-4}$$

gdje je:

- $d(a, c)$ – mjera udaljenost između uzorka sidra i uzorka klijenta,
- $d(a, i)$ – mjera udaljenost između uzorka sidra i uzorka uljeza i
- m – marginalna udaljenost koja definira koliko bi trebala iznositi razlika između uzoraka.

Cilj ove funkcije gubitka je minimizirati mjeru udaljenosti između uzoraka sidra i uzorka klijenta, koji trebaju biti slični, dok se maksimizira mjera udaljenosti između uzoraka sidra i uzorak uljeza koji trebaju biti različiti.



Slika 2.13. *Prikaz arhitekture sijamske neuronske mreže koja koristi tri grane predložene radom* [43]

Evaluacija mreža izvršena je na VoxCeleb-1 skupu podataka, a treniranje mreža na participama VoxCeleb-2 skupa podataka. Trening skup je podjednako podijeljen na pozitivne i negativne parove. Mreže su optimizirane korištenjem Adam optimizatora sa stopom učenja od 0.0001. Treniranje mreža je ograničeno do 500 epoha ili do stagniranja rezultata u slijednih pet epoha. Kod mreže koja koristi funkciju gubitka tripleta parametar *m* je postavljen na vrijednost 0.8. Budući da nije poznat broj pozitivnih parova, odabire se *k* uzoraka s najvećim kosinusnim vrijednostima koji će predstavljati pozitivni par. Mreže su trenirana s odabirom nekoliko različitih vrijednosti *k*, a tablicom 2.5. prikazani su rezultati mreže koja ima dvije i mreže koja ima tri grane s različitim vrijednostima parametra *k*. Također, istrenirane mreže su uspoređene s klasičnim metodama kao što je pristup naziva i-vektori/PLDA. Usporedba rada mreža odrađena je na temelju rezultata stope

jednake pogreške. Iz tablice 2.5. je vidljivo kako mreža s dvije grane ima najmanju stopu jednake pogreške od 6.90% za vrijednost parametra *k* jednake 10. Povećanjem vrijednosti parametra *k* ne dolazi se do značajnog poboljšanja rezultata. Za jednaku vrijednosti parametra *k*, mreža s tri grane ostvaruje gotovo sličan rezultat, s EER vrijednosti od 6.95%. U konačnici, najbolji rezultat ostvaruje pristup spoja mreža s dvije i tri grane na razini rezultata. Navedeni pristup, doveo je do poboljšanja rezultata s postignutom EER vrijednosti od 6.07%. Međutim, ni takav pristup nije mogla prestići model Baseline koji koristi AMSoftmax aktivacijsku funkciju.

Tablica 2.5. *Prikaz rezultata stope jednake pogreške kod različitih modela s različitom vrijednosti parametra k u radu* [43]

Model (Mreža)	k	EER
(1) i-vector/PLDA	-	9.54%
(2) Baseline (Sofmax)	-	6.81%
(3) Baseline (AMSoftmax)	-	5.71%
(4) 2 - grane	2	7.81%
(5) 2 - grane	5	7.73%
(6) 2 - grane	10	6.90%
(7) 3 - grane	10	6.95%
Spoj na razini rezultata mreža (6) i (7)	10	6.07%

3. PREDLOŽENO RJEŠENJE ZA VERIFIKACIJU GOVORNIKA TEMELJENO NA IZGOVORENOM SADRŽAJU

U ovom poglavlju detaljno je opisan proces izrade rješenja za verifikaciju govornika temeljenog na izgovorenom sadržaju. Pristup rješavanju navedenog problema temelji se na uporabi tipične sijamske neuronske mreže koja kao podmrežu za izvlačenje značajki koristi konvolucijsku neuronsku mrežu. Sijamskoj neuronskoj mreži kao ulaz se predaju parovi logaritamskih Mel spektrograma koji se dobivaju transformacijom zvučnih zapisa fraza govornika u vremensko-frekvencijsku domenu. Prvi element para predstavlja referentni uzorak, a drugi novi uzorak. Novi uzorak je onaj uzorak koji se unosi u postupku verifikacije govornika koji se uspoređuje s referentnim uzorkom. Treniranje mreže za verifikaciju govornika provedeno je na GRID skupu podataka čiji su zvučni zapisi prvotno obrađeni kako bi se dobio skup prikladan za ostvarenje verifikacije govornika. Skup je naknadno proširen dodatnim zvučnim zapisima drugih govornika s ciljem stvaranja raznolikosti skupa podataka. Navedeni skup koristi se za treniranje novih mreža. Nove mreže dobivene se procesom prijenosnog učenja na temelju mreža treniranih na GRID skupu podatka. U konačnici, izrađena je aplikacija za verifikaciju govornika koja korištenjem istreniranih mreža omogućuje verifikaciju govornika u stvarnom okruženju. U nastavku ovog poglavlja objašnjeni su postupci obrade GRID skupa podataka, proširenje GRID skupa podataka, treniranje mreža za verifikaciju govornika na temelju GRID skupa podataka, treniranje mreže prijenosnim učenjem i postupak izrade aplikacije za verifikaciju govornika.

3.1. GRID skup podataka

Za potrebe treniranja i evaluacije mreže odabran je GRID audio-vizualni (engl. *The Grid Audio-Visual Speech Corpus*) skup podataka verzije 1.0. Skup sadrži zvučne zapise 34 različita govornika, od kojih je 18 muških i 16 ženskih govornika. Svaki govornik je izgovorio i snimio 1000 zvučnih zapisa. Zvučni zapisi su snimljeni u okolini sa smanjenom prisutnosti šuma, frekvencijom uzorkovanja od 16000 Hz te su zvučni zapisi pohranjeni u *wav* formatu. Zbog različitog sadržaja i brzine izgovora fraza pojedinih govornika, duljina zvučnih zapisa varira unutar raspona od 1 do 2 sekunde [9].

Fraze u GRID skup podataka izgovorene su na engleskom jeziku. Svaka fraza formirana je od niza riječi koje su slijedom: naredba, boja, prijedlog, slovo engleske abecede, znamenka i prilog. Tablicom 3.1. prikazan je popis riječi korištenih za formiranje fraza. Kao primjer fraze može se navesti skup riječi „bin blue at A9 again“ [45].

Tablica 3.1. *Prikaz popisa riječi korištenih u GRID skupu podataka* [45]

Naredba	Boja	Prijedlog	Slovo	Znamenka	Prilog
bin	*blue*	*at*	*A-Z*	*0-9*	*again*
lay	*green*	*by*			*now*
place	*red*	*in*			*please*
set	*white*	*with*			*soon*

U nastavku je objašnjen postupak obrade zvučnih zapisa iz GRID skupa podataka koji omogućuje stvaranje pozitivnih i negativnih parova spektrograma. Također, detaljno su prikazane dvije podjele skupa podataka spektrograma, odnosno zvučnih zapisa pojedinih govornika.

3.1.1. Obrada zapisa GRID skupa podataka

Fraze GRID skupa podataka sastoje se od 6 različitih riječi koje zajedno čine jedinstvenu frazu koju govornik izgovara samo jedanput. Navedeno predstavlja problem u stvaranju pozitivnih parova, gdje pozitivan par predstavlja dva zvučna zapisa u kojem isti govornik izgovara istu frazu. Kako GRID skup nema takvih zvučnih zapisa, rješenje problema ostvareno je obradom zapisa na način da se uzima dio zvučnog zapisa koji odgovara prvim trima izgovorenim riječima fraze. Na ovaj je način došlo do povećanja broja zvučnih zapisa u kojima se izgovara ista fraza i izgovara je isti govornik.

Obrada zvučnih zapisa temelji se na izdvajanju dijelova zvučnog zapisa u kojima je detektiran izgovor prve tri riječi korištenjem metode detekcije glasovne aktivnosti (engl. *Voice Activity Detection* - VAD). VAD predstavlja algoritam za detekciju govorne aktivnosti na temelju energijske razine zvučnog signala. Metoda radi na način da se izdvoje dijelovi zvučnog zapisa iznad određenog praga koji se naziva energijski prag, a takve izdvojene dijelove algoritam detektira kao prisutnost govora [46]. Kod obrade zapisa izabran je energijski prag od 27 dB**.** Duljina zvučnog zapisa koji se promatra je 0.82 sekundi, a razlog odabira te vrijednosti je eksperimentalna procjena trajanja prosječnog izgovora prve tri riječi. Duljina izgovora pojedinih fraza nikada nije jednake duljine, ali odabranom vrijednosti od 0.82 sekunde uzima se prosjek kojim se promatra takav dio zvučnog zapisa u kojem se u većini slučajeva pojavljuju prve tri riječi. Nakon izdvajanja segmenta zvučnog zapisa koji odgovara intervalima zvučnog signala gdje se prve tri riječi izgovorene fraze pojavljuju u zapisu, izračunava se logaritamski Mel spektrogram sa stopom uzorkovanja od 16 kHz, veličinom prozora 512 brze Fourierove transformacije, preklapanjem između susjednih prozora od 256 uzoraka i veličinom Mel pojasa od 128. Spektrogram se zatim pretvara u logaritamsku domenu i skalira na dimenzije 100x50x1. Navedeno

je ostvareno pomoću programskog rješenja napisanog u Python3 programskom jeziku prikazanog programskim kodom 3.1. Za ostvarenje manipulacije nad zvučnim podacima korištena je Python biblioteka Librosa za glazbenu i zvučnu analizu [47]. Za brze operacije nad nizovima podataka korištena je biblioteka NumPy [48], a za promjenu dimenzije spektrograma biblioteka Skimage [49].

```
import librosa
import numpy as np
from skimage.transform import resize

def preprocess_audio(audio_file, sr=16000, target_duration=0.82):
    y, sr = librosa.load(audio_file, sr=sr)
    vad_segments = librosa.effects.split(y, top_db=27)
    # Odabir prve tri riječi
    word_boundaries = []
    word_count = 0
    for segment in vad_segments:
        if word_count >= 3:
            break
        if segment[1] - segment[0] > 0.16 * sr:
            word_boundaries.append(segment)
            word_count += 1
    combined_segment = np.concatenate([y[boundary[0]:boundary[1]] for boundary in
word_boundaries], axis=0)
    target_samples = int(target_duration * sr)
    combined_segment = combined_segment[:target_samples]
    # Izračun spektrograma
    spectrogram = librosa.feature.melspectrogram(
        y=combined_segment, sr=sr, n_fft=512, hop_length=256, n_mels=128 )
    log_spectrogram = librosa.power_to_db(spectrogram, ref=np.max)
    resized_spectrogram = resize(log_spectrogram, (100, 50))
    resized_spectrogram = np.expand_dims(resized_spectrogram, axis=-1)
    return resized_spectrogram
```

Programski kôd 3.1. *Prikaz programskog koda funkcije za izdvajanje dijela zvučnog zapisa koji odgovara prvim trima izgovorenim riječima i stvaranje spektrogama na temelju tog segmenta zvučnog zapisa*

3.1.2. Stvaranje skupa podataka koji se sastoji od parova spektrograma

Predobradom svih zvučnih zapisa u GRID skupu podataka funkcijom prikazanom programskim kodom 3.1. dobiveni su spektrogrami skraćenih zvučnih zapisa na temelju kojih je moguće stvoriti pozitivne i negativne parove koji su potrebni za treniranje i evaluaciju mreže za verifikaciju govornika. Zbog ograničenosti sklopovskog okruženja na kojem se trenira mreža, odabire se 600 zvučnih zapisa po govorniku od ukupno 1000 dostupnih. Ovo smanjenje broja dostupnih zvučnih zapisa utječe na broj parova spektrograma. Pozitivni parovi stvaraju se na način da se svi spektrogrami koji su dobiveni iz zvučnih zapisa istog govornika i u kojem se izgovara određena

fraza kombiniraju jedan s drugim. Primjerice, ako postoje dva spektrograma koja su dobivena iz zvučnog zapisa prvog govornika, a izgovara se ista fraza onda je moguće stvoriti pozitivni par spektrograma. Broj stvorenih negativnih parova približno je jednak broju pozitivnih parova. Razlog usklađivanja broja pozitivnih i negativnih parova je smanjenje pretjeranog usklađivanja rada mreže na velikom broju pozitivnih ili negativnih parova. Stvaranje negativnih parova spektrograma odrađeno je na sljedeći način:

- na temelju svih zvučnih zapisa različitih govornika koji izgovaraju istu frazu (primjerice, prvi i drugi govornik izgovaraju frazu „bin blue at“);
- na temelju zvučnih zapisa istog govornika koji izgovara frazu koja ima jednake prve dvije riječi pri čemu mogu postojati najviše 3 takva negativna para (primjerice, prvi govornik izgovara fraze „bin blue at“ i „bin blue by“);
- približno 20% negativnih parova dobiveno je iz zvučnih zapisa istog govornika koji izgovara različite fraze (primjerice, prvi govornik izgovara fraze „bin green at“ i „lay blue in“);
- preostali negativni parovi stvaraju se nasumičnom kombinacijom spektrograma dobivenih iz zvučnih zapisa u kojim frazu ne izgovara isti govornik ili se izgovara različita fraza (primjerice, prvi govornik izgovara frazu „set green at“, a drugi „lay blue in“).

U konačnici to rezultira s ukupno 184588 stvorenih parova spektrograma koji se pohranjuju na tvrdi disk u obliku NumPy datoteke kako bi se mogli naknadno koristiti za potrebe treniranja i evaluacije mreže. Datoteke se pohranjuju pod nazivom koji daje informacije o identitetima govornika i frazama koje oni izgovaraju. Popis indeksa govornika koji određuju identitet govornika prema spolu u GRID skupu podataka dan je u tablici 3.2.

Tablica 3.2. *Prikaz indeksa i broja govornika prema spolu govornika u GRID skupu podataka* [9]

	Spol govornika	
	Muški	Ženski
Indeksi govornika	1001, 1002, 1003, 1005, 1006, 1008, 1009, 1010, 1012, 1013, 1014, 1017, 1019, 1026, 1027, 1028, 1030 i 1032	1004, 1007, 1011, 1015, 1016, 1018, 1020, 1021, 1022, 1023, 1024, 1025, 1029, 1031, 1033 i 1034
Broj govornika	18	16

3.1.3. Podjela skupa podataka parova spektrograma

Prilikom treniranja mreže za verifikaciju govornika potrebno je prvotno podijeliti stvoreni skup podataka parova spektrograma na trening, validacijski i testni skup. Trening skup sadrži parove spektrograma koji služe za treniranje mreže na temelju sličnosti tih parova. Nadalje, validacijski skup služi za podešavanje odgovarajućih hiperparametara mreže tijekom treniranja mreže kao što je stopa učenja i slično, a testni skup služi za evaluaciju dobivene mreže [50].

Podjela GRID skupa podataka ostvarena je na dva načina. Kod prve podjele skupa podataka, zvučni zapisi određenog govornika mogu se pojaviti u trening, validacijskom i testnom skupu. Takav skup je nazvan *skup s kombiniranim podacima*. Međutim, takva podjela može predstavljati problem kod evaluacije mreže jer testni skup vjerojatno sadrži zvučni zapis govornika na kojem je mreža trenirana. Iz tog razloga je učinjena i drugačija podjela gdje skupovi podataka za treniranje, validaciju i testni skup nemaju niti jednog zvučnog zapisa govornika koji im je zajednički. Takav skup je nazvan *skup s razdvojenim podacima*.

Parovi spektrograma kombiniranog skupa podataka su podijeljeni u omjeru 70:10:20 gdje 70% skupa se odvaja na trening, 10% na validacijski i 20 % na testni skup. S druge strane, kod skupa s razdvojenim podacima podjelu nije moguće izvršiti na parovima spektrograma, već je podjela izvršena izravno nad zvučnim zapisima, na način da svi zvučni zapisi određenog govornika moraju biti u određenom skupu. Dijeljenjem zvučnih zapisa govornika na trening, validacijski i testni skup nastoji se održati ravnoteža spolova kao kod GRID skupa podataka. Raspodjela govornika prema spolu i skupu podataka kojem pripada prikazana je u tablici 3.3. U konačnici, za svaki skup zvučnih zapisa govornika su generirani parovi spektrograma na način kao kod skupa s kombiniranim podacima. Pri tome se vodi briga o ujednačenosti pozitivnih i negativnih parova, a generiranje negativnih parova ostvareno je na identičan način kao što je objašnjeno u potpoglavlju 3.1.2..

Tablica 3.3. *Prikaz raspodjele govornika na trening, validacijski i testni skup kod skupa s razdvojenim podacima*

	Skup podataka		
	Trening	Validacijski	Testni
Indeksi govornika	1001, 1002, 1003, 1004, 1005, 1006, 1007, 1008, 1009, 1010, 1011, 1012, 1013, 1014, 1015, 1016, 1017, 1018, 1020, 1021, 1022, 1023, 1024, 1025	1019, 1026, 1029	1027, 1028, 1030, 1032, 1031, 1033, 1034
Broj ženskih govornika	12	1	3
Broj muških govornika	12	2	4

3.2. Vlastiti skup podataka

Jedan od zadataka ovog rada je proširenje postojećeg skupa zvučnim zapisima novih govornika. Postojeći GRID skup podataka snimljen je u optimalnim uvjetima uz smanjenu prisutnost šuma. Razlog proširenja skupa podataka je stvaranje robusnije mreže za verifikaciju govornika koja je u stanju raditi u stvarnom okruženju uz prisutnost šuma, odnosno u neoptimalnim uvjetima rada. Vlastiti skup temelji se na GRID skupu podataka. Osim što su neke fraze preuzete iz GRID skupa podataka, nove fraze su snimljene na engleskom jeziku i formirane su tako da dijelom prate ideju formiranja fraze u GRID skupu podataka. Nove fraze sadrže tri riječi gdje prva riječ predstavlja naredbu (glagol), druga riječ pridjev, a treća predmet (imenicu). Zvučni zapisi su prikupljeni od 12 različitih govornika koji su izgovorili 12 različitih fraza, a svaka od tih fraza je ponovljena 8 puta. To je rezultiralo s 96 zapisa po govorniku i ukupno 1152 zvučnih zapisa. Duljina trajanja zvučnih zapisa varira, no kreće se u rasponu od 1 do 2 sekunde. Skup je ujednačen sa stajališta spolova govornika, sa 6 ženskih i 6 muških govornika. Zvučni zapisi su snimani pomoću mobilnih uređaja i spremljenih u *wav* formatu zvučne datoteke. Popis fraza proširenog skupa prikazan je u tablici 3.4. Skup sadrži 12 fraza, od kojih prvih 6 odgovara frazama iz GRID skupa podataka, dok preostalih 6 predstavlja nove fraze.

