Detekcija i označavanje zvukova na audio snimku korišćenjem metoda mašinskog učenja

Mihajlo Perendija
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
perendija.e274.2020@uns.ac.rs

Petar Bašić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad, Srbija
petar.basic@uns.ac.rs

***Apstrakt* — Detekcija pojavljivanja određenog zvuka na audio snimku i označavanje vremena njegovog početka i kraja predstavlja problem čije rešavanje bi imalo pozitivan uticaj na ponašanje raznih sigurnosnih i korisničkih sistema. Metodologija kojom se u radu pristupilo rešavanju ovog problema podrazumeva transformaciju audio snimaka u pogodniji oblik za mašinsku analizu, kreiranje modela za detekciju određenog zvuka i modela za označavanje intervala u kom je taj zvuk aktivan. Za potrebe oba modela korišćene su potpuno povezane neuronske mreže, rekurentne neuronske mreže i konvolutivne neuronske mreže. Nakon izvršenih eksperimenata dobijeni su obećavajući rezultati za oba zadatka gde se pri detekciji zvuka ističu konvolutivne neuronske mreže a rekurentna neuronska mreža pri označavanju dela sekvence gde je određeni zvuk aktivan. Primećeno je da dužina sekvence nad kojom modeli rade ima značajan uticaj na rezultate a izbor te dužine bi takođe zavisio i od upotrebe samog sistema koji integriše modul za rešavanje opisanih problema.**

Ključne reči* — *zvuk; detekcija; trajanje; sekvenca; klasifikacija

I. Uvod

Problem automatskog određivanja događaja na audio snimku obuhvata detekciju postojanja specifičnog zvuka i vremenskog intervala tokom kog je zvuk trajao. Rešavanjem ovog problema otvaraju se mogućnosti za značajno unapređenje tehnologija koje se upotrebljavaju u pametnim kućama, sigurnosnim sistemima, sistemima za prepoznavanje govora, za forenzičku analizu, kao i u raznim IoT (*internet of things*) uređajima.

Ovaj rad bavi se razvijanjem modela za detekciju zvuka i određivanje vremenskog intervala njegovog trajanja. Predložena metodologija obuhvata obradu audio snimaka preoblikovanjem u mašinski obradiv oblik, obučavanje tri vrste modela mašinskog učenja za navedene zadatke i naposletku evaluaciju modela i analizu dobijenih rezultata.

Podaci, odnosno audio snimci korišćeni prilikom implementacije predložene metodologije preuzeti su sa javno dostupnog izvora koji je objavljen u sklopu DCASE izazova (*IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*) [1].

U daljim poglavljima detaljnije su izloženi korišćeni podaci, načini njihove pripreme i kreiranja pogodne reprezentacije za primenu metoda mašinskog učenja. U poslednjim poglavljima dat je pregled rezultata primenjenih metoda i izveden zaključak rada.

