Automatsko prepoznavanje izgovorenih cifara na bosanskom jeziku korištenjem vještačkih neuronskih mreža

Muamer Parić, Isam Vrce, Armin Žunić
Odsjek za automatiku i elektroniku
Elektrotehički fakultet
Univerzitet u Sarajevu
Sarajevo, Bosna i Hercegovina
mparic2@etf.unsa.ba, ivrce1@etf.unsa.ba, azunic1@etf.unsa.ba

***Sažetak*—Prepoznavanje govora je oblast koja se kroz historiju zasnivala na različitim konceptima. Oni koji su se u prošlosti najviše izdvajali od drugih su obično bili bazirani na skrivenim Markovljevim modelima. U skorije vrijeme neuronske mreže, kao nezaobilazna stanica, su našle primjeru i u ovoj oblasti. Zajedno za ekstrakcijom pogodnih koeficijenata iz audio signala, pokazale su se mnogo uspješnije od prethodnih modela. Upravo će kroz rad biti pokazana tačnost neuronskih mreža na ograničenom skupu riječi koje je potrebno prepoznati. U radu je prikazan rezultat implementacije stabla odlučivanja, klasične umjetne neuronske mreže sa gusto povezanim slojevima, te dvije varijante konvolucione neuronske mreže za suočavanje sa problemom automatskog prepoznavanja govora.**

***Ključni pojmovi*—FF-ANN, CNN, automatsko prepoznavanje govora**

***Abstract*—Speech recognition is an area of study which has through history relied on various concepts. Those which have traditionally stood out the most have been based on hidden Markov chains. More recently artificial neural networks, as an evermore prevelant tool, have found a use in this area as well. Together with extraction of adequate coefficients from audio signals, they have shown to be much more successful than earlier models. This work will showcase neural networks and their accuracy on a limited dataset of words to be recognized. The work showcases the results of implementing a decision tree, a traditional artificial neural network with densely connected layers, as well as two variants of convolutional neural networks for tackling the problem of automatic speech recognition.**

***Keywords*—FF-ANN, CNN, automatic speech recognition**

I. Uvod

Već nekoliko decenija neuronske mreže se koriste za prepoznavanje uzoraka i rješavanje problema za koje je teško definisati algoritam rješavanja u klasičnom smislu [1]. Inspirisani inovativnim radom [2] koji je objavljen 1975. i koji je predstavio neuronsku mrežu treniranu za prepoznavanje cifara napisanih rukom, autori ovog rada su odlučili baviti se problemom prepoznavanja izgovorenih cifara. Riječ je o tipu prepoznavanja govora izoliranih riječi [[3], [4]].

Ovo je osnovni zadatak koji služi kao uvod u primjenu neuronskih mreža za automatsko prepoznavanje govora, no u sebi sadrži mnoge izazove i ideje za koje se autori nadaju da će čitaocima moći služiti kao odskočna daska za dalje bavljenje ovim problemom.

Uvidjevši da su javni resursi za treniranje sistema automatskog prepoznavanja govora na službenim jezicima Bosne i Hercegovine izuzetno ograničeni, autori su u sklopu rada mnogo vremena posvetili prikupljanju podataka, odnosno snimanju izgovorenih riječi ”nula”, ”jedan”, ”dva”, ”tri”, ”četiri”, ”pet”, ”šest”, ”sedam”, ”osam” i ”devet”. Ovaj naizgled jednostavan problem vrlo brzo postaje izazov sam za sebe.

Kako se ove riječi ne razlikuju na bosanskom, hrvatskom, srpskom i crnogorskom jeziku tako autori smatraju da je skup podataka primjenljiv na sve te jezike, prihvatajući da u snimcima nije predstavljen puni raspon naglasaka na ovim jezicima pri izgovaranju riječi cifara. Radi jednostavnosti autori se u nastavku teksta referiraju na jezik cifara kao bosanski jezik, u skladu sa područjem sa kojeg su učesnici koji su učestvovali u prikupljanju podataka.

Rad je koncipiran na način da smo prije svega objasnili postupak prikupljanja podataka. Nakon toga algoritme, odnosno postupke za dobijanje korisnih podataka koji će se koristiti u postupku treniranja neuronske mreže. Opis i strukura neuronskih mreža slijedi nakon toga, te rezultati koji su postignuti. Rad se završava zaključkom u kojem je sumirano ono urađeno u cijelom radu, gdje su predložena moguća poboljšanja.

A. Data collection

Neuronske mreže, ovisno od strukture, podešavaju veliki broj parametara da bi transformirali ulaznu sekvencu signala u koristan izlaz. U ovisnosti od strukture mreže i broja internih parametara, potreban je različit broj instanci skupa podataka za treniranje. Da bi omogućili da jednom istrenirana mreža može prepoznavati cifre izgovorene od strane nepoznatih ljudi, autori su se potrudili da skup podataka za treniranje bude što je moguće brojniji i raznovrsniji.