Tablica 3.4. *Popis fraza vlastitog skupa podataka*

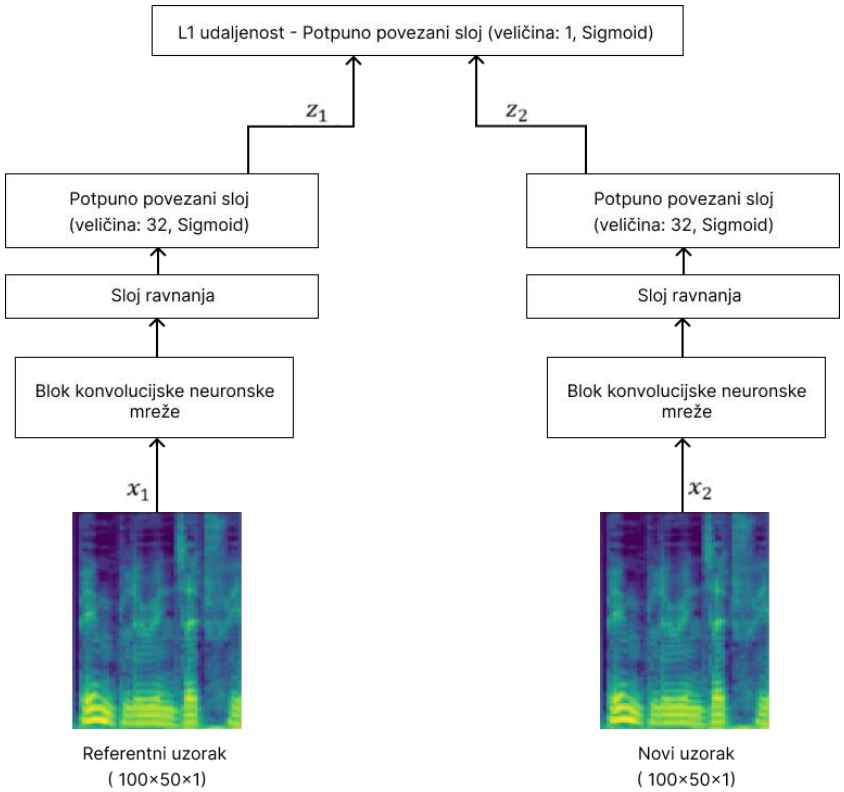
Redni broj fraze	Fraza	Oznaka
1.	*bin blue at*	*BBA*
2.	*bin blue by*	*BBB*
3.	*bin green at*	*BGA*
4.	*bin green by*	*BGB*
5.	*lay blue in*	*LBI*
6.	*lay green in*	*LGI*
7.	*open first door*	*OFD*
8.	*open second door*	*OSD*
9.	*open first window*	*OFW*
10.	*close first door*	*CFD*
11.	*close first window*	*CFW*
12.	*close third window*	*CTW*

3.3. Predložena arhitektura mreže za verifikaciju govornika

Arhitektura sijamske neuronske mreže koja je predložena ovim radom prikazana je na slici 3.1. Iz navedene slike moguće je uočiti grafičku strukturu sijamske neuronske mreže koja koristi dvije grane, odnosno podmreže. Svaka grana sadrži blok konvolucijske neuronske mreže, nakon kojeg slijedi jedan sloj ravnanja (engl. *flatten layer*) i potpuno povezani sloj s aktivacijskom funkcijom Sigmoid. Upravo taj potpuno povezani sloj izlaz je pojedine grane koja daje vektor značajki veličine 32. Na temelju tih vektora iz obiju grana, na slici 3.1. označenih sa z_1 i z_2, računa se mjera udaljenosti prema metrici L1, odnosno Manhattan udaljenosti. Izraz L1 udaljenosti dan je u nastavku:

$$L1 = \sum_{i=1}^{n} |z_{1_i} - z_{2_i}|, \tag{3-1}$$

gdje je parametar n broj elemenata vektora z_1 i z_2, parametar z_{1_i} je *i*-ti element vektora z_1, a z_{2_i} je *i*-ti element vektora z_2 [51]. Dobivena udaljenost predaje se jednom izlazom neuronu. Taj jedan neuron koristi Sigmoid aktivacijsku funkciju kako bi se skalirala vrijednost u rasponu od 0 do 1 (Slika 2.2.). Izlaz mreže daje vjerojatnost sličnosti ulaznih parova spektrograma pri čemu su vrijednosti bliže 1 pokazatelj veće sličnosti, dok vrijednosti bliže 0 ukazuju na različitost.



Slika 3.1. *Prikaz arhitekture sijamske neuronske mreže za verifikaciju govornika predložene ovim radom*

Arhitektura bloka konvolucijske neuronske mreže koja služi za izvlačenje značajki prikazana je tablicom 3.5. Navedenom tablicom prikazan je popis slojeva mreže s odgovarajućim brojem kanala, veličinom jezgre, korakom i parametrom očuvanja dimenzije (engl. *padding*). Svaki od konvolucijskih slojeva koristi ReLu aktivacijsku funkciju. Prva dva sloja konvolucijske mreže su upravo konvolucijski slojevi. Kako bi se smanjile dimenzije slika i ubrzalo treniranje mreže nakon tih slojeva slijedi sloj sažimanja koji primjenjuje sažimanje na temelju maksimalne vrijednosti promatranog prostora dimenzija 2x2.

Tablica 3.5. *Prikaz arhitekture bloka konvolucijske neuronske mreže koji se koristi u sijamskoj neuronskoj mreži prikazane slikom 3.1.*

Sloj	Broj kanala	Veličina filtra (jezgre)	Korak	Očuvanje dimenzije
Konvolucijski sloj – 1	12	3	1	NE
Konvolucijski sloj - 2	24	3	1	DA
Sloj sažimanja prema najvećoj vrijednosti - 1	-	2	2	DA
Konvolucijski sloj - 3	24	3	1	NE
Konvolucijski sloj - 4	36	3	1	DA
Sloj sažimanja prema najvećoj vrijednosti - 2	-	2	2	DA
Konvolucijski sloj - 5	64	2	1	NE
Konvolucijski sloj - 6	128	2	1	NE
Sloj sažimanja prema najvećoj vrijednosti - 3	-	2	2	DA

3.4. Treniranje predložene mreže za verifikaciju govornika

Mreža predložena ovim radom, čija je arhitektura prikazana slikom 3.1., trenira se korištenjem Adam optimizatora, a optimizacija parametara mreže temelji se na minimizaciji funkcije naziva binarni unakrsni entropijski gubitak. Zbog velikog broja i veličine parova spektrograma na kojima se mreža trenira, podaci se dijele na *n* dijelova jednake veličine zvanih *batch* blokovi na način da ih računalo može obraditi. Veličina jednakih dijelova naziva se *batch* veličina. Na dostupnom sklopovlju korištena je *batch* veličina iznosa 512 koja se pokazala najboljom. Korištenjem manjih veličina kao što su 32 ili 64, uočena je pogrešaka pri treniranju mreže odabirom samo pozitivnih ili samo negativnih parova. Iako je *batch* veličina 512, ne znači nužno da svi blokovi imaju 512 parova. Iznimka je moguća kod posljednjeg bloka koji sadrži preostali broj parova. Primjerice, ako skup ima 1000 parova, prvi blok će imati 512, a drugi preostalih 488 parova. Mreža se trenira na svim podacima trening skupa gdje se potpuni prolazak kroz sve podatke u skupu podataka naziva

epoha. Zbog boljih performansi konačne mreže, treniranje mreže provodi se u više epoha. Međutim, ovaj pristup može dovesti do pretjeranog usklađivanja na trening skup podataka. Pretjerano usklađivanje na trening skup podataka predstavlja nepovoljni događaj u kojem mreža ima odlične performanse na trening skupu podataka, a loše na onim podacima na kojima nije trenirana. Time se gubi osnovna ideja strojnog učenja, prema kojoj bi mreža trebala dobro raditi na podacima s kojima se nije susrela [52]. Kao pokazatelji pojave pretjeranog usklađivanja na trening podacima mogu biti vrijednosti gubitka i točnosti na trening i validacijskom skupu podataka. Ako je vrijednost gubitka na trening skupu značajno manja u odnosu na validacijski skup i ako je točnost značajno veća na trening skupu u odnosu na validacijski to upućuje da se mreža pretjerano usklađuje na trening skup podataka [52, 53].

Ovim radom kao jedno rješenje za problem pretjeranog usklađivanja na trening skupu podataka predlaže se izmjena stope učenja tijekom treniranja mreže. U tu svrhu stvoren je algoritam postupnog smanjenja stope učenja tijekom treniranja mreže koji je prikazan programskim kodom 3.2. Navedenim programskim kodom prikazana je klasa koja kroz konstruktor prima parametre: faktor smanjenja stope učenja, minimalna stopa učenja, prag razlike (predstavlja apsolutnu razliku između vrijednosti gubitka na trening i validacijskom skupu) i strpljenje koji govori koliko epoha treba proći do sljedeće promjene stope učenja. Algoritam radi na način da se svake *i*-te epohe izvrši smanjenje stope učenja uzimajući u obzir omjer gubitka na trening i validacijskom skupu. Ako omjer gubitaka trening skupa i validacijskog skupa premašuje postavljeni prag razlike, postavlja se nova vrijednost stope učenja prema danom izrazu:

$$su_{nova} = su_{prethodna} * faktor * \frac{gubitak_{trening}}{gubitak_{val}}, \tag{3-2}$$

gdje je:

- su_{nova} – nova stopa učenja,
- $su_{prethodna}$ – prethodna stopa učenja,
- $faktor$ – faktor smanjenja,
- $gubitak_{trening}$ – gubitak na trening skupu i
- $gubitak_{val}$ – gubitak na validacijskom skupu.

Hiperparametar prethodna stopa učenja u početku treniranja mreže predstavlja početnu stopu učenja koju je potrebno unaprijed definirati. Osim početne stope učenja, unaprijed je potrebno definirati hiperparametre koji se predaju objektu klase prikazanog programskim kodom 3.2., a to su faktor smanjenja, strpljenje i prag razlike gubitka na trening i validacijskom skupu.

```
class DynamicLearningRateScheduler(Callback):

    def __init__(self, factor=0.65, patience=6, threshold=0.1, minLr=1e-5):
        super(DynamicLearningRateScheduler, self).__init__()
        self.factor, self.patience, self.threshold = factor, patience, threshold
        self.minLr, self.wait, self.bestDivergence = minLr, 0, float('inf')
    def on_epoch_end(self, epoch, logs=None):
        trainLoss = logs.get('train_loss')
        valLoss = logs.get('val_loss')
        if trainLoss is not None and valLoss is not None:
            divergence = abs(trainLoss - valLoss)
            if divergence < self.bestDivergence - self.threshold:
                self.bestDivergence = divergence
                self.wait = 0
            else:
                self.wait += 1
                if self.wait >= self.patience:
                    oldLr = float(K.get_value(self.model.optimizer.lr))
                    if oldLr > self.min_lr:
                        newLr = oldLr * self.factor * trainLoss/valLoss
                        newLr = max(newLr, self.minLr)
                        K.set_value(self.model.optimizer.lr, newLr)
                        if self.verbose > 0:
                        self.wait = 0
```

Programski kôd 3.2. *Prikaz programskog koda klase koja pruža funkciju za izmjenu stope učenja tijekom treniranja mreže*

Dodatna tehnika koja je u radu korištena za sprječavanje pretjeranog usklađivanja na trening podatke je augmentacija podataka. Drugi razlog primjene augmentacije je dobivanje mreža sposobnih za rad u stvarnom okruženju. U ovom radu se prilikom treniranja predložene mreže koriste dva tipa augmentacije po uzoru na rad [15], a to su Gaussov šum i amplitudna augmentacija spektrograma. Gaussov šum jedan je od najpoznatijih šumova koji se primjenjuje u augmentaciji podataka pri postupku treniranja dubokih neuronskih mreža. Konkretno, na GRID skupu podataka Gaussov šum umjetno je dodan zvučnim zapisima na temelju kojih su izrađeni spektrogrami. S druge strane, amplitudna augmentacija spektrograma primjenjuje se izravno na spektrogramima gdje se radi se o promjeni intenziteta frekvencije, tj. amplitude spektrograma.

Postupak augmentacije temeljen na Gaussovom šumu je sljedeći. Odabranom zvučnom zapisu dodaje se Gaussov šum sa srednjom vrijednosti jednakom 0, a iznos standardne devijacije se nasumično odabire u rasponu od 0.0003 do 0.002. Programsko rješenje za umjetno dodavanje Gaussovog šuma zvučnom zapisu prikazano je programskim kodom 3.3. Za ostvarenje nasumičnog odabira vrijednosti standardne devijacije i nasumične vrijednosti normalne razdiobe koristi se NumPy biblioteka.

```
def applyRandomGaussianNoiseToAudio(audio, mean=0.0, min_std=0.0003,
max_std=0.002):
    std = np.random.uniform(min_std, max_std)
    noise = np.random.normal(mean, std, audio.shape).astype(np.float32)
    noisy_audio = audio + noise
    return noisy_audio, std
```

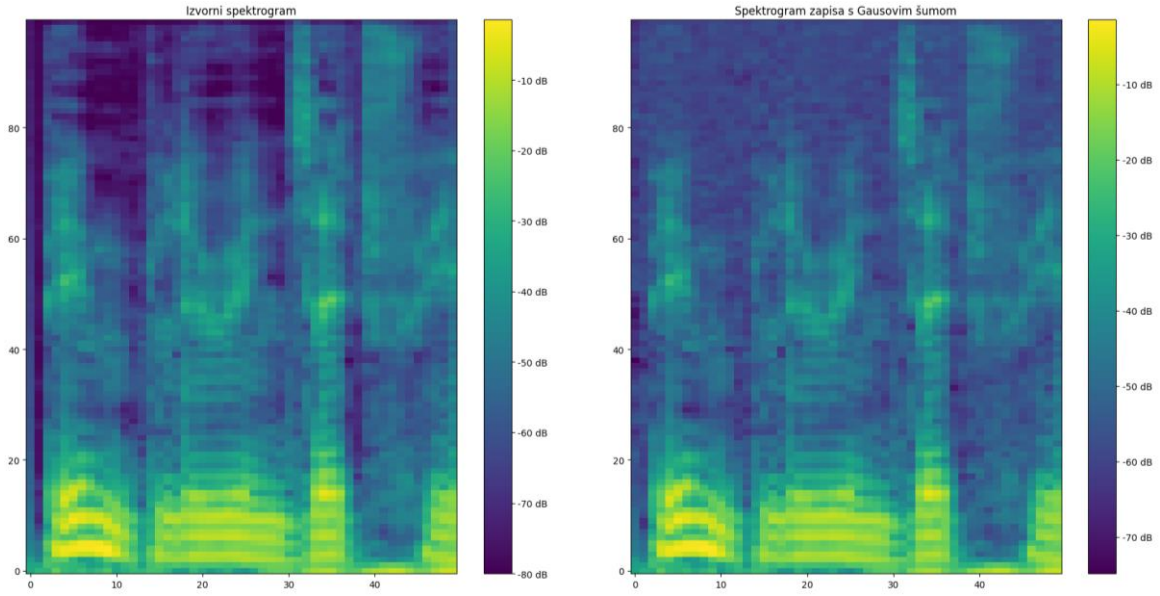
Programski kôd 3.3. *Prikaz programskog koda funkcije za dodavanje Gussovog šuma sa srednjom vrijednosti jednakom 0 i s nasumičnom vrijednosti standardne devijacije u rasponu od 0.0003 do 0.002*

Amplitudna augmentacija spektrograma vrši se promjenom intenziteta frekvencije, odnosno amplitude spektrograma. Budući da su vrijednosti piksela spektrograma zapravo iznosi intenziteta frekvencije, a nalaze se u rasponu od 0 dB do -80 dB gdje 0 dB predstavlja iznos s najvećom amplitudom. Promjenu je moguće izravno izvršiti na spektrogramu. Povećanje amplitude je izvršeno dijeljenjem pojedinačne vrijednosti s faktorom amplitudne augmentacije. Razlog tome je decibelna ljestvica gdje negativne vrijednosti koje su bliže nuli su zapravo veće amplitude (vrijednost -2 dB predstavlja veću amplitudu od -4 dB). Budući da se vrijednosti logaritamskog Mel spektrograma nalaze u rasponu od 0 do -80, osigurava se očuvanje vrijednosti unutar ovog raspona nakon augmentacije. Korišten je raspon faktora pojačanja od 0.9 (predstavlja smanjenje amplitude) do 1.2 (predstavlja povećanje amplitude). Programsko rješenje za primjenu amplitudne augmentacije spektrograma na predani spektrogram prikazano je programskim kodom 3.4.