II. Opis skupa podataka

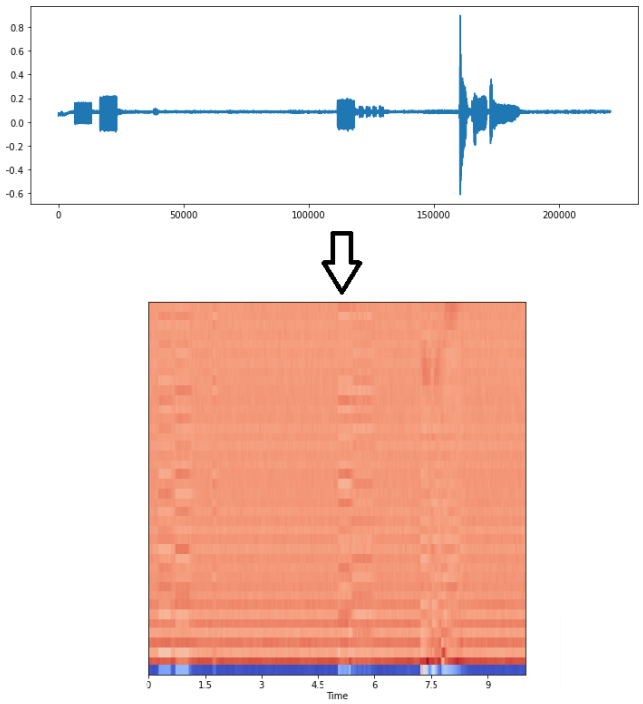
Iz razloga jednostavnosti a istovremeno stavljajući veći fokus na pronalazak najboljih modela mašinskog učenja, odabran je sintetički generisan skup podataka. Audio snimci su preuzeti sa izazova [1], a generisani su korišćenjem alata Scaper [2]. Trajanje generisanih snimaka je 10 sekundi, a na njima se može naći jedan ili više zvukova iz kućnog okruženja i to zvukovi: govora, psa, mačke, alarma, posuđa, blendera, tekuće vode, usisivača, električnog brijača i zvuk prženja hrane. Snimcima su pridružene oznake zvukova koji se na njima mogu čuti, kao i indikatori trenutka početka i kraja trajanja zvuka. Svaki generisani snimak je preslušan i njegove oznake su potvrđene od strane čoveka.

Originalni oblik i format audio snimaka nisu pogodni za mašinsku analizu i obradu, te je potrebno izvršiti transformaciju. Transformacija bi trebalo da bude takva da osobine iz talasnih oblika budu dobro reprezentovane, odnosno izdvojene. Jedna ovakva transformacija predložena je još 1980. godine od strane *Davis i Mermelstein* koji je u svom radu [3] nazivaju *Mel Frequency Cepstral Coefficients* (MFCCs) i pokazala se, tokom vremena, vrlo dobrom za predstavu zvučnih signala. MFCC tehnika obuhvata niz jednostavnijih transformacija:

1. Furijeova transformacija signala,
2. Mapiranje snage spektra dobijenog u koraku 1 na Mel-skalu,
3. Logaritmovanje snage svake frekvencije na Mel-skali,
4. Diskretna kosinusna transformacija liste iz koraka 3, kao da je signal,
5. MFCC vrednosti su amplitude rezultujućeg spektra.

Na slici 1 može se videti grafički prikaz polaznog zvučnog signala i grafički prikaz MFCC osobina tog zvuka.

Treba napomenuti da MFCC vrednosti nisu robusne u prisustvu šuma, što se može zanemariti u slučaju podataka korišćenih u ovom radu, a koji su generisani i šum im je unapred izdvojen. U slučaju realnih snimaka bilo bi poželjno primeniti tehnike kao što su normalizacija MFCC vrednosti i izmena samog algoritma za njihovo dobijanje.



Slika 1 Ilustracija ekstrakcije MFCC karakteristika iz zvučnog signala

Svaki audio primerak iz skupa je primenom algoritma za dobijanje MFCC vrednosti podeljen na (uzorkovan sa) 431 deo. Svaki pojedinačni deo reprezentuje ~0.02 sekunde snimka i sastoji se od niza MFCC vrednosti. Broj vrednosti je podesiv a najbolji rezultati su dobijeni sa 13 MFCC vrednosti po uzorku. Iako bi u opštem slučaju veći broj osobina možda doprineo dobijanju boljih rezultata, u konkretnom slučaju ovog rada povećanje broja vrednosti po uzorku davalo je lošije ili približno jednake rezultate.