Podaci za treniranje mogu biti označeni i ne označeni [5]. Autori su za treniranje neuronske mreže koristili označeni skup podataka. Ovo za posljedicu ima da će jednom prikupljene podatke biti potrebno označiti. Odnosno ekspertna osoba (u

ovom slučaju autori rada) će na neki način dati značenje svakom podatku korištenom za treniranje.

Proces prikupljanja podataka je u literaturi poznat pod nazivom: "Data collection"[6]. Pojmovi koji se koriste u ovoj oblasti nema smisla prevoditi na bosanski jezik. Dodatno pored toga se određeni pojmovi u našem jeziku često prevode istim riječima. Bilo kako bilo, Data collection se sastoji iz nekoliko poddijelova, a prvi od njih jeste "Data acquisition", ili u prijevodu prikupljanje podataka, tako da su podaci raznoliki i da postoji dovoljan broj podataka. Drugi dio predstavlja "Data labeling" ili označavanje podataka. Kao što je ranije naglašeno autori će za treniranje neuronskih mreža koristiti označene podatke, te konačno "Existing data", odnosno integracija s postojećim podacima.

S obzirom na to da je cilj kreirati model koji će biti u stanju raspoznavati cifre izgovorene na bosanskom jeziku, "Existing data" dio je nepostojeći, odnosno autori nisu uspjeli pronaći postojeći skup podataka. Zaključuje se da je potrebno prikupiti podatke "od nule", a u ovom poglavlju će biti govora upravo o načinu prikupljanja podataka.

Prije svega je bilo potrebno odabrati način akvizicije podataka. Prvobitno su autori akvizirali podatke korištenjem "podcast mikrofona" visokih performansi. Međutim, ubrzo je shvaćeno da s ovim pristupom postoji nekoliko problema. Prvi je činjenica da je potrebno prikupiti veliki broj snimaka od različitih osoba pa je akvizicija podataka otežana. Drugi problem je činjenica da uslovi u kojima se akviziraju podaci ne oslikavaju stvarne uslove. Zbog toga su se autori odlučili da se podaci prikupljaju preko diktafona na mobilnim uređajima. Na ovaj način se bolje oslikavaju stvarni uslovi korištenja, te je jednostavnije prikupiti veći broj uzoraka. Još jedan važan parametar prilikom prikupljanja podataka je frekvencija uzorkovanja. Većina mobilnih telefona koristi istu frekvenciju uzorkovanja (44100 Hz). Međutim, neki mobilni telefoni koriste veću frekvenciju uzorkovanja (48000 Hz), pa je takve uzorke bilo potrebno resemplirati na frekvenciju 44100Hz.

Nakon što je odabran način uzorkovanja (diktafon na mobilnom telefonu) bilo je potrebno osmisliti način prikupljanja podataka, koji je to broj uzoraka dovoljan da bi se osiguralo da će mreža biti adekvatno istrenirana, osigurati raznovrsnost uslova snimanja, pozadinska buka, ujednačenost spolova za skup za treniranje, raznovrsnost dobne skupine.

Prije svega 6 osoba (3 osobe muškog spola i tri osobe ženskog spola) su čitale prvih 400 decimalnih cifara broja π. Tako je prikupljeno 2400 cifara koje su manje-više uniformno raspodijeljene. Od 400 cifara koje je svaka osoba pročitala, 200 cifara je snimljeno u "idealnim uslovima". Pod pojmom "idealni uslovi" misli se na odsustvo pozadinske buke. Narednih 100 cifara je snimljeno uz postojanje male pozadinske buke, ostatak je snimljen uz pozadinsku buku veće amplitude. Pozadinska buka je u ovom slučaju predstavljala različito "šuškanje", "klepetanje", zvukovi kućanskih aparata (napa, zvuk s laptopa), zvuk automobila na ulici. Ispitanici su čitali cifre broja π da bi se otklonio fenomen poznavanja sljedeće cifre. Naime, ako osoba čita brojeve redoslijedom (0, 1, 2,...) onda postoji tendencija da se cifre "nastavljaju jedna na drugu" što nije poželjno. A cifre broja π su poredane u slučajnom redoslijedu. Dodatno kasnije označavanje cifara je mnogo jednostavnije jer unaprijed znamo cifre koje su ispitanici čitali.

Sa ovim skupom podataka autori su stvorili bazu koja ima prosječno 240 instanci svake cifre. Ovakvi podaci su dobro balansirani što se tiče spolova i postojanja različitih uslova (pozadinske buke). Međutim raznovrsnost nije bila zastupljena. Odnosno mreža koja bi se trenirala nad ovim podacima bi dobro prepoznavala kada neko od 6 ispitanika nanovo izgovori neku cifru. Međutim ukoliko bi osoba koja nije učestvovala u snimanju izgovorila neku cifru, vjerovatno ista ne bi bila dobro prepoznata, ili bi tačnost prepoznavanja bila manja.