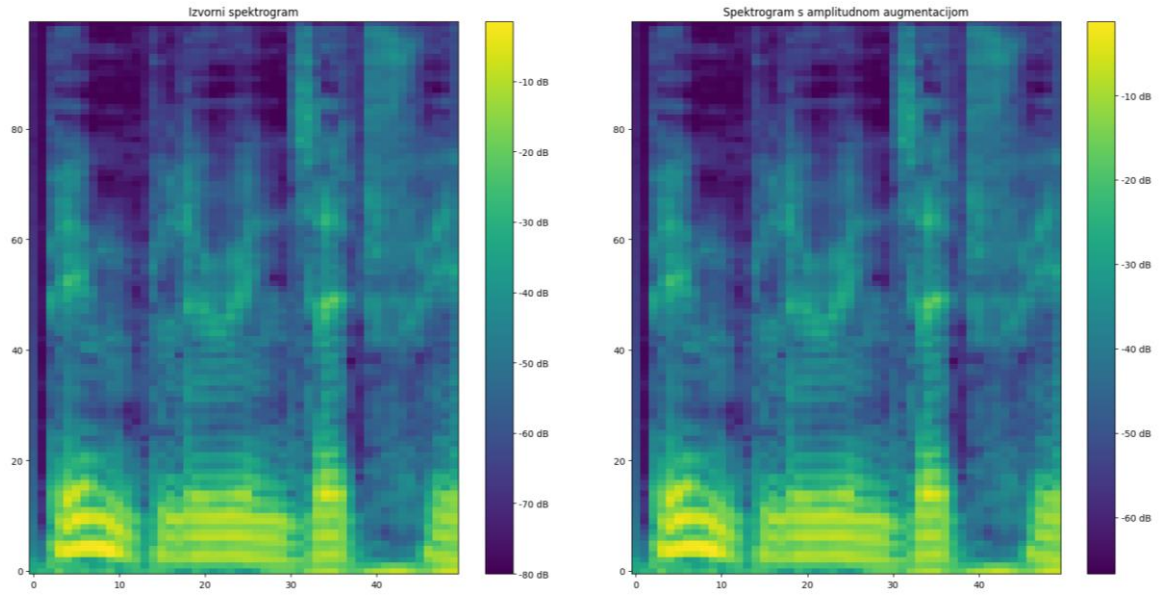
```
def apply_random_amplitude_scaling_to_spectrogram(spectrogram, min_scale=0.9,
max_scale=1.2):
    scale_factor = np.random.uniform(min_scale, max_scale)
    scaled_spectrogram = spectrogram / scale_factor
    scaled_spectrogram = np.clip(scaled_spectrogram, -80, 0)
    return scaled_spectrogram
```

Programski kôd 3.4. *Prikaz programskog koda funkcije za dodavanje amplitudne augmentacije spektrogramu s nasumičnom vrijednosti faktora pojačanja u rasponu od 0.9 do 1.2*

Umjetno dodavanje Gaussovog šuma i amplitudna augmentacija spektrograma izvršeni su zasebno na cijelom trening skupu podataka. Cilj je izraditi odvojene mreže i utvrditi utjecaj šuma i amplitudne augmentacije spektrograma na rad konačne mreže. Augmentacije su izvršene na skupu s kombiniranim i razdvojenim podacima. Slikama 3.2. i 3.3. prikazana je usporedba izvornog spektrograma, odnosno spektrograma na kojem nije primijenjena augmentacija, sa spektrogramom koji primjenjuje augmentaciju.



Slika 3.2. *Prikaz usporedbe izvornog spektrograma i spektrograma dobivenog iz zvučnog zapisa na koji je umjetno dodan Gaussov šum sa standardnom devijacijom vrijednosti 0.00129*



Slika 3.3. *Prikaz usporedbe izvornog spektrograma i spektrograma na koji je primijenjena amplitudna augmentacija s faktorom pojačanja vrijednosti 1.2*

U nastavku ovog poglavlja navedene su predložene mreže trenirane na skupa podataka s kombiniranim i razdvojenim podacima. Nakon treniranja tih mreža i korištenja istih, mehanizmom prijenosnog učenja treniraju se nove mreže na vlastitom skupu podataka.

3.4.1. Treniranje predložene mreže na temelju skupa s kombiniranim podacima

Treniranje mreže na temelju skupa s kombiniranim podacima provedeno je u 80 epoha. Treniranje je provedeno na skupu parova spektrograma bez primjene augmentacije podataka koji se naziva osnovni skup podataka kao i na one s primijenjenom augmentacijom pomoću Gaussovog šuma ili

amplitudne augmentacije spektrograma. Na taj način dobivaju se tri različite mreže od kojih je svaka izrađena na različitom skupu podataka. Vrijednosti hiperparametara treniranja mreže koji se predaju objektu klase prikazane programskim kodom 3.2., a o kojima ovisi brzina promjene stope učenja tijekom treniranja mreže podešavaju se unaprijed. Vrijednosti hiperparametara za svaku mrežu, tj. početne stope učenja, faktora smanjenja i strpljenja prikazani su u tablici 3.6. Vrijednost hiperparametra minimalna brzina učenja postavljena je na 10^{-6} i prag razlike na 0.1.

Tablica 3.6. *Prikaz hiperparametara početne stope učenje, faktora smanjenja, i strpljenja kod treniranja mreža s kombiniranim podacima na skupovima bez i s primjenom augmentacije*

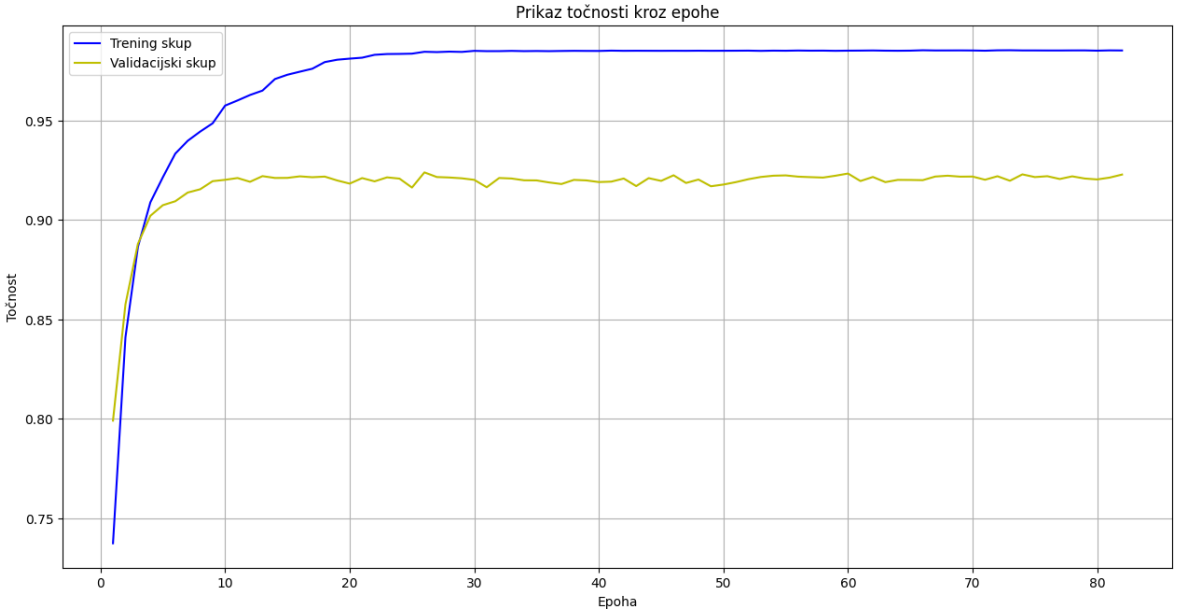
Augmentacija	Početna stopa učenja	Faktor smanjena	Strpljenje
Bez augmentacije	0.0008	0.9	3
Gaussov šum	0.000074	0.91	3
Amplitudna augmentacija spektrograma	0.00007	0.91	4

Treniranjem kroz epohe odabire se mreža na onoj epohi gdje je ostvarila najbolje rezultate točnosti i gubitka na validacijskom skupu. Tako se kao najbolja mreža koja je trenirana na podacima bez augmentacije pokazala ona na 62. epohi, dok je mreža s primijenjenom augmentacijom pomoću Gaussovog šuma najbolji rezultat pokazuje na 15. epohi. Mreža trenirana na podacima koji primjenjuju amplitudnu augmentaciju spektrograma najbolje rezultate ostvaruje na 26. epohi. Među tim mrežama najbolje rezultate pokazuje mreža trenirana na skupu koji primjenjuje amplitudnu augmentaciju spektrograma.

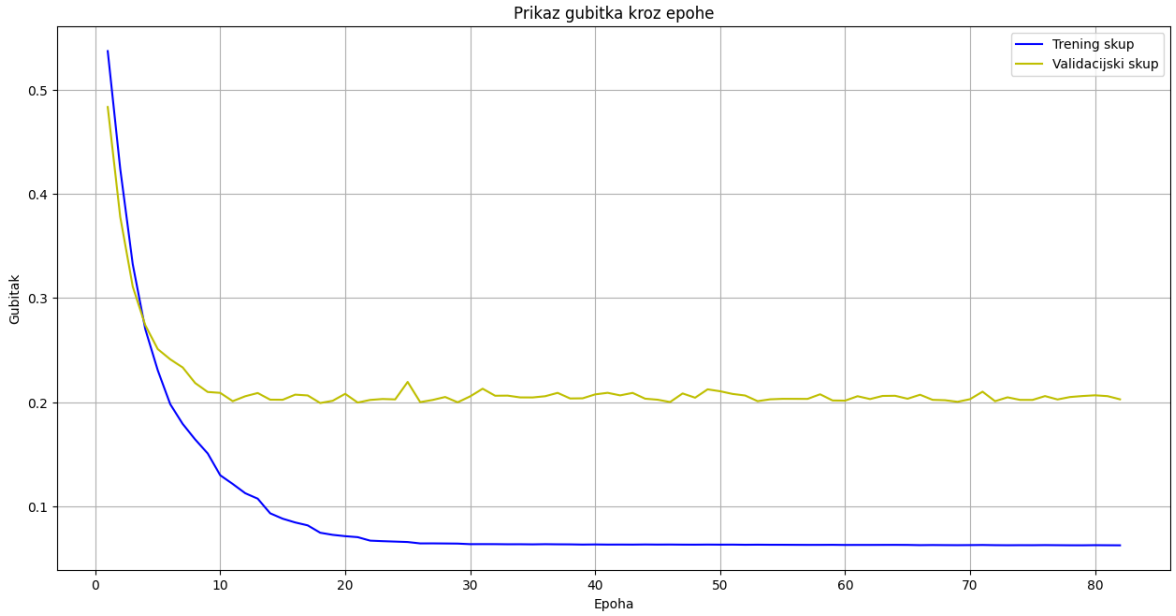
Za ilustraciju tijeka treniranja neuronske mreže odabran je slučaj gdje je primijenjena amplitudna augmentacija spektrograma. Slikom 3.4. prikazan je graf točnosti kroz epohe za mrežu treniranu na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Vidljivo je kako točnost na validacijskom skupu i trening skupu raste jednakom brzinom u prve 4 epohe, nakon čega točnost na validacijskom skupu počinje stagnirati. Točnost na trening skupu raste do 30. epohe gdje kreće stagnirati. U oba slučaja točnost mreže je preko 90%, ali ona na trening skupu je veća u odnosu na validacijski skup.

Slikom 3.5. prikazan je graf gubitka mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Iz grafa je vidljivo kako u prvim epohama gubitak naglo opada, ali već i pri niskim epohama dolazi do stagniranja stope učenja na validacijskom skupu i povećanja razlike između gubitka na trening i validacijskom skupu podataka. Gubitak trening

skupa manji je u odnosu na validacijski za vrijednost veću od 0.1, što uz razliku točnosti prikazane na slici 3.4., može biti pokazatelj pretjeranog usklađivanja na trening skup podataka.



Slika 3.4. *Prikaz točnosti kroz epohe na trening i validacijskom skupu podataka mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*



Slika 3.5. *Prikaz gubitka kroz epohe na trening i validacijskom skupu mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*

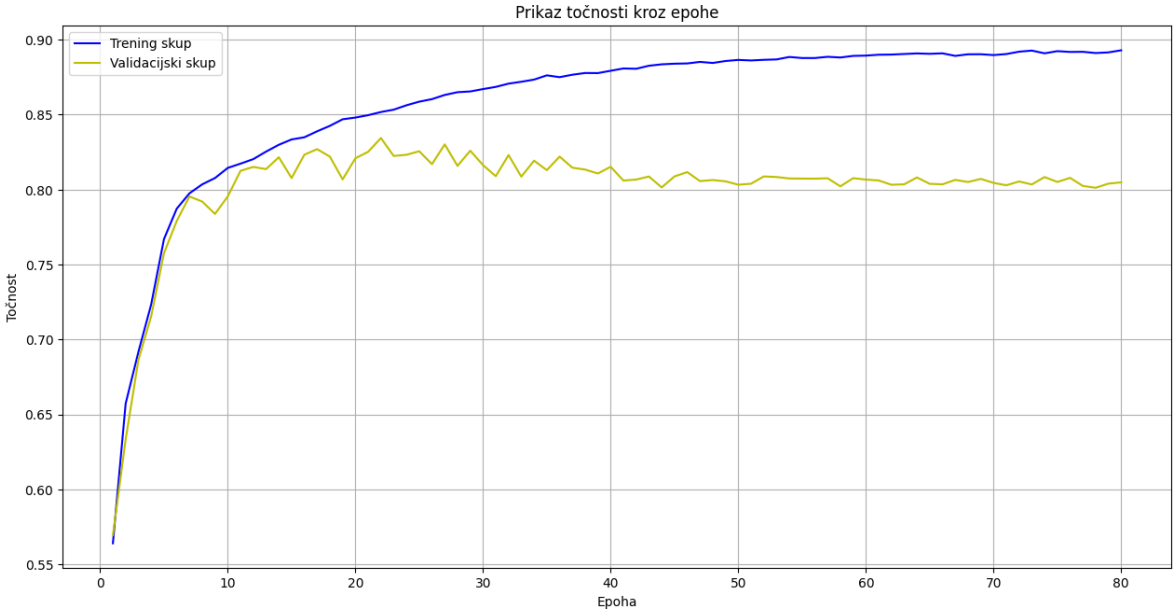
3.4.2. Treniranje predložene mreže na temelju skupa s razdvojenim podacima

Treniranje mreže na temelju skupa s razdvojenim podacima provedeno je u 80 epoha. Vrijednosti hiperparametara prikazane su u tablici 3.7., a vrijednost hiperparametra minimalne brzine učenja postavljena je na 10^{-6} i prag razlike na 0.1.

Tablica 3.7. *Prikaz hiperparametara početne stope učenje, faktora smanjenja, i strpljenja kod treniranja mreža s razdvojenim podacima na skupovima bez i s primjenom augmentacije*

Augmentacija	Početna stopa učenja	Faktor smanjena	Strpljenje
Bez augmentacije	0.00005	0.71	5
Gaussov šum	0.00005	0.8	8
Amplitudna augmentacija spektrograma	0.00005	0.8	8

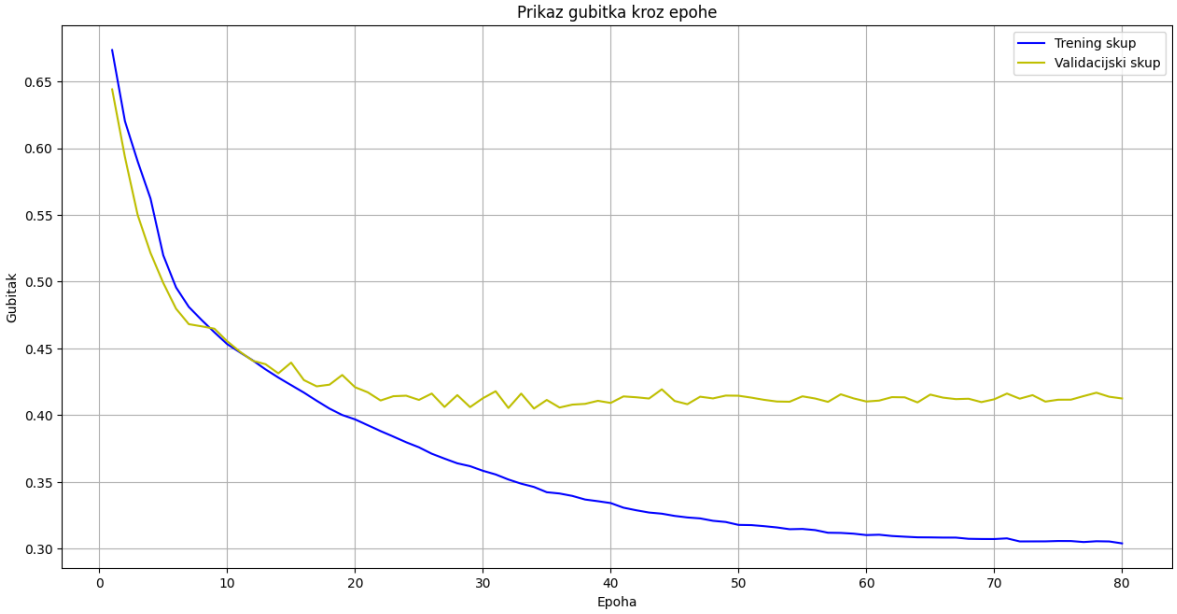
Na temelju rezultata točnosti i gubitka na validacijskom skupu kao najbolja mreža među razdvojenim podacima se pokazala mreža trenirana na podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Za tu mrežu su u nastavku prikazani grafovi točnosti i gubitka kroz epohe na trening i validacijskom skupu podataka. Na slici 3.6. prikazan je graf točnosti kroz epohe navedene mreže. U usporedbi sa grafom točnosti prikazanim slikom 3.4., koji prikazuje rezultat mreže trenirane na skupu s kombiniranim podacima, vidljiva je manja točnost mreže. Navedeno je dijelom posljedica smanjenja početne stope učenja jer postavljanjem veće stope učenja dolazi do izraženijeg pretjeranog usklađivanja na trening podacima već pri početnim epohama. Manja točnost mreže trenirane na skupu s razdvojenim podacima, očekivana je zbog načina podjele skupa podataka, budući da u validacijskom skupu nema nijednog zvučnog zapisa govornika iz trening skupa.



Slika 3.6. *Prikaz točnosti kroz epohe na trening i validacijskom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*

Na slici 3.7. prikazan je graf gubitka kroz epohe. Iz navedene slike uočava se kako smanjenje gubitka nije izraženo kao kod mreže trenirane na kombiniranim podacima (Slika 3.5.). Navedeno

je posljedica smanjenja stope učenja. Stopa učenja smanjena je zbog značajne razlike između iznosa gubitaka na trening i validacijskom skupu.



Slika 3.7. *Prikaz gubitka kroz epohe na trening i validacijskom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*

3.4.3. Treniranje mreže prijenosnim učenjem

Prijenosno učenje (engl. *transfer learning*) je proces gdje se već istrenirana mreža koristi kao osnova za treniranje nove mreže u cilju skraćenja vremena treniranja i poboljšanja točnosti mreže [54]. Takvu istreniranu mrežu moguće je nazvati izvornom mrežom. U ovom slučaju izvorne mreže su one koje su trenirane na GRID skupu podataka, na temelju kojih se dobivaju nove mreže treniranjem na vlastitom skupu podataka. Vlastiti skup dijeli se samo na način *razdvojenih podataka*, odnosno zvučni zapisi određenog govornika nalaze se samo u jednom od sljedećih skupova podataka: trening, validacijskom ili testnom. Drugim riječima, nije moguće da se zvučni zapisi određenog govornika nalaze i u trening i validacijskom skupu ili testnom skupu. Oznake zvučnog zapisa, odnosno parova spektrograma odgovaraju onima u GRID skupu podataka. Razlika je u indeksu govornika gdje u GRID skupu indeksi kreću početnom oznakom „1“ (primjerice, „1001“), a u vlastitom skupu kreću s „2“ (primjerice, „2001“). Raspored zvučnih zapisa odgovarajućih govornika prikazan je tablicom 3.8. Skup podataka organiziran je tako da se zvučni zapisi 7 govornika koriste za stvaranje parova spektrograma trening skupa, 2 govornika za validacijski skup, a 3 govornika za testni skup.