Nakon transformacije, odnosno generisanja reprezentacije zvučnih signala u nizove MFCC vrednosti, izvršena je priprema za obučavanje modela mašinskog učenja i to:

za zadatak detekcije postojanja zvuka na snimku:

- Pretvaranje oznaka pridruženih snimku u multi-hot vektore čija je dužina jednaka broju klasa, odnosno broju mogućih zvukova. Zatim, izvršeno je uparivanje generisanih vektora sa odgovarajućim snimcima.

za zadatak određivanja vremenskog intervala trajanja pojedinačnih zvukova na snimku:

- Grupisanje reprezentacija snimaka na osnovu jedinstvenih oznaka koje su im pridružene, a u cilju izgradnje modela mašinskog učenja za prepoznavanje konkretne oznake (zvuka sa snimka).
- Pretvaranje indikatora početka i kraja trajanja zvuka na snimku u multi-hot vektore dužine faktora uzorkovanja (431 u ovom slučaju), a koji sadrže jedinice na pozicijama koje odgovaraju vremenskom intervalu gde se zvuk može čuti i nule na svim ostalim pozicijama.
- Opciono, radi potencijalnog poboljšanja rada modela, izvršena je podela originalne sekvence na više manjih (npr. sekvenca od 431 uzorka na 10 sekvenci od po 44 uzorka). Ovakvom podelom se drastično povećava broj sekvenci nad kojima se model obučava, dok se njihova dužina smanjuje.

III. METODOLOGIJA

Nakon generisanja, obrade i pripreme podataka primenjeno je nekoliko metoda mašinskog učenja, koje se mogu grupisati na osnovu zadatka koji se rešava: detekcija postojanja zvuka (odnosno klasifikacija snimaka na osnovu zvukova koji se pojavljuju) i određivanje vremenskog intervala trajanja zvuka.

U slučaju postojanja više klasa, odnosno zvukova, postoje dva arhitekturalna pristupa pri kreiranju klasifikacionog modela. Prvi pristup jeste kreiranje *multi-class* modela, na čijem izlazu se nalazi jedna izabrana klasa od N mogućih (odnosno verovatnoće za svaku klasu pojedinačno), a drugi pristup je kreiranje ansambla od N modela gde se za svaku klasu obučava zaseban model koji nam govori da li taj podatak pripada toj klasi ili ne.

A. *Metode za zadatak detekcije postojanja zvuka*

Za rešavanja zadatka detekcije postojanja zvuka na snimku isprobana su tri modela mašinskog učenja u režimu *multi-class* klasifikacije: potpuno povezana višeslojna neuronska mreža, sequence-to-classification model zasnovan na rekurentnim neuronskim mrežama (RNN) i konvolutivna neuronska mreža (CNN) ukombinovana sa jednostavnom potpuno povezanom neuronskom mrežom.

B. *Metode za zadatak određivanja vremenskog intervala trajanja zvuka*

Pripremljeni podaci predstavljeni su u obliku sekvence reprezentacija zvučnog signala u različitim vremenskim trenucima. Zbog sposobnosti pamćenja istorijskih informacija, odnosno vremenske dimenzije iz podataka, za predstavljanje vremenskih sekvenci se često koriste rekurentne neuronske mreže (RNN). Usled postojanja problema eksplozije i nestajućeg gradijenta razvijen je LSTM (*long short-term memory*) model i njegova pojednostavljena verzija GRU (*gated recurrent unit*), koji se u ovom radu koristi tokom istraživanja.

GRU jedinice su iskorišćene u sklopu sequence-to-sequence modela čiji zadatak je da na osnovu sekvenci MFCC vrednosti zvučnog signala predvidi vremenske

intervale trajanja specifičnog zvuka, odnosno vektor nula i jedinica u kome jedinice predstavljaju pozicije trajanja zvuka u vremenu.

Drugi isprobani model za rešavanje zadatka iz ove grupe je konvolutivna neuronska mreža (CNN) ukombinovana sa jednostavnom potpuno povezanom neuronskom mrežom. Za razliku od sličnog modela iskorišćenog za zadatak detekcije zvuka, ovaj model je obučavan nad izmenjenim skupom podataka. Naime, zadatak modela postaje ekvivalentan zadatku prethodnog modela, ali model se obučava nad pojedinačnim uzorcima (reprezentacije) zvučnog signala umesto nad celokupnim audio snimkom. Na ovaj način, sam model ne prepoznaje kompletan interval trajanja konkretnog zvuka, već detektuje postojanje zvuka na dovoljno malim uzorcima sa snimka. Ovakav pristup rezultat je više eksperimenata sa obučavanjem modela nad reprezentacijama celokupnog audio snimka, gde je uočeno pogoršanje rezultata sa povećanjem vremenskog trajanja snimka.