Zbog toga se naredni skup podataka sastojao od 115 slučajno raspoređenih cifara koje je izgovaralo 20 osoba. Odabrano je 115 cifara da bi se izbalansirao broj cifara kojih ima više u data setu. Osobe koje su izgovarale ove cifre su bile različitih spolova i godina (od 7 godina do 77 godina). Također, polovina cifara je izgovorena u "idealnim uslovima", a ostatak u "vještački generiranom pozadinskom bukom". Vještački generirana pozadinska buka predstavlja buku koja je rezultat zvuka s televizora ili laptota. Ovim je u bazu dodato 2200 cifara.

Do sada snimljeni podaci bi trebali generirati neki generalni model sistema. Da bi se osigurala robusnost mreže, dodatno smo zamolili približno 70 ispitanika da izgovore 20 cifara koji su snimani u najtežim uslovima pozadinske buke. Govorimo o tome da smo 30 procenata ispitanika snimali u kafićima, 30 procenata u prostorijama gdje su drugi ljudi igrali stoni tenis, na ulici, na hodniku fakulteta. Ostatak podataka (40 procenata) je snimljeno u manje ekstremnim uslovima pozadinske buke. Pored ovoga kolega s drugog fakulteta nam je ustupio oko 500 cifara koje je izgovaralo 10 ljudi. Dakle, svaka osoba je izgovorila po 50 cifara.

Nakon podataka za treniranje, u različitim uslovima je snimljeno dodatnih 300 snimaka (svake cifre po 30 uzoraka) koje će služiti za testiranje. Važno je napomenuti da osobe koje su izgovarale cifre u skupu za testiranje se ne nalaze u skupu za treniranje. Na ovaj način smo željeli simulirati rad mreže u stvarnom svijetu.

Sve u svemu, na kraju su autori na raspolaganju imali oko 2 sata izdvojenih snimaka, te 1.8 GB audio materijala gdje je odstranjena tišina. Odnosno, ono što se na ovim fajlovima čuje su samo korisni signali izgovora cifara. U skupljanju podataka je učestvovalo više od 120 različitih osoba kojima se ovim putem svima zahvaljujemo.

B. Izdvajanje pojedinačnih cifara iz snimka

Snimci pojedinačnih cifara se izdvajaju iz dužeg snimka sa više cifara heurističnim algoritmom koji se zasniva na thresholdingu. Izdvajanje se zasniva isključivo na amplitudi pritiska zraka, odnosno glasnoće zvuka. Prelazak amplitude preko vrijednosti koju određuje parametar algoritma označava početak riječi, odnosno cifre. Završetak cifre je određen tra-

janjem snimka bez ijednog uzorka u kojem glasnoća zvuka prelazi taj isti prag glasnoće. Ovo rezultira granicama riječi.

Dva dodatna koraka su učinjena kako bi se poboljšalo izdvajanje snimka. Ukoliko je trajanje izdvojene riječi previše kratko to ukazuje na potencijalan šum koji je izdvojen, a ne izgovorena riječ. Također, ova metoda u prisustvu šuma rezultira pretjerano agresivnom izdvajanju koje odsječe početak i kraj riječi. Iz ovog razloga, ako je snimak dovoljno dug da bi se smatrao da sadrži riječ onda se granice snimka simetrično proširuju u oba smjera za parametrom zadano trajanje kako bi se obuhvatila kompletna izgovorena riječ.

Adekvatno izdvajanje riječi ovakvim algoritmom ovisi od parametara za minimalnu dužinu riječi, maksimalnu glasnoću šuma, minimalno trajanje između izgovorenih riječi i vrijeme simetričnog proširivanja snimka. Ove parametre je bilo potrebno ručno podešavati u ovisnosti od specifičnosti snimka. Kako su korišteni snimci napravljeni u raznim lokacijama, pod raznim uvjetima i sa raznim uređajima nije bilo moguće sa ovom metodom odrediti univerzalni skup vrijednosti parametara.

Autori su primjetili da je adekvatan vremenski razmak između riječi bio ključan faktor za olakšano izdvajanje riječi, a da je amplituda šuma, donekle binarno, potpuno onemogućavala izdvajanje riječi. Ukoliko je šum veće amplitude od najtiše izgovorene riječi onda ovaj algoritam nije mogao izdvojiti riječi bez izdvajanja šuma.

C. Struktura skupa za treniranje i testnog skupa

Nakon što su podaci prikupljeni i iz njih adekvatno izdvojeni korisni dijelovi snimaka, dobijen je konačni skup podataka koji će biti korišten za treniranje, validaciju i testiranje neuronskih mreža. Svi podaci za treniranje se u ovom trenutku nalaze u folderu "treniranje" u kojem se nalazi 10 foldera sa nazivima svih 10 cifara. U folderima se u prosjeku nalazi 630 ".wav" fajlova na kojima se nalaze izgovorene cifre.