Tablica 3.8. *Prikaz raspodjele zvučnih zapisa govornika iz vlastitog skupa podataka na trening, validacijski i testni skup*

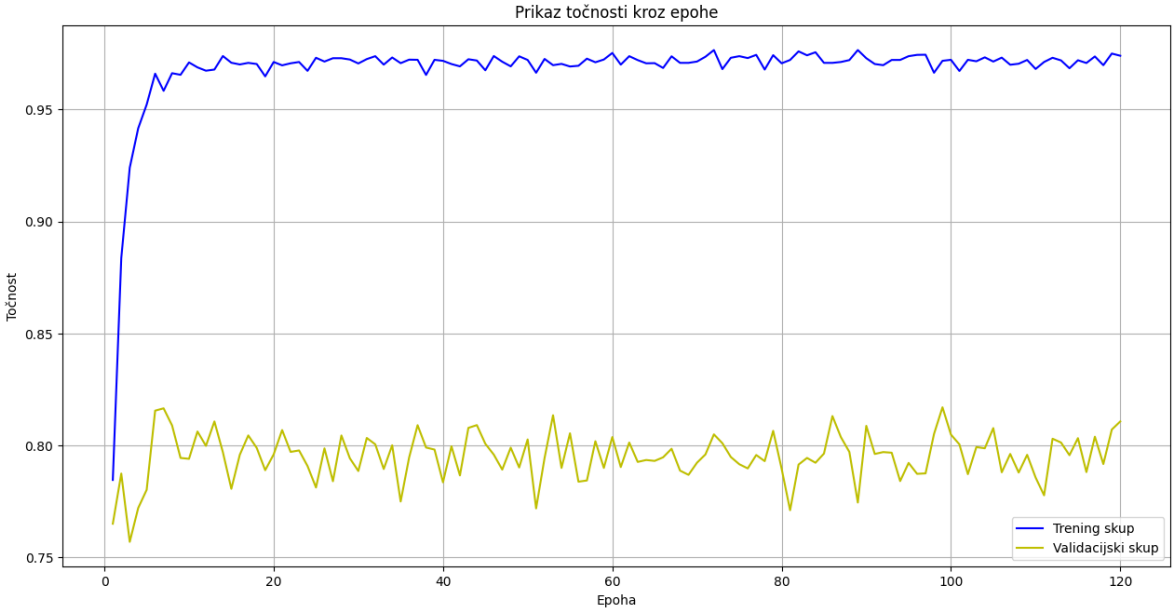
	Skup podataka		
	Trening skup	Validacijski skup	Testni skup
Indeksi govornika	2002, 2003, 2004, 2005, 2006, 2007, 2011	2008, 2009	2001, 2010, 2012
Broj ženskih govornika	3	1	2
Broj muških govornika	4	1	1

Mreže koje se treniranju procesom prijenosnog učenja trenirane su u 120 epoha na temelju svih 6 izvornih mreža dobivenih treniranjem na GRID skupu podataka. Na temelju tih mreža dobiva se novih 6 mreža koje su trenirane na vlastitom skupu podataka. Nakon treniranja tih mreža, odabiru se one mreže koji su najbolje među kombiniranim i razdvojenim podacima. Kao i kod treniranja mreže na GRID skupu, potrebno je unaprijed definirati hiperparametre kao što su početna stopa učenja, faktor smanjenja i slično. Sljedećom tablicom su prikazani korišteni hiperparametri, a minimalna stopa učenja je 10^{-6} i prag razlike je 0.1.

Tablica 3.9. *Hiperparametri nove mreže trenirane na vlastitom skupu podataka na temelju izvornih mreža GRID skupa podataka*

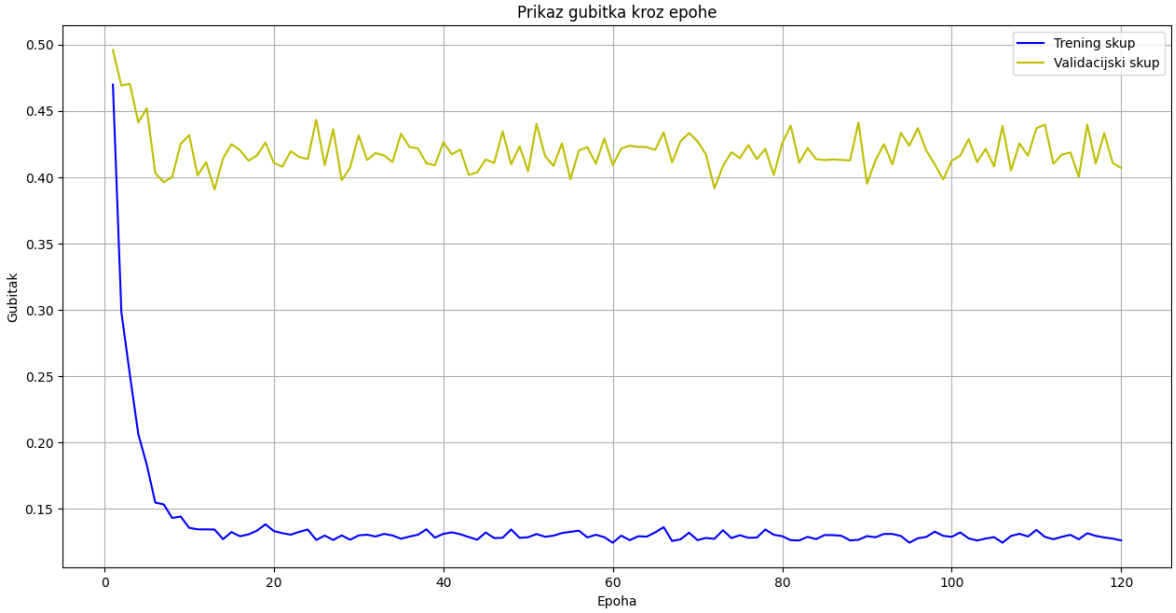
Izvorna mreža	Početna stopa učenja	Faktor smanjena	Strpljenje
Mreža s kombiniranim podacima – Bez augmentacije	0.001	0.9	2
Mreža s kombiniranim podacima – Gaussov šum	0.001	0.98	2
Mreža s kombiniranim podacima – Amplitudna augmentacija spektrograma	0.001	0.98	3
Mreža s razdvojenim podacima – Bez augmentacije	0.00005	0.98	6
Mreža s razdvojenim podacima – Gaussov šum	0.00005	0.98	6
Mreža s razdvojenim podacima - Amplitudna augmentacija spektrograma	0.00005	0.98	6

Prema vrijednostima točnosti i gubitka na validacijskom skupu spektrograma dobivenih na vlastitom skupu podataka odabiru se najbolje mreže. Kao najbolja mreža pokazala se ona dobivena prijenosnim učenjem na temelju izvorne mreže trenirane na skupu s kombiniranim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma. Na slici 3.8. je prikazan graf točnosti kroz epohe na trening i validacijskom skupu navedene najbolje mreže. Mreža na validacijskom skupu ostvaruje točnost koja varira, ali je približno 80%, dok je na trening skupu preko 95%.



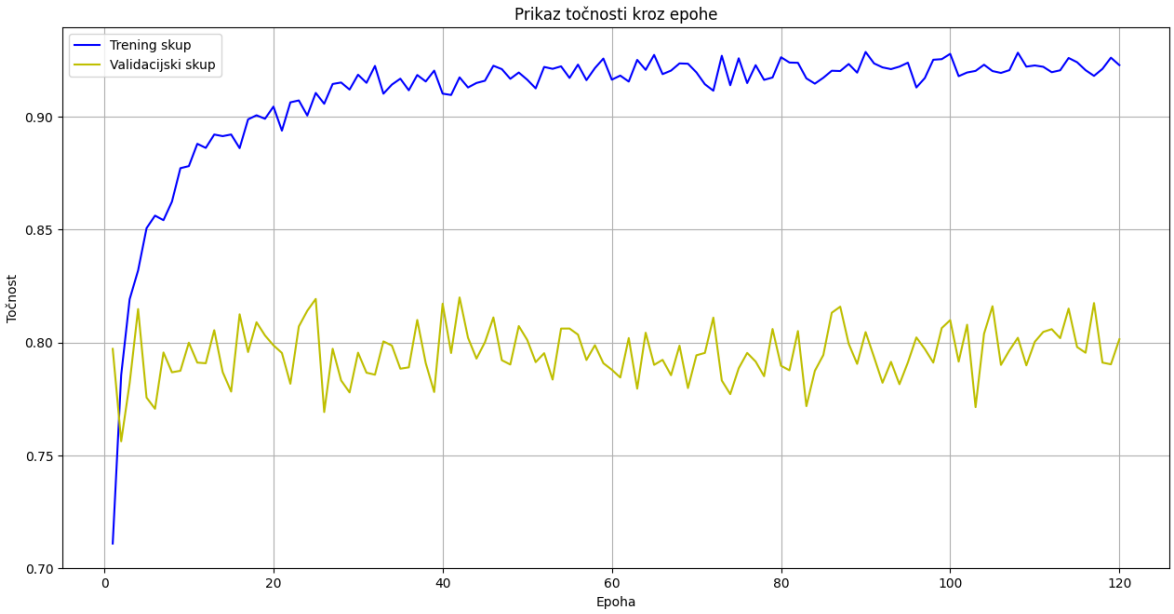
Slika 3.8. *Prikaz točnosti kroz epohe na trening i validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum*

Na slici 3.9. je prikazan graf gubitka kroz epohe na trening i validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum. Na validacijskom skupu ne dolazi do velikog smanjenja gubitka kao kod trening skupa. Na temelju navedene slike i slike 3.8, zbog velike razlike gubitka i točnosti, takvog da je gubitak na trening skupu značajno manji u odnosu na validacijski i točnost značajno veća, zaključuje se da dolazi do pretjeranog usklađivanja na trening skup podataka.

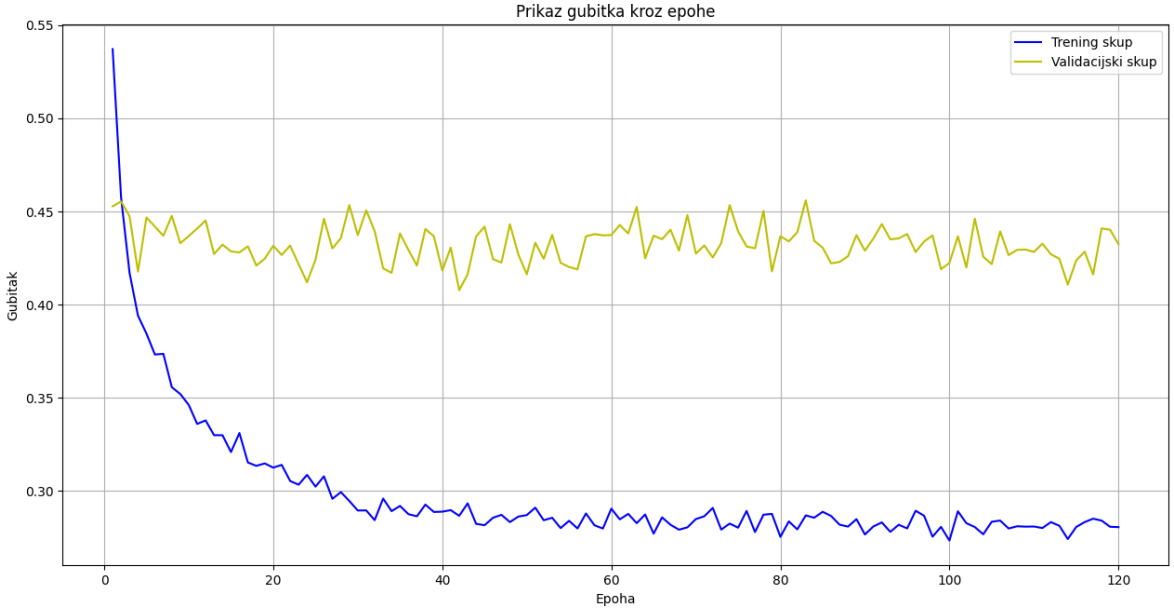


Slika 3.9. *Prikaz gubitka kroz epohe na trening i validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum*

Kod mreža treniranih na temelju izvornih mreža s razdvojenim podacima najbolja se pokazala ona mreža dobivena prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum. Na slici 3.10. je prikazan graf točnosti kroz epohe na trening i validacijskom skupu navedene mreže. S druge strane, slika 3.11. prikazuje gubitak kroz epohe na trening i validacijom skupu podataka navedene mreže.



Slika 3.10. *Prikaz točnosti kroz epohe na trening i validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum*



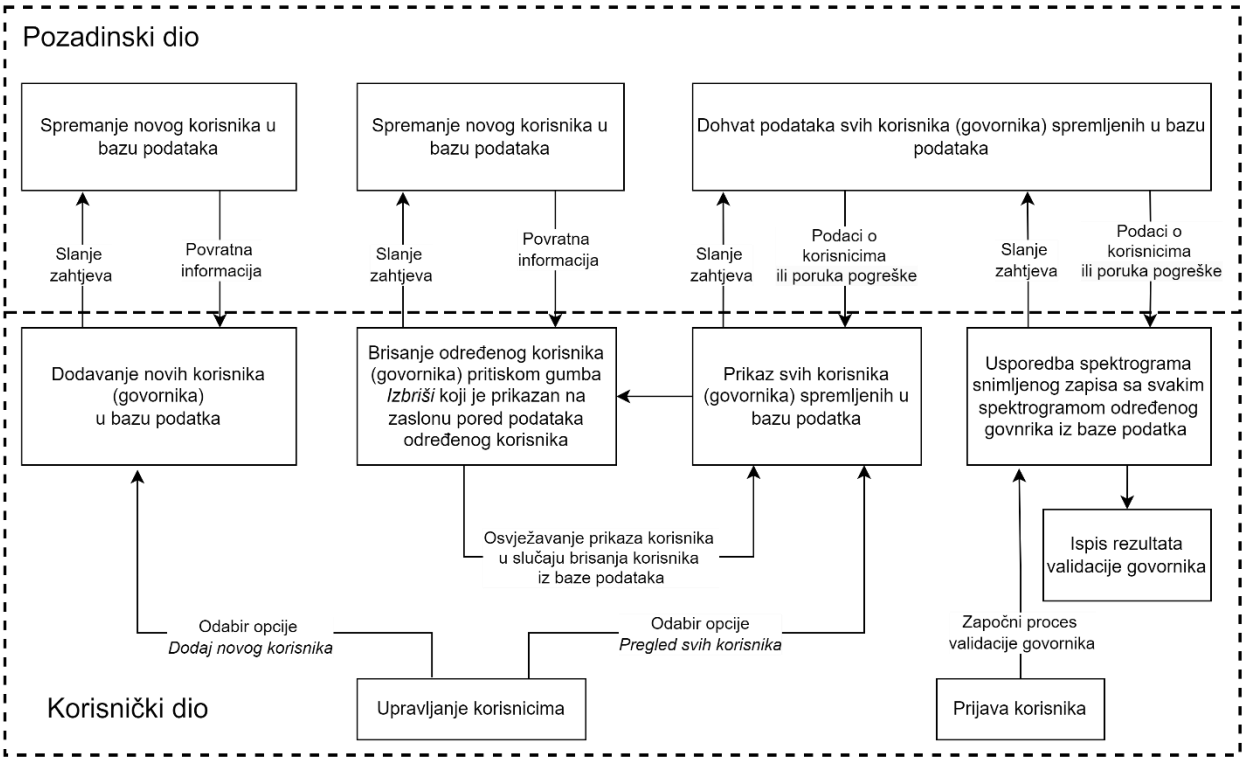
Slika 3.11. *Prikaz gubitka kroz epohe na trening i validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum*

U usporedbi s mrežom dobivenoj na temelju izvorne mreže trenirane na skupu s kombiniranim podacima (Slika 3.8.), točnost na trening skupu je manja, ali na validacijskom je približno ista.

Odnosno, dolazi do smanjenja razlike točnosti između trening i validacijskog skupa. Iz slike 3.11. vidljiva je smanjena razlika gubitaka na trening i validacijskom skupu podataka u odnosu na mrežu s izvornom mrežom treniranom na kombiniranim podacima (Slika 3.9.), ali je razlika gubitaka i dalje prisutna.

3.5. Izrada aplikacije za verifikaciju govornika

U okviru rada izrađena je i aplikacija za verifikaciju govornika na temelju izgrađenih neuronskih mreža. Aplikacija je implementirana u Python3 programskom jeziku i sastoji se od korisničkog i pozadinskog dijela. U korisničkom dijelu pruža se mogućnost postupka verifikacije govornika, ali dodavanja, brisanja i pregleda svih govornika dodanih u bazu podataka. Pozadinski dio programske podrške bavi se logikom dohvata i pohrane podataka korisnika u bazu podataka. Na slici 3.12. prikazana je povezanost pozadinskog i korisničkog dijela aplikacije za verifikaciju govornika kroz funkcijski blok dijagram.



Slika 3.12. *Prikaz funkcijskog blok dijagrama pozadinskog i korisničkog dijela aplikacije za verifikaciju govornika s funkcijama koje omogućuju i njihovu međusobnu interakciju*

Za potrebe simulacije, baza podataka je pokrenuta na računalu s Linux operacijskim sustavom, a izvodi se pomoću MariaDB poslužitelja. Radi se o MySQL bazi podataka nazvanoj *speaker_verification* koja sadrži tablicu *govornici*. Navedena tablica predstavlja entitet govornika čije informacije se unose u bazu podataka. Entitet *govornici* sadrži atribute:

- *id* – jedinstvena oznaka cjelobrojnog tipa podataka, predstavlja primarni ključ,
- *ime* – ime govornika, tip podatka je zapis niza znakova,
- *prezime* – prezime govornika, tip podatka je zapis niza znakova i
- *spektrogram_url* – putanja do datoteke koja predstavlja spektrogram referentnog uzorka govornika, tip podatka je zapis niza znakova.