Treći i poslednji isprobani model iz ove grupe je model čiji je cilj određivanje tačnog trenutka početka i trenutka završetka trajanja zvuka na snimku. Strukturu modela čini enkoder sloj na ulaznom delu i dva dekoder sloja na izlazu iz modela. Pretpostavka je da bi pojedinačni dekoderi mogli biti uspešno obučeni nad kodiranim MFCC vrednostima, a za zadatke otkrivanja početka i kraja trajanja zvuka. Specifičan zadatak otkrivanja konkretnog vremenskog trenutka nekog događaja, zdravorazumski, nalaže da je potrebno uočiti karakterističnu promenu iz trenutka kada se zvuk ne nalazi na snimku u trenutak kada se on može čuti, i obrnuto. Iz ovog razloga za enkoder deo modela izabrana je GRU jedinica, koja može pamtiti istorijske informacije, ali pokazaće se da priroda korišćenog modela ipak ne odgovara zadatku. Za izgradnju oba dekoder sloja modela isprobane su potpuno povezane neuronske mreže i GRU jedinice.

Svi modeli iz ove grupe isprobani su u režimu jednoklasne klasifikacije, odnosno kreirani su pojedinačni modeli za otkrivanje vremenskog intervala trajanja svakog zvuka, koji su zatim objedinjeni u ansambl modela.

IV. Rezultati i diskusija

Koristeći prethodno opisane pripremljene podatke i primenjujući navedene metode klasifikacije i metode detekcije sekvenci izvršeni su brojni eksperimenti u potrazi za najboljim rešenjem zadataka ovog rada.

U svakom od eksperimenata podaci su podeljeni na trening (64%), validacioni (16%) i test skup (20%), dok su za evaluaciju rezultata nad pojedinačnim klasama korišćene mere preciznosti, odziva i F1 mera. Na nivou svih klasa zajedno korišćene su iste mere sračunate *micro* i *macro average* metodama kojima se može zaključiti odnos rezultata modela za najzastupljenije i najmanje zastupljene klase.

A. *Eksperimenti sa zadatkom detekcije postojanja zvuka*

Generalna procedura izvršena u eksperimentima iz ove grupe obuhvata pripremu podatka, kao što je ranije navedeno, obučavanje različitih predloženih modela i evaluaciju njihovih rezultata.

Eksperimenti sa neuronskom mrežom su vršeni ručno i započeti su sa jednostavnom mrežom, koja je davala nezadovoljavajuće rezultate. Isprobavanjem različitih parametara modela uočene su neznatne promene rezultata, a najoptimalnija arhitektura sastoji se od: dva sakrivena sloja (1024 i 256 čvorova po sloju), *dropout*-a od 20 % između svakog sloja, „selu“ aktivacione funckije na skrivenim slojevima, „binary crossentropy” loss funkcije, „sigmoid” aktivacione funkcije na izlaznom sloju i „adam” optimizatora.

Obučavanjem sequence-to-classification modela dobijeni su lošiji rezultati u odnosu na prethodni eksperiment, a kao najbolji pokazao se vrlo jednostavan model sa dva GRU sloja koji kodiraju ulazne podatke u vektor dužine 512, jednim sakrivenim potpuno povezanim slojem sa 512 čvorova i jednim izlaznim klasifikacionim slojem koji daje predikciju za neke od 10 klasa. Ostali parametri jednaki su modelu iz prethodnog eksperimenta.