D. Pretprocesiranje podataka

Prije nego li se krene sa bilo kakvom obradom audio snimaka, iste je potrebno pretprocesirati. To bi značilo da je svaki snimak potrebno normalizirati, te iz njega izvući određeni set podataka koji će se kasnije koristiti za treniranje, validaciju i testiranje.

Prije svega, za učitavanje audio snimaka u Python okruženje se koristila biblioteka *librosa* [7]. Ista biblioteka je bila okosnica rada i za preostalo pretprocesiranje. Nakon učitanog audio snimka, isti je normaliziran min-max normalizacijom, korištenjem funkcijom *util.normalize* iz pomenute biblioteke. Ovo bi značilo da se amplitude svakog snimka kreću između -1 i 1. Dalje se iz svakog normaliziranog snimka izvlačio određeni broj koeficijenata koji su bili ulaz u neuronsku mrežu.

Riječ je o Mel-frequency cepstral coefficients (MFCCs) [8]. Izdvajanje ovih koeficijenata se bazira na malo kompleksnijoj matematici, čija je osnova Fourierova transformacija, ali uz korištenje gotovih funkcija iz već pomenutih biblioteka ni taj dio nije problematičan. Formiranje konačnih koeficijenata koji će činiti ulaz u mrežu se može podijeliti u nekoliko koraka:

- Podjela ulaznog signala na dijelove (20 dijelova za potrebe rada je bilo sasvim dovoljno), pri čemu će svaki dio činiti određeni broj uzoraka (frejmova) signala. Preveliki broj dijelova signala bi značio izdvajanje mnogo nepotrebnih karakteristika signala kao što su šum prije i poslije "bitnog" dijela signala, dok bi premali broj dijelova signala značio mali broj ulaza u mrežu, što se pokazalo kao veoma nepoželjno.
- Ispravan odabir parametra Nyquistove frekvencije uzevši u obzir glasovne mogućnosti osoba koje su posuđivale glas u ovom radu, a eksperimentalno je utvrđeno da je dovoljno uzeti da je ta frekvencija jednaka 8192 Hz.
- Za svaki tih dijelova signala se računaju MFCC koeficijenti, a broj tih koeficijenata se obično uzima 13, što se u ovom radu pokazalo kao sasvim dovoljno. Autori su pokušali uzeti nešto manji broj koeficijenata, međutim, sa smanjenjem tog broja procenat uspješnosti opada.
- Ako se želi postići veća tačnost (kroz izradu rada se moglo primijetiti) poželjno je pored osnovnih MFCC koeficijenata koristiti i njihov prvi izvod, pa se tako nakon računanja koeficijenata poziva funkcija koja prima te koeficijente te traži njihov izvod. Ovim se broj ukupnih koeficijenata za jedan dio signala udvostručava.
- Nakon dobijenih svih koeficijenata za svaki od dijelova signala, vrši se normalizacija ovih koeficijenata. Prvi koeficijenti su mnogo veće amplitude od posljednjih. Ovim je svaki od 13, odnosno, 26 koeficijenata imao istu ili sličnu ulogu (u svrhu ovoga je korišten *StandardScaler* iz biblioteke *sklearn.preprocessing*).

Nakon prethodno provedenog postupka, koeficijente svakog signala je moguće zapisati u csv fajl koji je najpogodniji oblik zapisa podataka za kasnije algoritme, što omogućava biblioteka *csv*. Nakon ovoga podaci su spremni. Pretprocesiranje je završeno, te se može krenuti sa klasifikacijom audio snimaka.

II. Decision tree

A. Korišteni algoritam

Prethodno normalizirani koeficijenti su korišteni u nastavku za klasifikaciju cifara. Prije korištenja neuronske mreže kao klasifikatora testirane su mogućnosti klasifikacije korištenjem drveta odlučivanja. Ovo je korišteno kako bismo imali neku refernecu. Korištena je implementacija iz Pythona *DecisionTreeClassifier*.

B. Simulacijski rezultati

Autori su pokušali različite kriterije, te različite postavke parametara, također, i bez normalizacije audio snimaka i bez normalizacije koeficijenata. Naravno, normalizacija je pokazala mnogo bolje rezultate, dok je najbolji kriterij u samom stablu odlučivanja bio "gini", gdje se kao parametar *splitter* koristio "best". Uz ovakvu postavku klasifikatora procenat uspješnosti nad testnim podacima je $\approx 80\%$. Od ovog procenta se krenulo u realizaciju neuronskih mreža, i on je služio kao

neki vid donje granice ispod koje neuronska mreža ne bi imala smisla.