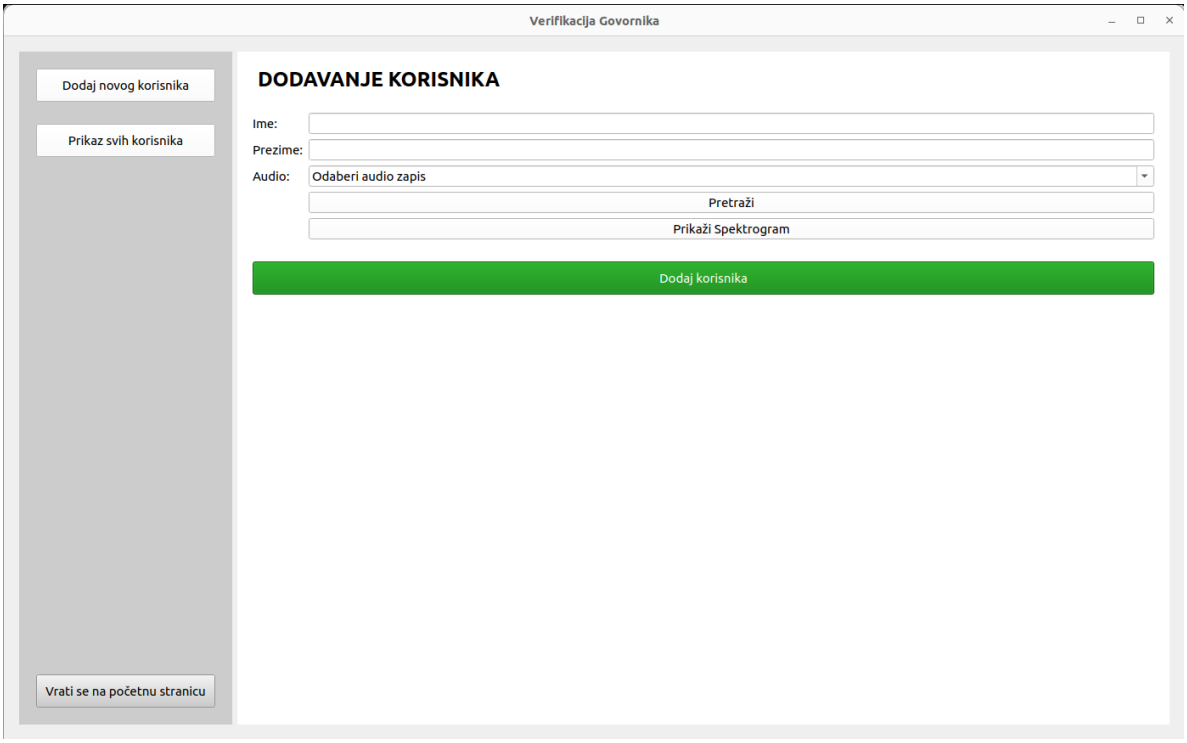
O logici povezivanja na bazu podataka i dohvatu podataka brigu vodi Python skripta koja svoj rad ostvaruje pomoću funkcija pruženih u MySQLdb biblioteci. Primjer korištenja jedne takve funkcije prikazan je programskim kodom 3.5. Navedenim programskim kodom prikazana je funkcija *dohvati_govornike* unutar koje se pozivaju funkcije za povezivanje na bazu i dohvat svih govornika iz tablice *govornici*.

```
def dohvati_govornike():
    try:
        vezaNaBazu = MySQLdb.connect(**db_config)
        kursorBaze = vezaNaBazu.cursor(MySQLdb.cursors.DictCursor)
        kursorBaze.execute("SELECT * FROM govornici")
        govornici = kursorBaze.fetchall()
        return govornici
    except Exception as e:
        print(f"Greska pri dohvatu govornika: {str(e)}")
        return []
    finally:
        kursorBaze.close()
        vezaNaBazu.close()
```

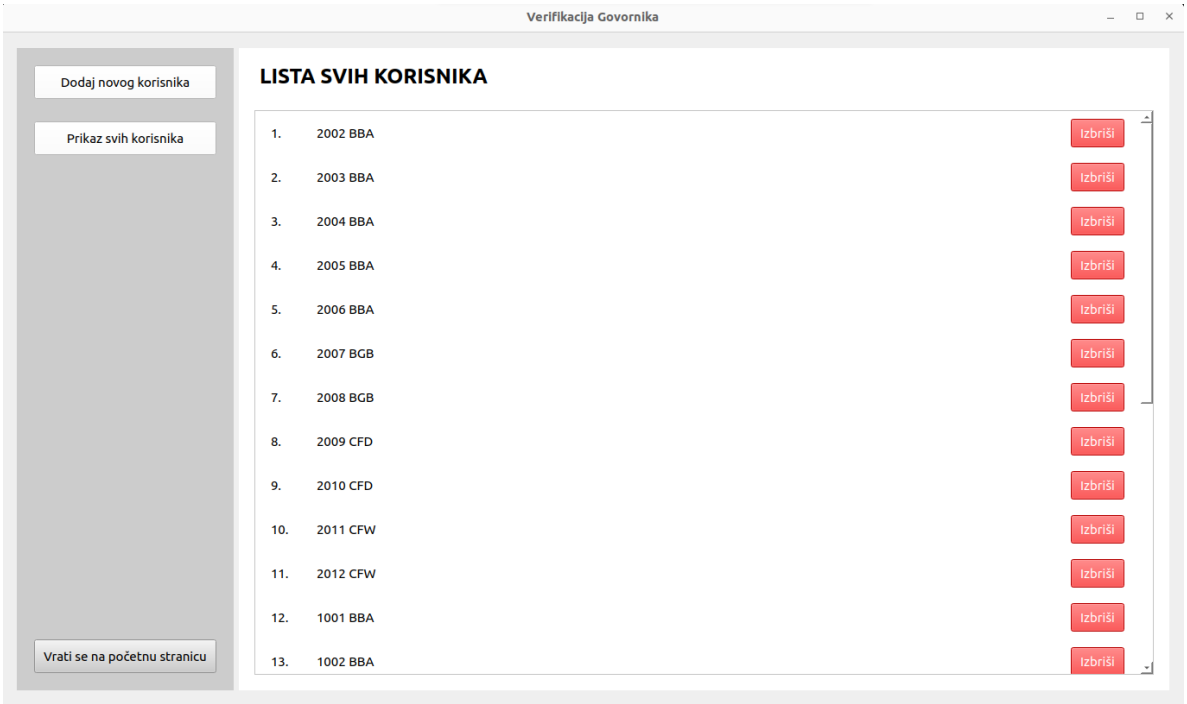
Programski kôd 3.5. *Prikaz programskog koda za dohvat svih govornika iz baze podataka*

S druge strane, korisnički dio programske podrške predstavlja grafičko sučelje koje korisniku aplikacije pruža dvije osnovne mogućnosti: upravljanje korisnicima i prijava korisnika. Upravljanje korisnicima uključuje funkcionalnost pohrane novih govornika (Slika 3.13.) te pregleda i brisanja postojećih korisnika pohranjenih u bazu podataka (Slika 3.14.). Druga mogućnost, prijava korisnika je pak glavna funkcionalnost aplikacije, odnosno postupak verifikacije govornika (Slika 3.15.). Proces verifikacije govornika unutar aplikacije izvodi na način objašnjen u nastavku. Korisnik aplikacije pritiskom na gumb „Započni snimanje“ počinje snimanje zvučnog zapisa i treba izgovoriti odgovarajuću frazu. Snimanje uzorka se zaustavlja nakon isteka 5 sekundi ili pritiskom gumba „Zaustavi snimanje“. Nakon toga, snimljeni zvučni zapis se obrađuje korištenjem funkcije prikazane programskim kodom 3.1. Funkcija izdvaja dio snimljenog zvučnog zapisa u kojem je detektirana pojava prve tri riječi i na temelju tog zvučnog

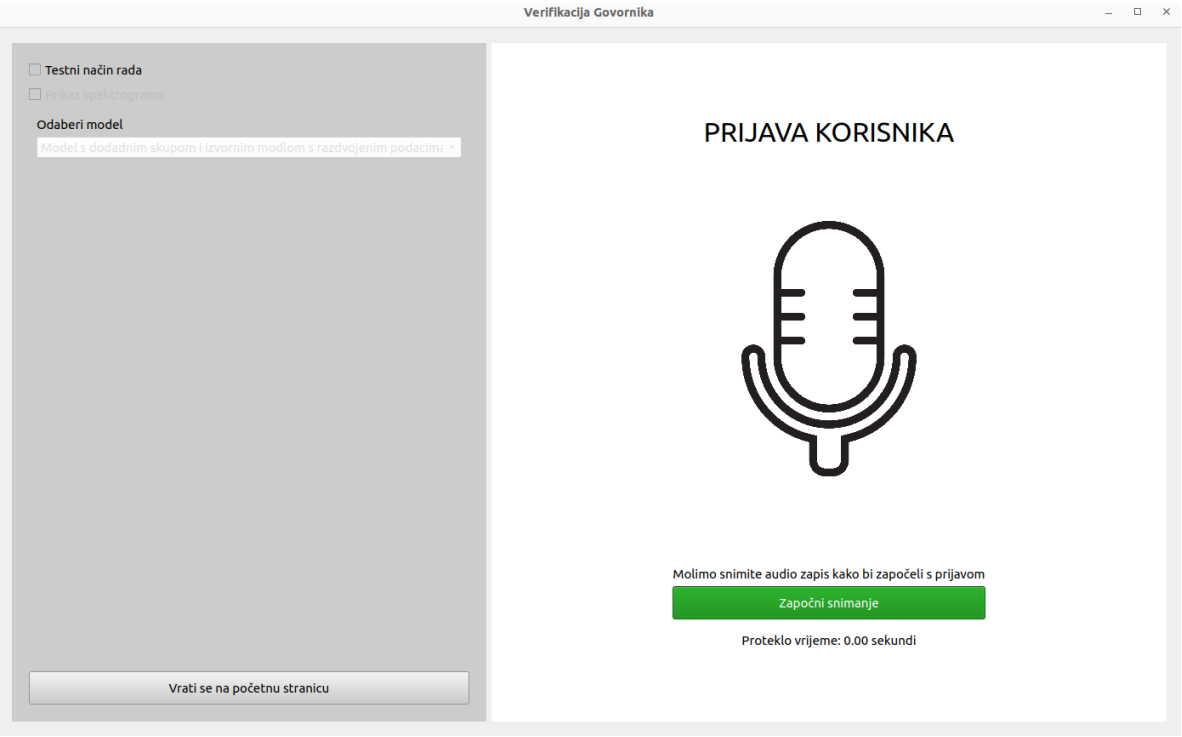
zapisa stvara se logaritamski Mel spektrogram. Budući da nije poznat identitet govornika, s dobivenim spektrogramom rade se parovi sa svim spektrogramima govornika pohranjenih u bazu. Takvi parovi spektrograma predaju se kao ulaz mreže za verifikaciju govornika s ciljem donošenja zaključka na temelju sličnosti predanih ulaza. Onaj govornik čiji spektrogram ima najveću sličnost sa spektrogramom snimljenog zvučnog zapisa i čija je vjerojatnost predikcije sličnosti mreže iznad određenog praga odluke, odabire se kao rezultat. Unutar dijela aplikacije prepoznavanja govornika, moguće je odabrati testni način rada koji pruža dodatne funkcionalnosti izbora mreže kojom se vrši predikcija i prikaz spektrograma snimljenog zvučnog zapisa. Grafičko sučelje programske podrške oslanja se na biblioteku PyQt5, a snimke zaslona aplikacije moguće je vidjeti u nastavku.



Slika 3.13. *Prikaz snimke zaslona prozora aplikacije za pohranu novog korisnika (govornika) u bazu podataka*



Slika 3.14. *Prikaz snimke zaslona prozora aplikacije za pregled svih korisnika u bazi podataka*



Slika 3.15. *Prikaz snimke zaslona prozora prijave korisnika aplikacije za verifikaciju govornika*

4. EVALUACIJA PREDLOŽENOG RJEŠENJA ZA VERIFIKACIJU GOVORNIKA TEMELJENOG NA IZGOVORENOM SADRŽAJU

U ovom se poglavlju mreže koje su istrenirane u trećem poglavlju evaluiraju pomoću testnih skupova podataka. Na taj način dobiva se uvid koliko dobro predloženo rješenje daje performanse na neviđenim podatkovnim primjerima. Nadalje, testira se i demonstrira rad predložene aplikacije za verifikaciju govornika.

4.1. Testiranje dobivenih mreža

Metrike evaluacije mreže koje se koriste su matrica zabune, iz koje se računa vrijednost točnosti (engl. *accuracy*), preciznosti (engl. *precision*), odziva (engl. *recall*) i specifičnosti (engl. *specificity*). Tablicom 4.1. prikazan je izgled matrice zabune koja se koristi u binarnim klasifikacijskim zadacima u svrhu određivanja performansi modela.

Tablica 4.1. *Prikaz izgleda matrice zabune*

Matrica zabune		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	*TP*	*FP*
	NE (-)	*FN*	*TN*

Zadatak verifikacije govornika predstavlja binarnu klasifikaciju, gdje mreža na temelju sličnosti dva uzorka vrši klasifikaciju. Ako su dva uzoraka slična mreža ih klasificira kao pozitivan par, odnosno kao istog govornika. U suprotnom, u slučaju niske sličnosti dva uzorka, mreža uzorke klasificira kao negativan par, odnosno kao različite govornike. Uz navedeno prema tablici 4.4., *TP* predstavlja broj istinito pozitivnih rezultata (engl. *true positive*), odnosno broj stvarno pozitivnih parova koje mreža klasificira kao pozitivne parove. S druge strane, *FP* predstavlja broj lažno pozitivnih rezultata (engl. *false positive*), odnosno broj stvarno negativnih parova koje mreža klasificira kao pozitivne parove. U području verifikacije govornika, *FP* označava broj govornika kojima je omogućen neovlašten pristup. Nadalje, *TN* predstavlja broj istinito negativnih rezultata (engl. *true negative*), gdje se *TN* vrijednost odnosi na broj stvarno negativnih parova koje mreža klasificira kao negativne parove. U konačnici, *FN* predstavlja broj lažno negativnih rezultata (engl. *false negative*) i označava broj stvarno pozitivnih parova koje mreža klasificira kao negativne parove.

Pomoću vrijednosti varijabli *TP*, *FP*, *TN* i *FN* računaju se metrike točnosti, preciznosti, odziva i specifičnosti [55, 56]. Izrazi navedenih metrika dani su u nastavku:

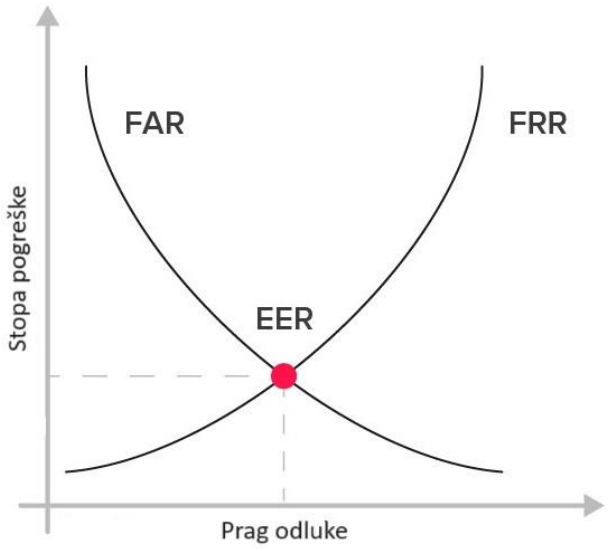
$$Točnost = \frac{TP + TN}{TP + FP + TN + FN}, \tag{4-1}$$

$$Preciznost = \frac{TP}{TP + FP}, \tag{4-2}$$

$$Odziv = \frac{TP}{TP + FN}, \tag{4-3}$$

$$Specifičnost = \frac{TN}{FP + TN}. \tag{4-4}$$

Osim navedenih metrika, za evaluaciju mreže koristi se i stopa jednake pogreške. Stopa jednake pogreške je mjera performansi rada mreže koja se često kositi u području biometrijske sigurnosti kao što je verifikacija govornika. Ona odgovara vrijednosti stope pogreške u trenutku kada su stopa lažnog prihvaćanja (engl. *False Acceptance Rate* - FAR) i stopa lažnog odbijanja (engl. *False Rejection Rate* - FRR) jednake [57, 58]. Na slici 4.1. grafički je prikazana FAR i FRR krivulja te točka u kojoj se one sijeku koja predstavlja EER.



Slika 4.1. *Prikaz izgleda grafa FAR i FRR krivulja s prikazom stope jednake pogreške* [59]

Stopa lažnog prihvaćanja je stopa kojom sustav pogrešno prihvaća lažno pozitivne rezultate, a stopa lažnog odbijanja je stopa kojom sustav netočno odbija istinito pozitivne rezultate [59]. Prema tome može se reći kako vrijednost stope jednake pogreške pokazuje udio lažnih odbijanja koji je

jednak udjelu lažnih prihvaćanja. Niža vrijednost stope jednake pogreške pokazuje veću točnost biometrijskog sustava [58]. Slikom 4.1. prikazan je graf ovisnosti stope pogreške o pragu odluke. Ako je prag odluke nizak, FAR vrijednost je visoka, a FRR je niska. Povećanjem praga odluke dolazi do smanjenje FAR vrijednosti, a povećanja FRR vrijednosti [59].

U nastavku su prikazani rezultati evaluacije na testnom skupu mreža istreniranih na GRID skupu podataka i rezultati mreža dobivenih prijenosnim učenjem. U postupku evaluacije, za svaku mrežu prikazan je graf predikcija na testnom skupu koji prikazuje koliko se predikcije razlikuju od stvarnih vrijednosti. Numerički je iskazana matrica zabune na testnom skupu te metrike točnosti, preciznosti, odziva i specifičnosti. U konačnici je prikazana FAR i FRR krivulja mreže na validacijskom skupu iz koje je vidljiva vrijednost stope jednake pogreške, a na temelju koje se određuje prag odluke koji se koristi za izračun matrice zabune na testnom skupu.

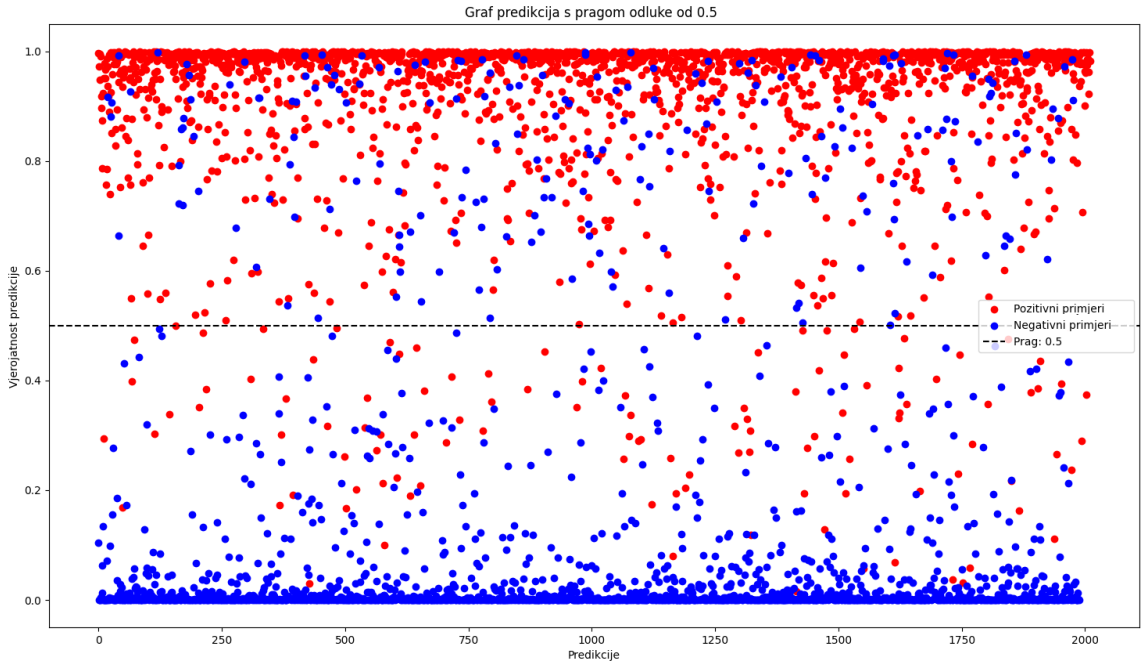
4.1.1. Testiranje mreže trenirane na GRID skupu podataka s kombiniranim podacima

Tijekom treniranja mreže na temelju točnosti i gubitka na validacijskom skupu, najbolja se pokazala ona mreža trenirana na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma. U nastavku su prikazani rezultati evaluacije svih mreža treniranih na kombiniranim podacima. Mreže su evaluirane na testnom skupu na temelju metrika: točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške (EER). U tablici 4.2. prikazani su rezultati navedenih metrika svih mreža uz prag odluke od 0.5. Na temelju rezultata prikazanih tablicom 4.2. uočava se kako mreža trenirana na podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma, ostvaruje lošije rezultate u odnosu na mrežu treniranu na podacima na koje nije primijenjena augmentacija. Međutim, mreža trenirana na podacima koji primjenjuju amplitudnu augmentaciju spektrograma, postiže malo rezultate od mreže trenirane na podacima bez primjene augmentacije te ostvaruje najbolje rezultate među svim analiziranim mrežama.

Tablica 4.2. *Prikaz točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške na testnom skupu mreža treniranih na kombiniranim podacima s pragom odluke od 0.5*

Mreža (Kombinirani podaci)	Točnost	Preciznost	Odziv	Specifičnost	EER
Bez augmentacije	0.9209	0.9007	0.9461	0.8956	0.09
Gaussov šum	0.8571	0.8168	0.9209	0.7933	0.15
Amplitudna augmentacija spektrograma	0.9209	0.9019	0.9446	0.8972	0.09

Slika 4.2. prikazuje graf predikcija navedene najbolje mreže s kombiniranim podacima gdje je prikazan raspored vjerojatnosti predikcija 2000 slučajno izabranih podataka iz testnog skupa podataka. Prag odluke iznosi 0.5, a crvenim točkama na grafu označeni su primjeri pozitivne klase, dok plave označavaju primjere negativne klase. Iz grafa je moguće uočiti kako većina primjera teži ispravnoj vrijednosti, ali i dalje postoje primjeri koji su neispravno klasificirani. Veliki problem mogu predstavljati primjeri koji su s visokom vjerojatnošću neispravno klasificirani, posebice lažno pozitivne vrijednosti. Na taj način dolazi do neispravnog prepoznavanja govornika s velikom sigurnošću što je u stvarnoj primjeni sigurnosni rizik. Za navedenu najbolju mrežu tablicom 4.3. prikazana je matrica zabune na testnom skupu podataka uz prag odluke od 0.5.



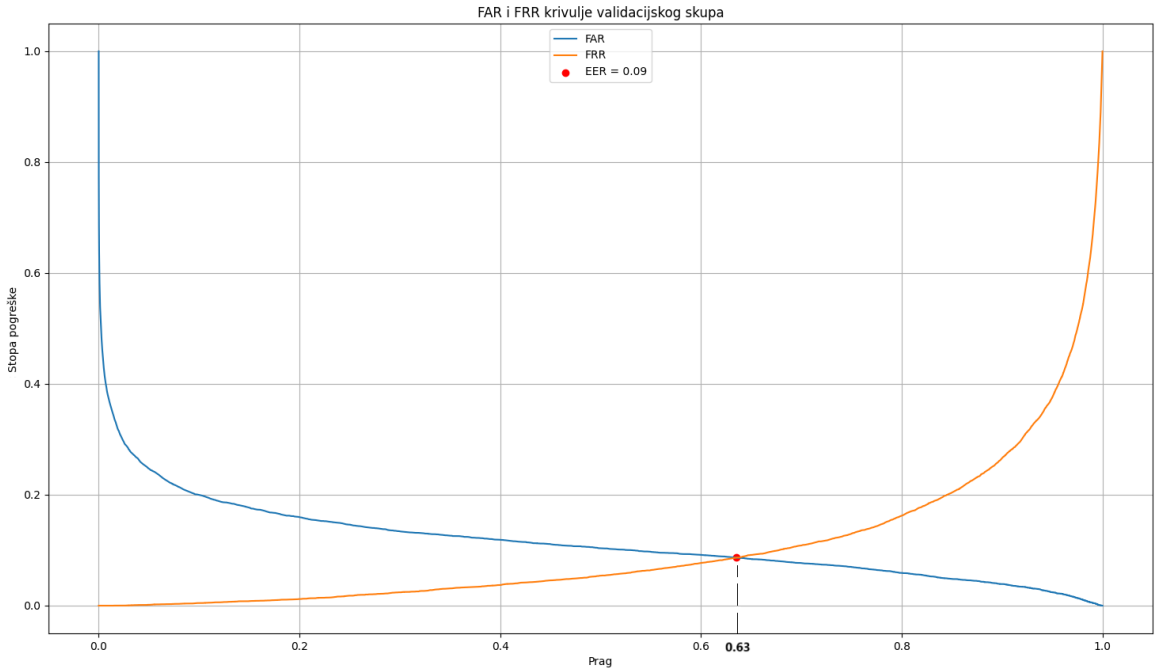
Slika 4.2. *Prikaz grafa predikcija na testnom skupu mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.5*

Tablica 4.3. *Prikaz matrice zabune na testnom skupu mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.5*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	17441	1897
	NE (-)	1023	16559

Slika 4.3 prikazuje graf FAR i FRR krivulja na validacijskom skupu podataka mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Iz slike je vidljivo sjecište FAR i FRR krivulja. Na temelju sjecišta, vrijednost EER dobiva se očitanjem

vrijednosti s y-osi, dok se očitanjem s x-osi dobiva vrijednost praga odluke pri EER koji iznosi 0.63.



Slika 4.3. *Prikaz grafa FAR i FRR krivulja na validacijskom skupu mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*

Matrica zabune navedene najbolje mreže dobivene na testnom skupu uz prag odluke od 0.63 prikazana je tablicom 4.4. U usporedbi s matricom zabune prikazane tablicom 4.3., dolazi do očekivanog smanjenja lažno pozitivnih rezultata, ali i povećanja lažno negativnih rezultata zbog povećanja iznosa praga odluke s 0.5 na 0.63.

Tablica 4.4. *Prikaz matrice zabune na testnom skupu mreže trenirane na kombiniranim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.63*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	16831	1562
	NE (-)	1633	16894

4.1.2. Testiranje mreže trenirane na GRID skupu podataka s razdvojenim podacima

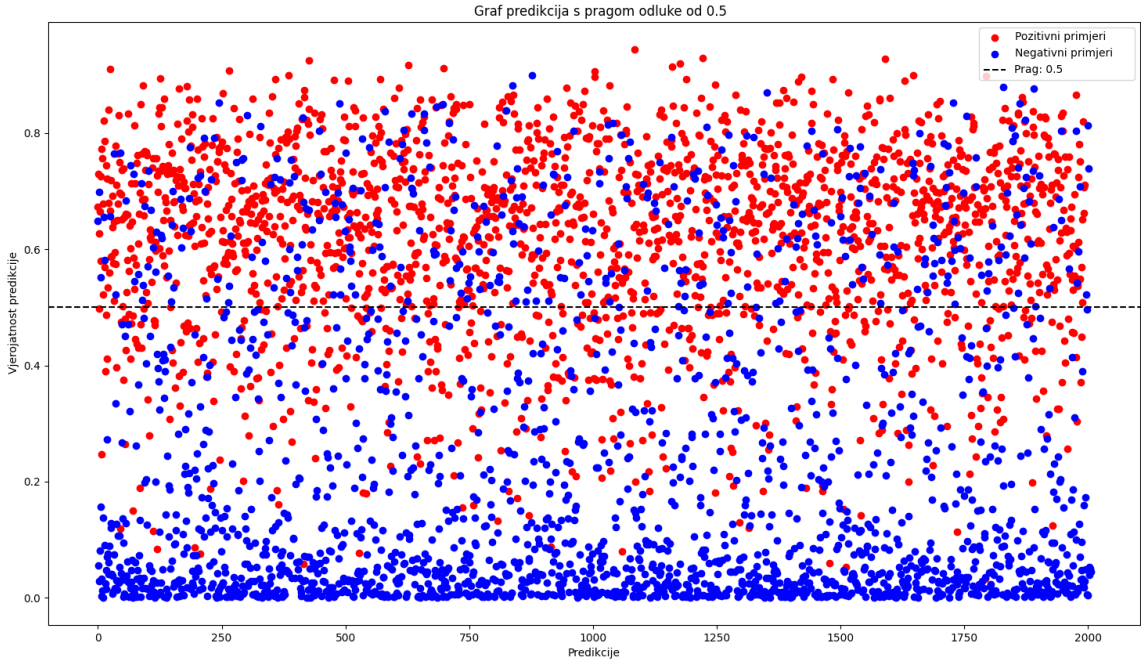
U tablici 4.5. prikazani su rezultati točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške svih mreža treniranih na razdvojenim podacima uz prag odluke od 0.5. Iz rezultata prikazanih navedenom tablicom uočava se kako mreža trenirana na razdvojenim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma, ostvaruje slične rezultate kao mreža trenirana na podacima bez primijenjene augmentacije. Najbolje rezultate ostvaruje mreža trenirana na

podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Upravo ova mreža najbolje rezultate je pokazala na validacijskom skupu podataka tijekom treniranja mreže.

Tablica 4.5. *Prikaz točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške mreža treniranih na razdvojenim podacima s pragom odluke od 0.5*

Mreža (Razdvojeni podaci)	Točnost	Preciznost	Odziv	Specifičnost	EER
Bez augmentacije	0.7878	0.7632	0.8357	0.7398	0.23
Gaussov šum	0.7872	0.7788	0.8032	0.7711	0.22
Amplitudna augmentacija spektrograma	0.7981	0.8088	0.7817	0.8146	0.19

Slika 4.4. prikazuje graf predikcija navedene najbolje mreže s razdvojenim podacima gdje je prikazan raspored vjerojatnosti predikcija 2000 slučajno izabranih podataka iz testnog skupa podataka. Prag odluke iznosi 0.5, a oznake su jednake onima kao kod kombiniranih podataka (Slika 4.2.). Iz grafa je moguće uočiti kako mreža u prosjeku s manjom vjerojatnošću ispravno klasificira primjere pozitivne klase u usporedbi s mrežom treniranom na kombiniranim podacima. Primjeri negativne klase dijelom su ispravno klasificirane, ali i kod njih se javlja raspršenje predikcija, odnosno postoje neispravno klasificirani negativni primjeri koji se klasificiraju kao pozitivni.



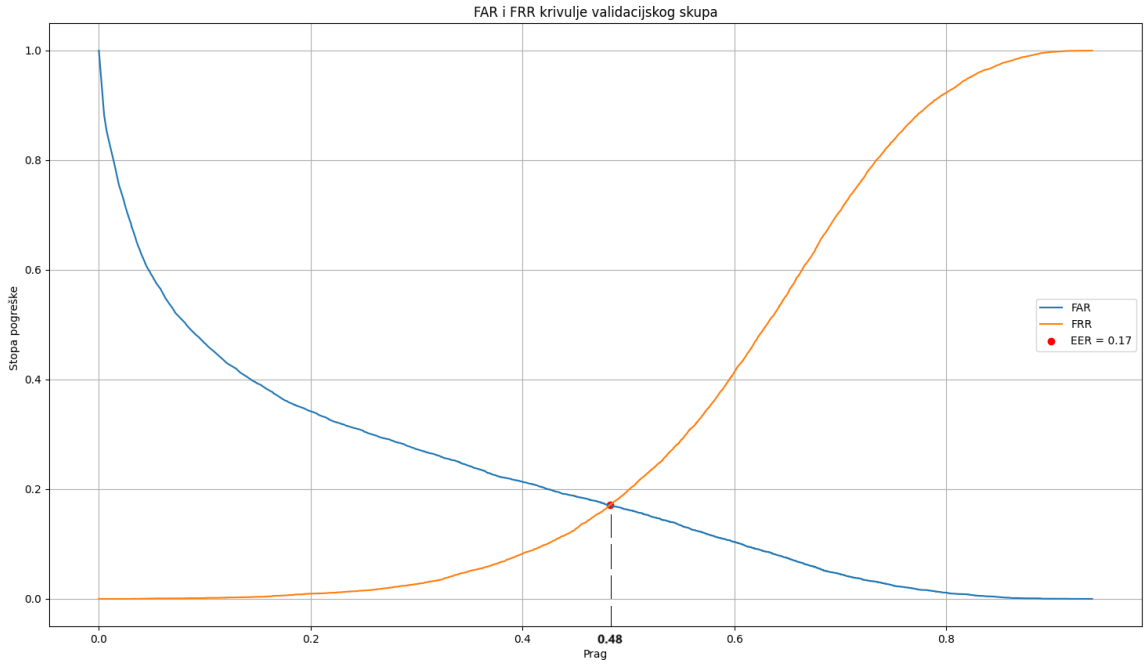
Slika 4.4. *Prikaz grafa predikcija na testnom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.5*

Za navedenu najbolju mrežu tablicom 4.6. prikazana je matrica zabune na testnom skupu podataka uz prag odluke od 0.5.

Tablica 4.6. *Prikaz matrice zabune na testnom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.5*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	14927	3529
	NE (-)	4169	15508

Slika 4.5 prikazuje graf FAR i FRR krivulja na validacijskom skupu podataka mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma. Prag odluke koji se očitava iz slike 4.5. pri EER iznosi 0.48. Matrica zabune navedene mreže dobivene na testnom skupu uz prag odluke od 0.48 prikazana je tablicom 4.7.



Slika 4.5. *Prikaz grafa FAR i FRR krivulja na validacijskom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma*

Tablica 4.7. *Prikaz matrice zabune na testnom skupu mreže trenirane na razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma s pragom odluke od 0.48*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	15438	3732
	NE (-)	3658	15305

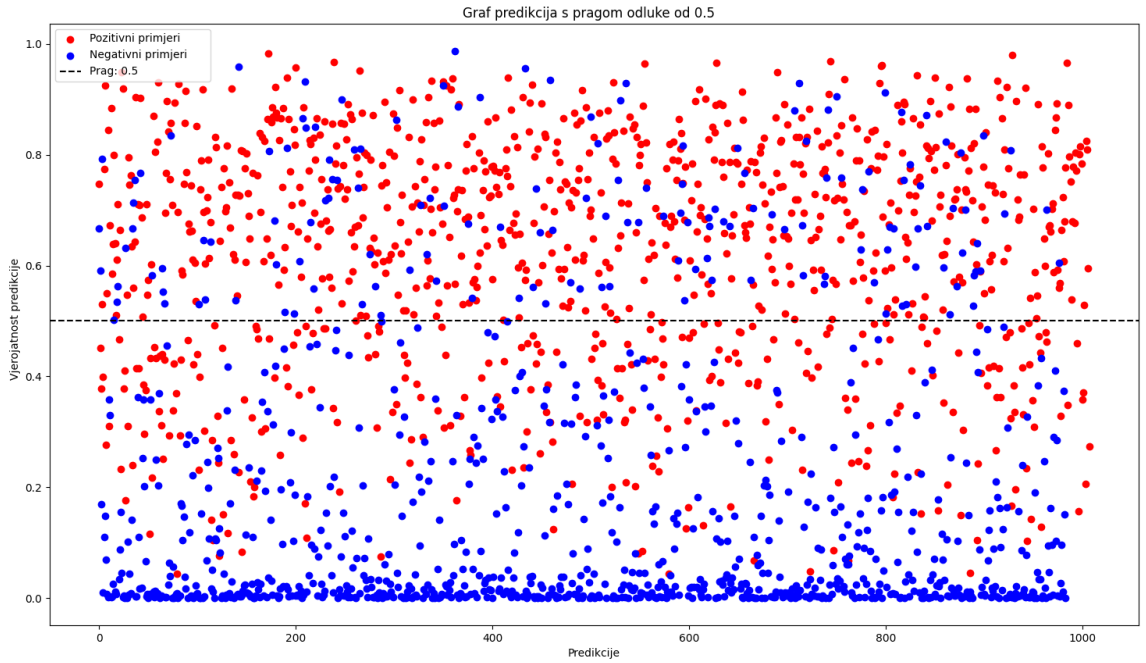
4.1.3. Testiranje mreža dobivenih prijenosnim učenjem

Kod mreža dobivenih procesom prijenosnog učenja, koristi se ista metrika procjene kao i kod onih mreža treniranih na GRID skupu podataka. Međutim, rad mreže se evaluira na testnom skupu vlastitog skupa podataka. Razlog korištenja vlastitog skupa je taj što je mreža trenirana na vlastitom skupu podatka. U tablici 4.8 prikazani su rezultati navedenih metrika za sve mreže dobivene na temelju izvornih mreža treniranih na kombiniranim podacima. Mreža dobivena prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma ostvaruje najbolje rezultate i koja je također prema točnosti i gubitku na validacijskom skupu u procesu treniranja, pokazala najbolje rezultate.