III. FF-ANN neuronska mreža

A. Korišteni algoritam

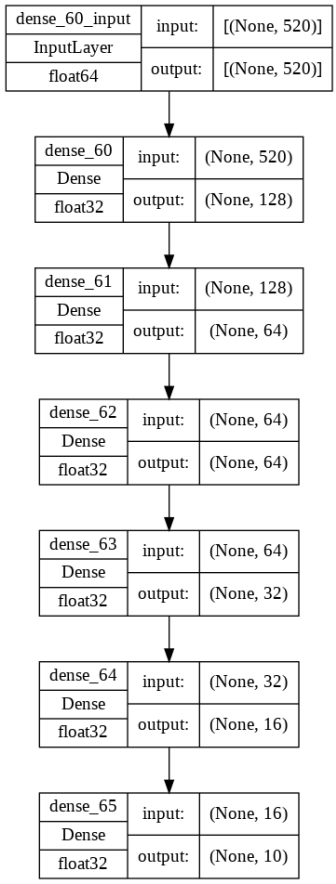
Nakon što je izvršena klasifikacija korištenjem drveta odlučivanja, sljedeći korak bio je klasifikacija cifara korištenjem guste vještačke neuronske mreže (FF-ANN). U radu je izvršena klasifikacija korištenjem guste neuronske mreže, kao i klasifikacija korištenjem konvolucionih neuronskih mreža (CNN).

Prilikom pokušaja klasifikacije podataka pomoću FF-ANN korištene su različite strukture, različite aktivacijske funkcije, te različit broj parametara FF-ANN. Autori su testirali više od 30 različitih FF-ANN da bi pronašli onu koja daje najbolje rezultate. Nakon prethodno pročitane rečenice odmah se može postaviti pitanje kako odrediti koja je struktura FF-ANN bolja od druge. Prije svega bitno je naglasiti da se procjena "bolje" neuronske mreže vrši isključivo nad test skupom podataka koje istrenirana mreža nije nikada prije "vidjela". Jasno, niti validacijske podatke mreža nikada prije nije vidjela, ali u validacijskim podacima se potencijalno nalaze i uzorci istih osoba koje izgovaraju iste cifre koje se nalaze u trening skupu. U test skupu se nalaze podaci od osoba koje mreža nikada nije "vidjela". Nakon što je odabran skup podataka nad kojim će se vršiti testiranje potrebno je odabrati metriku po kojoj će se birati bolja mreža. Često se treniranjem iste neuronske mreže dobiju različite tačnosti nad test skupom podataka (jer se kao krening skup podataka uzimaju slučajni uzorci). Zbog toga je treniranje vršeno više puta uzastopno (29 puta za svaku odabranu neuronsku mrežu) i računata je prosječna tačnost. Tako da će se kao metrika za određivanje bolje neuronske mreže koristiti upravo srednja tačnost nad višestrukim treniranjem.

Obzirom da je testiran veliki broj različitih neuronskih mreža, sa različitim parametrima (broj MFCC koeficijenata koji se računaju, broj dijelova izvornog snimka, da li se koriste izvodi MFCC koeficijenata, Nyquistova frekvencija, različiti hiperparametri FF-ANN), autori ne smatraju za bitno da se ovdje navode rezultati koji su lošiji od najboljeg. Svakako, ukoliko čitaoca zanimaju neki razultati neka bude slobodan da se obrati putem maila bilo kojem autoru.

Kao što je u pogavlju prije iznesena pretpostavka da je najbolje koristiti 13 MFCC koeficijenata, a da se pri tome ulazni signal nad kojim se računaju MFCC koeficijenti podijeli na 20 dijelova, te Nyquistovom frekvencijom od 8192 Hz, ovdje su te pretpostavke potvrđene. Nakon što je nad tri različite strukture neuronskih mreža potvrđeno da ovi ulazi daju najbolje rezultate, u ostatku testiranja je mijenjana struktura, te hiperparametri mreže. Nakon testiranja većeg broja struktura i hiperparametara najbolje rezultate je dala mreža koja ima strukutru kao na slici (Slika 1).

FF-ANN posjeduje 520 ulaza sa 4 skrivena sloja, sve aktivacijske funkcije su *relu* [9], osim posljednje koja ima aktivacijsku funkciju *softmax* [9]. *Softmax* se koristi kao izlazna aktivacijska funkcija, jer je u pitanju klasifikacija iz



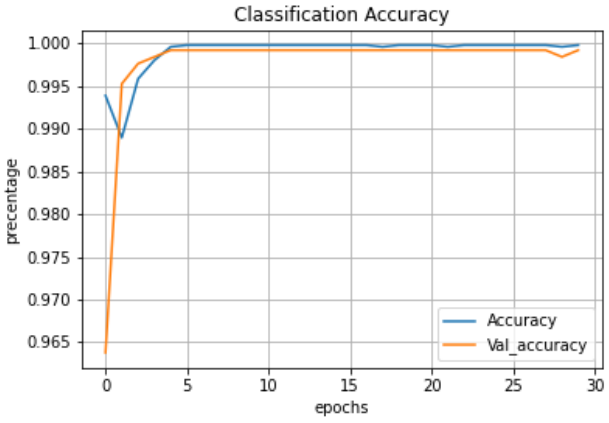
Slika 1. Struktura FF-ANN koja je dala najbolje rezultate

konačnog broja zvučnih audio signala od kojih neki mogu biti "slični", pa se na ovaj način vrši normiranje izlaza. Prilikom odabira broja neurona u pojedinim slojevima, uzeto je u obzir da na raspolaganju imamo oko 6000 audio signala. Pa broj unutrašnjih parametara nije mogao biti prevelik. Zbog toga je, nakon više iteracija, odabran je broj neurona po slojevima kao na slici (Slika 1).