Tablica 4.8. *Prikaz točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške mreža dobivenih prijenosnim učenjem na temelju izvornih mreža treniranih na kombiniranim podacima s pragom odluke od 0.5*

Izvorna mreža trenirana na kombiniranim podacima	Točnost	Preciznost	Odziv	Specifičnost	EER
Bez augmentacije	0.7806	0.8226	0.7222	0.8404	0.21
Gaussov šum	0.7781	0.8245	0.7133	0.8445	0.18
Amplitudna augmentacija spektrograma	0.7545	0.7842	0.7103	0.7998	0.22

Slikom 4.6. prikazan je graf predikcija navedene najbolje mreže uz prag odluke od 0.5.



Slika 4.6. *Prikaz grafa predikcija na testnom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.5*

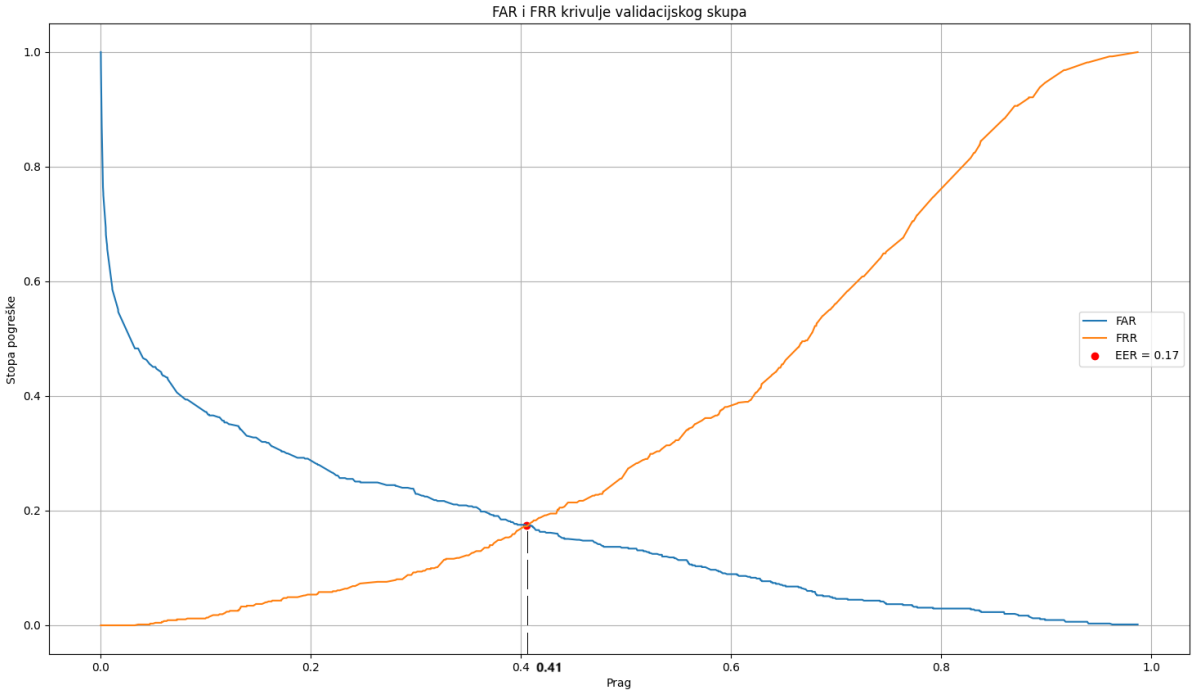
Iz grafa (Slika 4.6.) se uočava kako dolazi do nasumičnog razmještaja primjera pozitivne klase. Pozitivni primjeri se klasificiraju s različitom razinom vjerojatnosti predikcija što dovodi do određenog broja neispravnih klasifikacija i malog broja ispravno klasificiranih vrijednosti s viskom vjerojatnosti predikcije. S druge strane, određeni dio negativnih primjera mreža ispravno klasificira s velikom sigurnošću, ali i dalje postoji određeni dio lažno pozitivnih primjera.

Za mrežu dobivenu prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum tablicom 4.9. prikazana je matrica zabune dobivena na testnom skupu vlastitog skupa podataka s pragom odluke od 0.5.

Tablica 4.9. *Prikaz matrice zabune na testnom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.5*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	719	153
	NE (-)	289	831

Slika 4.7. prikazuje FAR i FRR krivulje na validacijskom skupu podataka navedene mreže iz kojih se očitava vrijednost praga odluke pri EER od 0.41. Na temelju tog praga odluke dobiva se matrica zabune na testnom skupu koja je prikazana tablicom 4.10.



Slika 4.7. *Prikaz grafa FAR i FRR krivulja na validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum*

Tablica 4.10. *Prikaz matrice zabune testnog skupa mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.41*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	825	182
	NE (-)	183	802

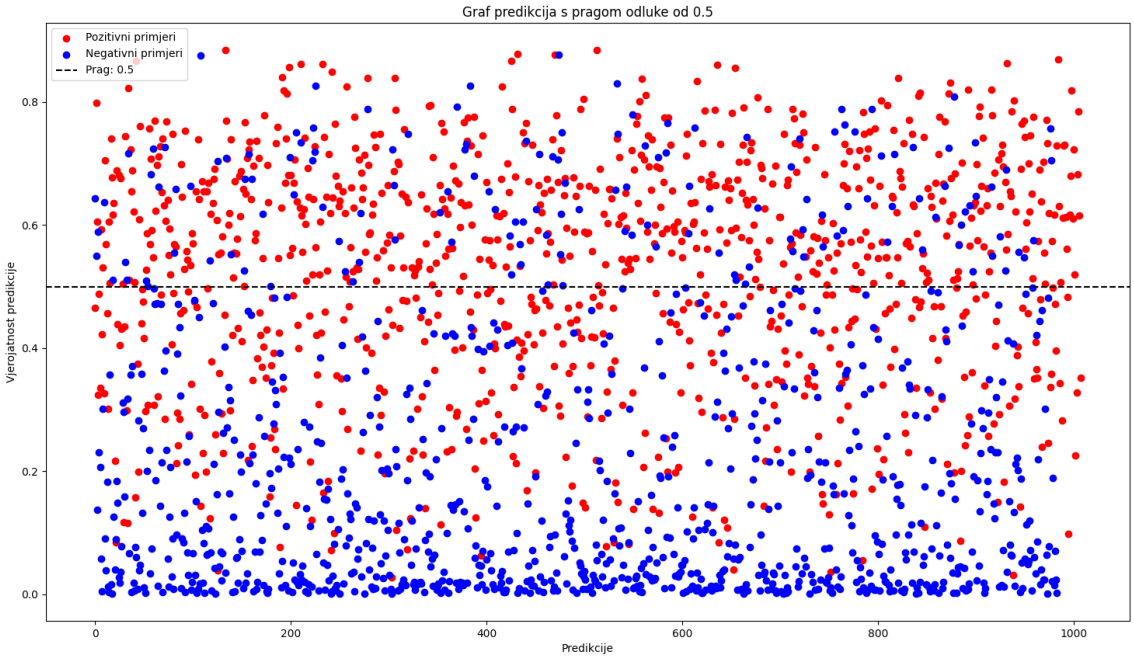
Kod mreža dobivenih na temelju izvornih mreža s razdvojenim podacima najbolje rezultate na validacijskom skupu u procesu treniranja ostvaruje ona mreža temeljena na izvornoj mreži treniranoj na podacima koji primjenjuju augmentaciju Gussov šum. Međutim, iz rezultata prikazanih u tablici 4.11. dobivenih na testnom skupu najveću točnost ostvaruje mreža s izvornom mrežom treniranoj na podacima bez primijenjene augmentacije i ima najveću EER vrijednost. S druge strane, mreža dobivena na temelju izvorne mreže trenirane na podacima s primjenom amplitudne augmentacije spektrograma ima najmanju EER vrijednost, ali najmanju točnost i odziv. Kao mreža koja ostvaruje rezultate sa srednjom vrijednosti ove dvije je upravo ona dobivena na izvornoj mreži s razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum, te ona predstavlja mrežu s najboljim rezultatima.

Tablica 4.11. *Prikaz točnosti, preciznosti, odziva, specifičnosti i stope jednake pogreške mreža dobivenih prijenosnim učenjem na temelju izvornih mreža treniranih na razdvojenim podacima s pragom odluke od 0.5*

Izvorna mreža trenirana na razdvojenim podacima	Točnost	Preciznost	Odziv	Specifičnost	EER
Bez augmentacije	0.7701	0.7429	0.8343	0.7043	0.26
Gaussov šum	0.7686	0.7617	0.7897	0.7470	0.23
Amplitudna augmentacija spektrograma	0.7555	0.8654	0.6121	0.9024	0.19

Slikom 4.8. prikazan je graf predikcija navedene najbolje mreže na testnom skupu uz prag odluke od 0.5. Iz navedenog grafa predikacija moguće je zaključiti kako mreža ostvaruje slične rezultate kao ona koja je dobivena na izvornoj mreži s kombiniranim podacima (Slika 4.6.). Međutim, primjeri pozitivne klase u prosjeku se klasificiraju s manjim iznosom vjerojatnosti predikcije.

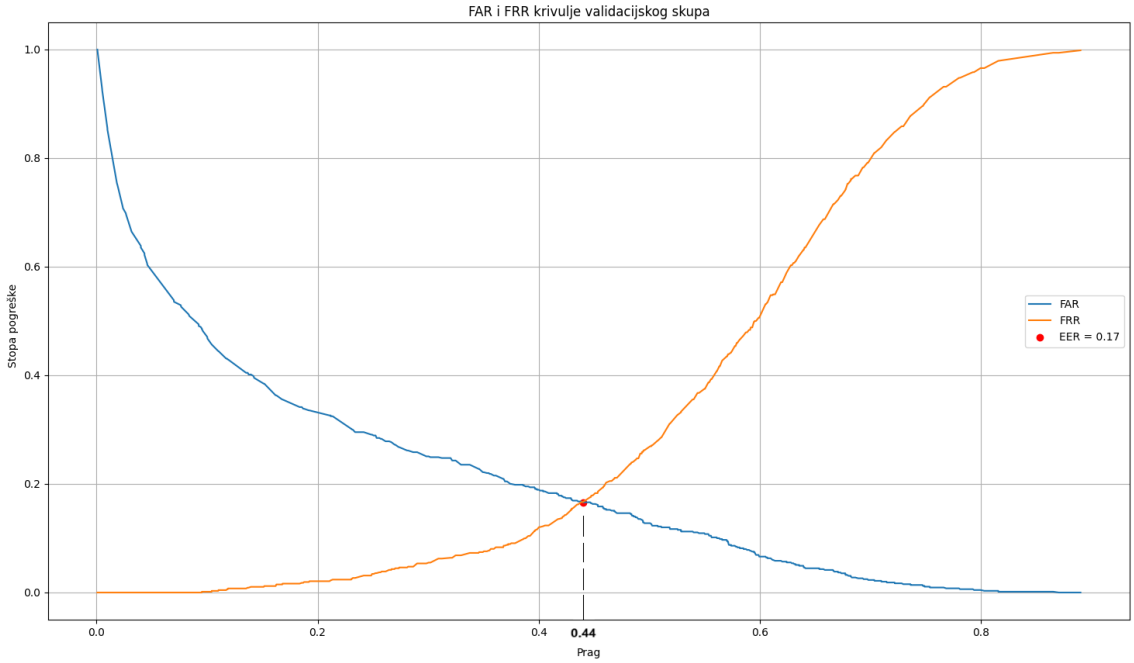
Za mrežu dobivenu prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma tablicom 4.12. prikazana je matrica zabune dobivena na testnom skupu uz prag odluke od 0.5. Slika 4.9. prikazuje graf FAR i FRR krivulja na validacijskom skupu podataka navedene mreže.



Slika 4.8. *Prikaz grafa predikcija na testnom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.5*

Tablica 4.12. *Prikaz matrice zabune na testnom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.5*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	611	147
	NE (-)	397	837



Slika 4.9. *Prikaz grafa FAR i FRR krivulja na validacijskom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum*

Iz grafa FAR i FRR krivulja prikazanog slikom 4.9. očitava se vrijednost praga odluke pri EER od 0.44. Na temelju tog praga odluke dobiva se matrica zabune na testnom skupu koja je prikazana tablicom 4.13.

Tablica 4.13. *Prikaz matrice zabune na testnom skupu mreže dobivene prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima koji primjenjuju augmentaciju Gaussov šum s pragom odluke od 0.44*

Matrica zabune – Testni skup		Stvarna klasa	
		DA (+)	NE (-)
Predviđena klasa	DA (+)	706	187
	NE (-)	302	797

4.2. Testiranje i demonstracija rada aplikacije

Proces testiranja rada aplikacije provodio se na način objašnjen u nastavku. Prvotno, baza podataka popunjena je korisnicima gdje su korisnici tester aplikacije i određeni izabrani govornici iz GRID i vlastitog skupa podataka. Svaki govornik određen je frazom koju izgovara. Popis fraza prema govorniku koji ih izgovora i pripadajućem skupu govornika dan je u nastavku:

- frazu „bin blue at“ izgovara 4 govornika iz GRID skupa, 5 govornika iz vlastitog skupa i tester aplikacije,
- frazu „bin green by“ izgovara 4 govornika iz GRID skupa i 2 govornika iz vlastitog skupa,
- frazu „close first door“ izgovara 2 korisnika iz vlastitog skupa,
- frazu „close first window“ izgovara 2 korisnika iz vlastitog skupa i
- frazu „try get in“ izgovara tester aplikacije.

U konačnici to rezultira s ukupno 21 spremljenim uzorkom gdje tester aplikacije jedini izgovara dvije različite fraze.

Nakon prvog koraka slijedi proces testiranja verifikacije govornika kroz aplikaciju. Prethodno je navedeno kako aplikacija ima testni način rada. U tom načinu rada moguće je odabrati jednu od 12 ponuđenih mreža za verifikaciju govornika:

- 6 mreža treniranih na GRID skupu podataka na kombiniranim i razdvojenim podacima,
- 3 mreže dobivene procesom prijenosnog učenja na temelju izvornih mreža treniranih na kombiniranim podacima,
- 3 mreže dobivene procesom prijenosnog učenja na temelju izvornih mreža treniranih na razdvojenim podacima.

Kada korisnik odabere jednu od tih mreža započinje s procesom verifikacije izgovaranjem odgovarajuće fraze. S ciljem ograničavanja lažne klasifikacije i smanjenja lažno pozitivnih vrijednosti, prag donošenja odluke postavljen je na 0.6. U prvom testnom slučaju, tester aplikacije izgovara frazu „bin blue at“ koja je prisutnu u sva tri skupa podataka. Cilj je utvrditi koliko dobro pojedina mreža radi kada se izgovori fraza koja predstavlja istu zaporku za različite korisnike. Postupkom je utvrđeno kako mreže koje su trenirane na GRID skupu podataka vrše neispravno prepoznavanje testera kao govornika iz GRID skupa podataka. Jedina iznimka je mreža trenirana na skupu razdvojenim podacima koji primjenjuju amplitudnu augmentaciju spektrograma uz ispravno prepoznavanje testera u jednom od tri uzastopna testiranja. Kod mreža koje su dobivene prijenosnim učenjem, treniranih na vlastitom skupu podataka vidljivo je poboljšanje. Međutim, i dalje dolazi do neispravnog prepoznavanja testera kao govornika iz GRID skupa podataka. Dvije mreže pokazale su najbolje rezultate:

- mreža dobivena na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma i
- mreža dobivena na temelju izvorne mreže trenirane na razdvojenim podacima koji ne primjenjuju augmentaciju.

U drugom testnom slučaju, tester izgovara frazu koja se ne nalazi niti u GRID niti u vlastitom skupu podataka „try get in“. Kod mreža treniranih na GRID skupu podataka, one mreže koje su trenirane na kombiniranim podacima vršile su lažno odbijanje, ali i ispravno prepoznavanje govornika. Mreže trenirane na GRID skupu podataka s razdvojenim podacima često su lažno prepoznale testera kao govornika iz vlastitog skupa podataka. S druge strane, mreže koje su dobivene prijenosnim učenjem na temelju izvornih mreža treniranih na kombiniranim podacima u većini slučajeva vrše lažno odbijanje, uz rijetko ispravno prepoznavanje govornika. Izuzetak je mreža dobivena prijenosnim učenjem na temelju izvorne mreže trenirane na kombiniranim podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma. Navedena mreža u većini slučajeva vrši ispravno prepoznavanje govornika. Kod mreža dobivenih prijenosnim učenjem na temelju izvornih mreža treniranih na kombiniranim podacima prihvatljive rezultate jedino ostvaruje mreže čija je izvorna mreža trenirana na skupu s razdvojenim podacima koji ne primjenjuju augmentaciju podataka.

Važno je istaknuti kako je proces testiranja proveden u prostoriji s malom prisutnosti šumova. Uočeno je kako promjena okruženja, ali i amplituda snimljenog zvučnog zapisa govornika utječu na kvalitetu prepoznavanja govornika. Kod testiranja rada aplikacije najboljom se pokazala mreža

dobivena prijenosnim učenjem na temelju izvorne mreže trenirane na razdvojenim podacima bez primjene augmentacije. Međutim navedena mreža vrši predviđanje s niskom razinom vjerojatnosti predikcija prosječenog iznosa od 63%. Navedeno predstavlja problem jer je prag odluke postavljeni na 60%, a njegovo smanjenje povećalo bi broj neovlaštenih pristupa. U konačnici se dolazi do zaključka kako aplikacija funkcionira na način zamišljen ovim radom.

5. ZAKLJUČAK

U okviru ovoga rada je proučen i riješen zadatak verifikacije govornika primjenom dubokih neuronskih mreža. Za uspješnu verifikaciju govornika ovisne o izgovornom sadržaju potreban je velik broj podataka za treniranje mreže. Ovaj izazov se kroz rad nastojao riješiti obradom zvučnih zapisa GRID skupa podataka izdvajanjem dijela zvučnog zapisa u kojem se izgovaraju prve tri riječi. Budući da je navedeni skup snimljen u optimalnim uvjetima, primjenom augmentacija kao što je umjetno dodavanje Gaussovog šuma zvučnim zapisima i amplitudna augmentacija spektrogram nastojao se stvoriti skup na temelju kojeg bi se treniranjem mreže omogućio bolji rad u stvarnoj primjeni. Osim augmentacija, raznolikost skupa podataka poboljšana je proširenjem GRID skupa vlastitim skupom podataka.