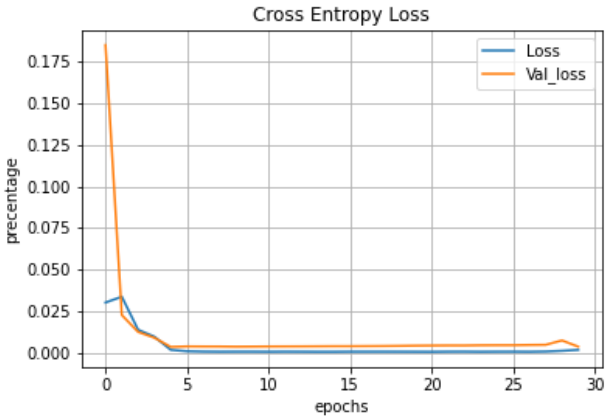
1) Odnos uzoraka za treniranje, validaciju i testiranje: Već je u poglavljima koje su prethodile bilo govora o setu podataka za treniranje i setu (odnosno skupu) podataka za testiranje. Od ukupnog broja uzoraka koji je snimljen, a koji je planiran za "treniranje", 80% uzoraka je korišteno ta treniranje, 20% podataka je korišteno za validiranje. Ovo znači da se za treniranje koristilo 5077, a za validaciju 1269. Obzirom da smo željeli testirati mrežu nad ciframa izgovorenim od nepoznatih osoba snimljeno je dodatnih 300 uzoraka. Ovo je otprilike 5% ukupnog broja uzoraka.

B. Simulacijski rezultati

Sa grafika (Slika 2) vidimo da je je tačnost nad trening skupom podataka nakon 5 epoha dospjela na 100%, te da je tačnost nad validacijskim skupom podataka također blizu 100%. Zaključak koji se dodatno može donijeti jeste činjenica da je treniranje moglo stati na 5. epohi, jer se nakon pete epohe tačnost nije promijenila. Slični zaključci se mogu donijeti i za



Slika 2. Procenat uspješnosti FF-ANN kroz epohe



Slika 3. Greška FF-ANN kroz epohe

grešku (Slika 3). Nakon pete epohe greška ostaje konstantna. U konkretnom slučaju nad testnim skupom podataka tačnost je bila 97.33% što je autorima bio zapanjujući podatak, jer je FF-ANN pogrešno klasificirala samo 8 uzoraka od njih 300. Uslovi u kojima su snimljeni ti uzorci su bili relativno mnogo otežani. Autori pozivaju radoznale čitaoce da se jave putem maila da poslušaju neke od uzoraka koje je mreža pogrešno klasificirala.

Da bi se dobio bolji uvid u tačnost mreže, treniranje je pokrenuto 29 puta i izračunata je srednja tačnost koja je iznosila 95.07%. Razlog zbog kojeg je tačnost nakon 29 treniranja manja nego najbolja od tih 29, je taj što je skup snimaka za treniranje svaki put odabran slučajnim putem. Zbog toga se desi da je skup podataka za treniranje "lošije odabran".

IV. Konvolucione neuronske mreže

A. Jednodimenzionalne konvolucione neuronske mreže

1) Korišeni algoritam: Još jedan pristup koji su autori istražili jeste korištenje konvolucionih neuronskih mreža. Konvolucione neuronske mreže se primarno koriste za prepoznavnje uzoraka unutar slika [10]. Bez obzira na to, autori su istražili primjenu konvolucionih neuronskih mreža na prepoznavanje cifara sa dva različita pristupa.

Jednostavniji pristup je direktno pohranjivanje zvučnog zapisa u neuronsku mrežu. Jedina obrada koja je učinjena prije ovoga je normalizacija trajanja zvučnih zapisa na jednu sekundu, te normalizacija amplitude.

Postupak normalizacije trajanja je razdvojen na dva slučaja. Ukoliko je izvorni snimak cifre, izdvojen na prethodno opisani način, kraći od jedne sekunde onda se vrši simetrično širenje signala na obje strane tišinom, odnosno uzorcima nulte amplitude. U drugom slučaju, ukoliko je izvorni snimak cifre duži od jedne sekunde onda se simetrično skraćuje snimak na jednu sekundu, odnosno na 44100 uzoraka.

Normalizacija po amplitudi je učinjena tako što su vrijednosti svih pozitivnih uzoraka podijeljene sa pozitivnim uzorkom najveće amplitude, a svi negativni uzorci podijeljeni sa apsolutnom vrijednosti amplitude negativnog uzorka sa najvećom apsolutnom vrijednosti. Ovo rezultira signalom čiji svi uzorci nisu manji od -1 niti veći od 1.

2) Simulacijski rezultati: Rezultirajući signali, svi sa 44100 uzoraka su predstavljali ulaze u neuronsku mrežu. Autori su probali razne arhitekture neuronske mreže koristeći do četiri konvoluciona sloja prije, do četiri gusta sloja, ali nisu postigli tačnost veću od 26%. Glavni problem koji je identificiran je *overfitting*. Magnituda ovog problema je ovisila i od veličine skupa podataka za treniranje, ali ni korištenje svih dostupnih podataka za treniranje nije dovelo do zadovoljavajućih performansi mreže.

B. Dvodimenzionalne konvolucione neuronske mreže

1) Korišteni algoritam: U svrhu postizanja boljih performansi konvolucionim neuronskim mrežama svi MFCC koeficijenti dobijeni iz snimka jedne riječi su poredani u dvodimenzionalnu mrežu. Svaka tačka na mreži predstavlja vrijednost jednog izračunatog koeficijenta, gdje jedna dimenzija predstavlja vremenski pomak od početka, a druga dimenzija predstavlja indeks koeficijenta koji je izračunat.

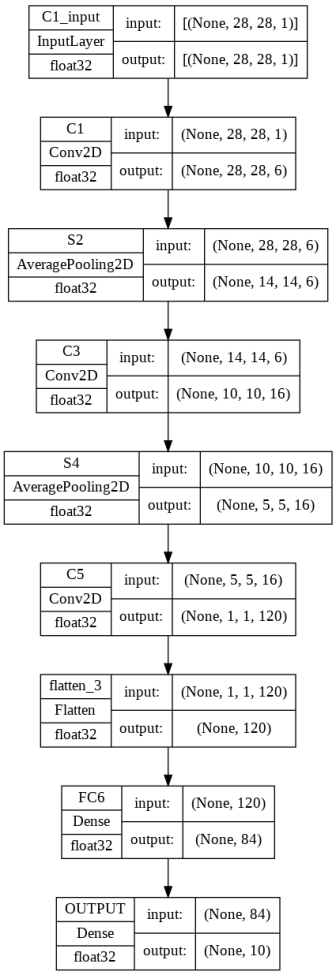
Na primjer, ukoliko se pri računanju MFCC koeficijenata snimak dijeli na 20 dijelova jednakog trajanja i iz svakog se izdvaja 13 koeficijenata onda se formira mreža, odnosno slika, dimenzija 20×26. U prvom redu se nalaze svi koeficijenti vezani za najraniji dio snimka, u drugom za idući i tako dalje do posljednjeg, 20. reda gdje se nalaze svi koeficijenti izračunati iz posljednjeg izdvojenog dijela snimka.

Autori su podijelili implementaciju dvodimenzionalne konvolucione neuronske mreže za rješavanje problema na dva različita pristupa. Prvi pristup koristi samo osnovne MFCC koeficijente, dok drugi pristup uz njih koristi i izvode MFCC koeficijenata.

2) Simulacijski rezultati: Prvi pristup nije pokazao zadovoljavajuće rezultate, ali drugi pristup je generisao mreže sa prosječnom tačnošću od 94%, a najbolja dobijena mreža je imala tačnost od 96%. Mreža se delta spektralnim koeficijentima je postavljena uz mrežu sa statičnim koeficijentima,

čime je dobijena mreža, odnosno slika, sa duplo više kolona od pojedinačnih slika.

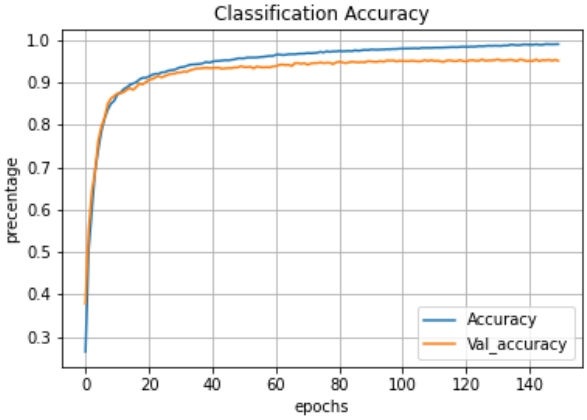
U ovu svrhu korištena je adaptirana LeNet arhitektura neuronske mreže [11]. Adaptacija autora se ogleda u prilagodbi dimenzija ulaznih parametara kako bi se poklapale sa dimenzijama generisanih slika koje sadrže odgovarajuće MFCC koeficijente. Struktura mreže je data na slici (Slika 4).