Evaluacijom na testnom skupu GRID skupa podatka uočava se kako mreže trenirane na GRID skupu podataka na čijem trening skupu je primijenjena amplitudna augmentacija spektrograma postižu najbolje rezultate. Uz prag odluke 0.5, mreže trenirane na kombiniranim podacima na testnom skupu postižu točnost iznad 90%, dok mreže trenirane na razdvojenim podacima točnost približno 80%. Međutim, kod mreža treniranih na razdvojenim podacima testiranje je provedeno na podacima dobivenih iz zvučnih zapisima govornika koji ne pripadaju trening skupu, što je uveliko utjecalo na rezultate. Evaluacijom na testnom GRID skupu podatka, mreža trenirana na podacima koji primjenjuju augmentaciju pomoću Gaussovog šuma očekivano je dala lošije rezultate u odnosu na ostale mreže upravo zato što je GRID skup snimljen s minimalnom prisutnosti šuma. Međutim, evaluacijom na testnom skupu vlastitih podataka, mreže dobivene prijenosnim učenjem temeljene na mrežama treniranim na GRID skupu podataka koji primjenjuju augmentaciju pomoću Gaussovog šuma, pokazale su najbolje rezultate. Ovo je dijelom očekivano jer su te mreže testirane na vlastitom skupu podataka snimljenom uz određenu prisutnost šuma.

Testiranjem aplikacije u stvarnom okruženju uočava se kako mreže trenirane na GRID skupu često vrše neispravnu klasifikaciju testera kao govornika iz GRID skupa ako izgovara frazu koja se nalazi u navedenom skupu. Najbolje mreže pokazale su se one dobivene procesom prijenosnog učenja, što naglašava važnost raznolikosti skupa podataka na kojem se mreža trenira. Mreža dobivena prijenosnim učenjem temeljena na mreži treniranoj na razdvojenim podacima bez augmentacije podatka pokazala se najboljom.

LITERATURA

[1] W. Shen, M. Surette i R. Khanna, Proceedings of the IEEE, Evaluation of automated biometrics-based identification and verification systems, br. 9, svez. 85, str. 1464-1478, rujan 1997.

[2] J. M. Stewart, The Three Types of Multi-Factor Authentication(MFA) [Mrežno], Global Knowledge, 2018., dostupno na: https://www.globalknowledge.com/us-en/resources/resource-library/articles/the-three-types-of-multi-factor-authentication-mfa/#gref [22. kolovoza 2024.].

[3] One Span, Identity Verification [Mrežno], One Span, dostupno na: https://www.onespan.com/topics/identity-verification [22. kolovoza 2024.]

[4] J. Ashok, V. Shivashankar i P.V.G.S. Mudiraj, An Overview of Biometrics, International Journal on Computer Science and Engineering, svez. 2, str. 2402 - 2408, listopad 2010.

[5] J. N. Pato i L. I. Millett, Biometric Recognition: Challenges and Opportunitie, National Academies Press (US), Washington (DC), 2010.

[6] A. K. Jain, P. Flynn i A. A. Ross, Handbook of Biometrics, Springer Science & Business Media, New York, 2007.

[7] X.-L. Zhang i Z. Bai, Speaker Recognition Based on Deep Learning: An Overview, Neural Networks, svez. 140, str. 65-99, travanj 2021.

[8] L. J. Jie, M. M. A. Zabidi, S. Sadiah i A. A.-H. A. Rahman, Siamese Networks for Speaker Identification on Resource-Constrained Platforms, Journal of Physics: Conference Series, svez. 2622, br. 1., str. 12- 14, listopad 2023.

[9] European Language Grid, The Grid Audio-Visual Speech Corpus [Mrežno], European Language Grid, 2021., dostupno na: https://live.european-language-grid.eu/catalogue/corpus/7769/overview/ [19. siječnja 2024.]

[10] F. Millstein, Convolutional Neural Networks in Python: Beginner's Guide to Convolutional Neural Networks in Python, North Charleston: Createspace Independent Publishing Platform, 2018.

[11] History of Information, The Inspiration for Artificial Neural Networks, Building Blocks of Deep Learning [Mrežno], History of Information, dostupno na: https://historyofinformation.com/detail.php?entryid=4726 [20. lipnja 2024]

[12] Educative, How to perform convolution in matrix multiplication [Mrežno], Educative, dostupno na: https://www.educative.io/answers/how-to-perform-convolution-in-matrix-multiplication [23. kolovoza 2024.]

[13] F. Feng, S. Wang, C. Wang i J. Zhang, Learning Deep Hierarchical Spatial–Spectral Features for Hyperspectral Image Classification Based on Residual 3D-2D CNN, Sensors, br. 23, svez. 19, listopad 2019.

[14] D. H. T. Hien, Ml Review [Mrežno], Medium, 2017., dostupno na: https://blog.mlreview.com/a-guide-to-receptive-field-arithmetic-for-convolutional-neural-networks-e0f514068807 [23. kolovoza 2024.]

[15] A. M. Jalil, F. S. Hasan i H. A. Alabbasi, Speaker identification using convolutional neural network for clean and noisy speech samples, First International Conference of Computer and Applied Sciences, Baghdad, Irak, 2019., str. 57-62

[16] Roboflow, What is a Convolutional Neural Network? [Mrežno], Roboflow, dostupno na: https://blog.roboflow.com/what-is-a-convolutional-neural-network/ [28. lipanja 2024.]

[17] Convolutional Neural Network — Lesson 9: Activation Functions in CNNs [Mrežno], Medium, dostupno na: https://medium.com/@nerdjock/convolutional-neural-network-lesson-9-activation-functions-in-cnns-57def9c6e759 [20. lipnja 2024.]

[18] S. Sharma, Activation Functions in Neural Networks [Mrežno], Medium, dostupno na: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6 [20. lipnja 2024.]

[19] P.-H. Kuo, S.-T. Lin i J. Hu, DNAE-GAN: Noise-free acoustic signal generator by integrating autoencoder and generative adversarial network, International Journal of Distributed Sensor Networks, svez. 16, svibanj 2020.

[20] M. Yani, S. Irawan i C. Setianingsih, Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail, Journal of Physics: Conference Series, svez. 1201, str. 012052, svibanj 2019.

[21] EITCA, What is the role of the fully connected layer in a CNN? [Mrežno], European Information Technologies Certification Academy, dostupno na: https://eitca.org/artificial-intelligence/eitc-ai-dlptfk-deep-learning-with-python-tensorflow-and-keras/convolutional-neural-networks-cnn/introduction-to-convolutional-neural-networks-cnn/examination-review-introduction-to-convolutional-neural-networks-cnn/ [22. lipnja 2024.]

[22] Kears, Dropout layer [Mrežno], Kears, dostupno na: https://keras.io/api/layers/regularization_layers/dropout/ [22. lipnja 2024.]

[23] K. O'Shea i R. Nash, An Introduction to Convolutional Neural Networks, ArXiv, studeni 2015.

[24] P. Kumar, Siamese Networks Introduction and Implementation [Mrežno], Medium, dostupno na: https://medium.com/@prabhattgs12345789/siamese-neural-network-enhancing-ai-capabilities-with-pairwise-comparisons-4f00e2dd8256 [20. lipnja 2024.]

[25] G. Koch, R. Zemel i R. Salakhutdinov, Siamese Neural Networks for One-shot Image Recognition, Toronto, 2015.

[26] Y. Zhang, M. Yu, N. Li, C. Yu, J. Cui i D. Yu, Seq2Seq Attentional Siamese Neural Networks for Text-dependent Speaker Verification, u ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019., str. 6131-6135

[27] J. Atanbori i S. Rose, MergedNET: A simple approach for one-shot learning in siamese networks based on similarity layers, Neurocomputing, svez. 509, str. 1-10, listopad 2022.

[28] S. Soonshin, K. Changmin i K. Ji-Hwan, Convolutional Neural Networks Using Log Mel-Spectrogram Separation for Audio Event Classification with Unknown Devices, Journal of Web Engineering, svez. 21, str. 497-522, svibanj 2022.

[29] H. Jeon, Y. Jung, S. Lee i Y. Jung, Area-Efficient Short-Time Fourier Transform Processor for Time–Frequency Analysis of Non-Stationary Signals, Applied Sciences, svez. 10, str. 7208, listopad 2020.

[30] K. A. Khadarnawas, B. Manish i K. Nayeemulla, Speaker Recognition using Random Forest, ITM Web of Conference, svez. 37, siječanj 2021.

[31] USCD, Mel Vs. Hertz [Mrežno], USCD, dostupno na: http://musicweb.ucsd.edu/~trsmyth/pitch2/Mel_Vs_Hertz.html [14. lipnja 2024.]

[32] MATLAB, Spoken Digit Recognition with Custom Log Spectrogram Layer and Deep Learning [Mrežno], MATLAB, dostupno na: https://www.mathworks.com/help/signal/ug/spoken-digit-recognition-with-custom-log-spectrogram-layer-and-deep-learning.html [2024. kolovoza 22.]

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren i V. Zue, TIMIT Acoustic-Phonetic Continuous Speech Corpus [Mrežno], dostupno na: https://catalog.ldc.upenn.edu/LDC93S1 [3. siječnja 2024.]

[34] NOISEX-92 noise dataset [Mrežno], dostupno na: http://spib.linse.ufsc.br/noise.html [3. siječnja 2024.]

[35] Robots, The VoxCeleb2 Dataset [Mrežno], Robots, dostupno na: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/vox2.html [11. siječnja 2024.]

[36] Microsoft, Speaker Verification: Text-Dependent vs. Text-Independent [Mrežno], Microsoft, dostupno na: https://www.microsoft.com/en-us/research/project/speaker-verification-text-dependent-vs-text-independent/ [22. kolovoza 2024.]

[37] Y. Tu, W. Lin i M.-W. Mak, A Survey on Text-Dependent and Text-Independent Speaker Verification, IEEE Access, svez. 10, str. 99038-99049, rujan 2022.

[38] S. Soleymani, A. Dabouei, S. M. Iranmanesh, H. Kazemi, J. Dawson i N. M. Nasrabadi, Prosodic-Enhanced Siamese Convolutional Neural Networks for Cross-Device Text-Independent Speaker Verification, u 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), Redondo Beach, Kalifornija, SAD, 2018., str. 1-7

[39] L. Mary i B. Yegnanarayana, Prosodic Features For Language Identification, u IEEE-International Conference on Signalprocessing, Communications and Networking, Chennai, veljača 2008.

[40] A. Fox, Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals, Oxford University Press, New York, 2000.

[41] University of Amsterdam (2018), Praat, dostupno na: http://www.fon.hum.uva.nl/praat/ [22. kolovoza 2024.]

[42] VOICEBOX: Speech Processing Toolbox for MATLAB [Mrežno], dostupno na: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html [22. kolovoza 2024.]

[43] U. Khan i J. Hernando, Unsupervised Training of Siamese Networks for Speaker Verification, Interspeech, Shanghai, ISCA, 2020., str. 3002-3006

[44] Javapoint, Unsupervised Machine Learning [Mrežno], dostupno na: https://www.javatpoint.com/unsupervised-machine-learning [28. lipnja 2024.]

[45] T. a. H. Zhang, X. Lun and Li i G. Feng, Efficient End-to-End Sentence-Level Lipreading with Temporal Convolutional Networks, Applied Sciences, svez. 11, str. 69-75, srpanj 2021.

[46] T. Bäckström, O. Räsänen, A. Zewoudie, P. Pérez Zarazaga, L. Koivusalo, S. Das, E. Gómez Mellado, M. Bouafif Mansali i D. Ramos, Voice Activity Detection (VAD) [Mrežno], Aalto University, dostupno na: https://speechprocessingbook.aalto.fi/Recognition/Voice_activity_detection.html [18. veljače 2024.]

[47] Librosa [Mrežno], dostupno na: https://librosa.org/doc/latest/index.html [18. veljače 2024.]

[48] NumPy [Mrežno], dostupno na: https://numpy.org/ [18. veljače 2024.]

[49] Scikit-image [Mrežno], dostupno na: https://scikit-image.org/ [20. travnja 2024.]

[50] T. Shah, About Train, Validation and Test Sets in Machine Learning [Mrežno], Towards Dana Science, 2017., dostupno na: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 [28. srpnja 2024.]

[51] F. E. Szabo, M, The Linear Algebra Survival Guide, Academic Press, Boston, 2015., str. 219-233

[52] IBM, What is overfitting? [Mrežno], IBM, dostupno na: https://www.ibm.com/topics/overfitting [18. srpnja 2024.]

[53] K. Muralidhar, Learning Curve to identify Overfitting and Underfitting in Machine Learning [Mrežno], Towards Dana Science, 2021., dostupno na: https://towardsdatascience.com/learning-curve-to-identify-overfitting-underfitting-problems-133177f38df5 [14. lipnja 2024.]

[54] IBM, What is transfer learning? [Mrežno], IBM, dostupno na: https://www.ibm.com/topics/transfer-learning [14. srpnja 2024.]

[55] P. Singh, N. Singh, K. K. Singh i A. Singh, Chapter 5 - Diagnosing of disease using machine learning, Machine Learning and the Internet of Medical Things in Healthcare, Academic Pres, 2021., str. 89-111

[56] M. Khanna, Classification Problem: Relation between Sensitivity, Specificity and Accuracy [Mrežno], Analytics Vidhya, 2021., dostupno na: https://www.analyticsvidhya.com/blog/2021/06/classification-problem-relation-between-sensitivity-specificity-and-accuracy/ [14. lipnja 2024.]

[57] M. N. Yaacob, S. Z. Syed Idrus, W. N. A. Wan Ali, W. Mustafa, M. Jamlos i M. H. Abd Wahab, Decision Making Process in Keystroke Dynamics, Journal of Physics: Conference Series, svez. 1529, str. 022087, travanj 2020.

[58] Webopedia, EER – equal error rate [Mrežno], Webopedia, dostupno na: https://www.webopedia.com/definitions/equal-error-rate/ [5. siječnja 2024.]

[59] Recfaces, The False Rejection Rate: The Importance of FRR & FAR Rates [Mrežno], Recfaces, dostupno na: https://recfaces.com/articles/false-rejection-rate#2 [5. siječnja 2024.]

SAŽETAK

U okviru ovog rada predstavljen je način rješavanja zadatka verifikacije govornika primjenom dubokih neuronskih mreža. Prvotno su definirani zadaci identifikacije i verifikacije govornika te su predstavljena moderna rješenja za rješavanje tih problema, pri čemu se ističu rješenja ostvarena korištenjem konvolucijske i sijamske neuronske mreže. Ovim radom zadatak verifikacije govornika se rješava treniranjem sijamske neuronske mreže na temelju logaritamskih Mel spektrograma stvorenih iz zvučnih zapisa GRID skupa podataka. Primjenom augmentacija Gaussovog šuma i amplitudne augmentacije spektrograma na trening skupu podataka treniraju se nove mreže kojima se procjenjuje utjecaj augmentacija na rad verifikacije govornika. S ciljem povećanja raznolikosti skupa podataka, stvoren je vlastiti skup podataka s 12 različitih govornika. Korištenje tog skupa i procesom prijenosnog učenja na temelju već istreniranih mreža na GRID skupu podataka dobivene su nove mreže. Nadalje, izrađena je aplikacija koja omogućuje primjenu istreniranih mreža za verifikaciju govornika u stvarnom okruženju. U konačnici, sve izgrađene mrže su detaljno evaluirane na testnom skupu podataka.

Ključne riječi: Identifikacija govornika, Konvolucijska neuronska mreža, Sijamska neuronska mreža, Spektrogram, Verifikacija govornika

ABSTRACT

Speaker recognition using deep neural networks

This master's thesis presents a method for solving the speaker verification task by using deep neural networks. Firstly, the tasks of speaker identification and verification are defined, followed by presenting modern solutions for these problems, with a focus on solutions achieved using Convolutional Neural Networks and Siamese Neural Networks. In this paper, the speaker verification task is solved by training a Siamese Neural Network based on log-Mel spectrograms created on audio recordings of the GRID dataset. By applying data augmentations on training dataset, such as Gaussian noise and amplitude augmentation of spectrograms, new networks are trained to evaluate the impact of these augmentations on speaker verification performance. To increase the diversity of the GRID dataset, a custom dataset with 12 different speakers was created. Using this dataset and a transfer learning process based on networks already trained on the GRID dataset, new networks were trained. Furthermore, a desktop application was developed that enables the use of trained networks for speaker verification in real-world environment. In the end, all constructed networks were thoroughly evaluated on the test dataset.

Key words: Convolutional Neural Network, Siamese Neural Network, Speaker identification, Speaker verification, Spectrogram

ŽIVOTOPIS

Luka Markić rođen je 14. siječnja 2000. godine u Mostaru, 2014. godine završava Treću osnovnu školu Mostar nakon čega upisuje Srednju elektrotehničku školu Ruđera Boškovića u Mostaru. Završetkom srednjoškolskog obrazovanja upisuje prijediplomski studij računarstva Fakulteta strojarstva, računarstva i elektrotehnike Sveučilišta u Mostaru, odobrenim prelaskom se upisuje kao student prijediplomskog sveučilišnog studija računarstva Fakulteta elektrotehnike, računarstva i informacijski tehnologija Osijek. Završetkom prijediplomskog sveučilišnog studija računarstva upisuje diplomski studij računarstva, izborni blok Informacijske i podatkovne znanosti na Fakultetu elektrotehnike, računarstva i informacijski tehnologija Osijek. Tijekom studiranja volontira kao član neprofitne studentske udruge HUMRSPTZ (engl. *IAESTE Croatia*) u kojoj je obnašao funkciju Fundraising koordinatora lokalnog odbora Osijek te je sudjelovao u izradi i provedbi raznih studentskih projekata.

_____

Potpis autora

PRILOZI

P.3.1. Programska podrška vezana uz obradu podataka i treniranje mreže, dostupna putem poveznice: https://github.com/LukaMarkic/speaker-verification.git.

P.3.2. Programska podrška aplikacije za verifikaciju govornika, dostupna putem poveznice: https://github.com/LukaMarkic/speaker-verification-desktop-app.git.