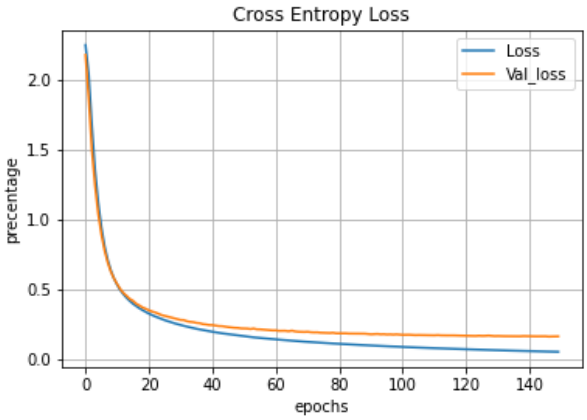
Slika 4. Struktura CNN koja je dala najbolje rezultate

Sa grafika (Slika 5) vidimo da se tačnost nad trening skupom podataka nakon 140 epoha približila 100%, te da je tačnost nad validacijskim skupom podataka približila 95%. Bilo je moguće pustiti da se treniranje vrši još veći broj epoha. Međutim, primjetno je da je tačnost nad validacijskim podacima u posljednih 50-ak epoha konstantna, te bi daljnje treniranje bilo suvišno. Slični zaključci se mogu donijeti i za grešku (Slika 6). Nakon 100. epohe greška nad validacijskim podacima se ne smanjuje, ostaje konstantna. U konkretnom slučaju nad testnim skupom podataka tačnost je bila 96%, što je dobar rezultat, ali ne kao kod predložene FF-ANN mreže.

Ono što se dodatno može zaključiti jeste činjenica da CNN mreža ima manju razliku između prosječne i najveće tačnosti. Pa je tako prosječna tačnost nad 29 različitih treniranja 94%.



Slika 5. Procenat uspješnosti CNN kroz epohe



Slika 6. Greška CNN kroz epohe

V. Zaključak

Unatoč velikim naprecima u oblasti u prošlih par decenija problem automatskog prepoznavanja govora još uvijek predstavlja oblast sa velikim potencijalom za napredak i istraživanje. U radu je prikazano nekoliko pristupa ovom problemu sa različitim stepenima uspješnosti.

Pristupi korištenjem *mel-frequency capstral* koeficijenata su dominantni u odnosu na pristupe koji se baziraju na vremenskim uzorcima ili jednostavnim manipulacijama frekventnog spektra signala [12]. Autori su pokazali uspješnu izvedbu automatskog prepoznavanja govora koristeći mel-frequency capstral koeficijente, ali ne i uspješnu izvedbu koristeći vremenske uzorke signala.

Kako bi se moglo izvesti prepoznavanje govora na bosanskom jeziku autori su snimili i obradili preko 6000 snimaka cifara koje smatraju značajnim dijelom ovog rada zato što prije objave ovog rada nisu mogli naći adekvatan skup podataka za treniranje automatskog prepoznavanja govora na bosanskom jeziku.

LITERATURA

[1] Bohdan Macukow. "Neural networks–state of art, brief history, basic models and architecture". *IFIP international conference on computer information systems and industrial management*. Springer. 2016., str. 3–14.

[2] Kunihiko Fukushima. "Cognitron: A self-organizing multilayered neural network". *Biological cybernetics* 20.3 (1975.), str. 121–136.

[3] Santosh K Gaikwad, Bharti W Gawali i Pravin Yannawar. "A review on speech recognition technique". *International Journal of Computer Applications* 10.3 (2010.), str. 16–24.

[4] Geoffrey Hinton i dr. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". *IEEE Signal processing magazine* 29.6 (2012.), str. 82–97.

[5] Oludare Isaac Abiodun i dr. "State-of-the-art in artificial neural network applications: A survey". *Heliyon* 4.11 (2018.), e00938.

[6] Yuji Roh, Geon Heo i Steven Euijong Whang. "A survey on data collection for machine learning: a big data-ai integration perspective". *IEEE Transactions on Knowledge and Data Engineering* 33.4 (2019.), str. 1328–1347.

[7] Brian McFee i dr. "librosa: Audio and music signal analysis in python." *In Proceedings of the 14th python in science conference* (2015.), str. 18–25.

[8] S. Suksri C. Ittichaichareon i T. Yingthawornsuk. "Speech Recognition using MFCC". *International Conference on Computer Graphics, Simulation and Modeling* (2012.), str. 4.

[9] Sagar Sharma, Simone Sharma i Anidhya Athaiya. "Activation functions in neural networks". *towards data science* 6.12 (2017.), str. 310–316.

[10] Keiron O'Shea i Ryan Nash. "An introduction to convolutional neural networks". *arXiv preprint arXiv:1511.08458* (2015.).

[11] Yann LeCun i dr. "LeNet-5, convolutional neural networks". *URL: http://yann. lecun. com/exdb/lenet* 20.5 (2015.), str. 14.

[12] Lindasalwa Muda, Mumtaj Begam i Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques". *arXiv preprint arXiv:1003.4083* (2010.).