Preostali eksperimenti iz ove grupe vršeni su sa modelom koji kombinuje konvolutivne (CNN) slojeve sa potpuno povezanom neuronskom mrežom. Najbolji rezultati, uzimajući u obzir performativnost modela, dobijeni su konfiguracijom koja se sastoji od: jednog ulaznog i četiri sakrivena konvolutivna sloja (128, 128, 64, 64, 64 čvora po sloju, respektivno) sa „relu“ aktivacionom funcijom, max pooling i batch normalizacijom između svakog konvolutivnog sloja, jednog sakrivenog potpuno povezanog sloja sa 256 čvorova i relu aktivacijom i jednog izlaznog sloja za određivanje jedne od 10 klasa. Preostali parametri jednaki su parametrima modela iz prethodnih eksperimenata.

Analizom rezultata iz tabele 1 može se videti superiornost rada poslednje opisanog modela u odnosu na ostale isprobane, što se može objasniti činjenicom da konvolutivni slojevi prave vrlo dobre i karakteristične unutrašnje reprezentacije podataka što omogućava lakšu distinkciju između zvukova. Takođe, važno je napomenuti da se sekvence MFCC vrednosti mogu posmatrati kao fotografija sa jednim kanalom boje, a poznato je da se konvolutivni slojevi često i uspešno koriste u sistemima za prepoznavanje i klasifikaciju slika.

Korišćeni model	Mera	Micro avg			Macro avg		
		Precission	Recall	F1	Precission	Recall	F1
Neuronska mreža		0.57	0.48	0.52	0.50	0.40	0.43
Seq-to-class (GRU + NN)		0.46	0.39	0.42	0.33	0.28	0.30
CNN + NN		**0.79**	**0.71**	**0.75**	**0.81**	**0.64**	**0.70**

Tabela 1 Prikaz rezultata izvršenih eksperimenata za zadatak detekcije postojanja zvuka na snimku.

B. *Eksperimenti sa zadatkom određivanja vremenskog intervala trajanja zvuka*

Zadatak koji je rešavan pri vršenju ovih eksperimenata se razlikuje u odnosu na prethodno opisani u tome što je ovde bilo potrebno pronaći periode u audio snimcima u kojima je određeni zvuk prisutan. Kao što je ranije predstavljeno, za rešavanje ovog problema korišćeni su naizgled drugačiji pristupi ali konačni rezultat svih metoda

jeste vreme početka trajanja određenog zvuka i njegov kraj, odnosno više takvih parova ukoliko se taj zvuk više puta pojavio u audio snimku.

Svi eksperimenti su vršeni ručno a evaluirani su bibliotekom sed_eval [4]. Da bismo dobili rezultate modela ovoj biblioteci potrebno je proslediti podatke o nazivu snimka, nazivu zvučnog događaja, početak trajanja događaja u sekundama i njegov kraj, za dobijene podatke uz pomoć korišćenja modela i tačne podatke uz pomoć kojih se rad tih modela evaluira. Ova biblioteka nam daje dve skupine metrika. Prva se naziva *Segment based* i daje nam odgovor na pitanje da li je u određenom segmentu (jedna sekunda) tačno pretpostavljeno postojanje zvuka određenog događaja, dok druga (*Event based*) posmatra trenutke početka i kraja i govori nam koliko se dobro poklapaju tačni i pretpostavljeni podaci.

U prvoj podgrupi ovih eksperimenata pristupilo se rešavanju *sequence-to-sequence* problema i to u ansambl režimu, odnosno kreiranjem posebnog modela za svaki zvuk pojedinačno (za otkrivanje vremenskog intervala njegovog trajanja). Ulaz u model predstavlja sekvenca MFCC vrednosti, a izlaz je sekvenca nula i jedinica, gde jedinice ukazuju na prisutnost određenog zvuka u tom vremenskom trenutku, a nule obrnuto, da zvuk nije prisutan. Rešavanju ovog problema pristupilo se korišćenjem dva sloja rekurentne mreže, odnosno njene varijante zvane GRU. Ulaz u prvi sloj predstavlja sekvenca MFCC vrednosti, izlaz iz tog sloja se prosleđuje na drugi sloj iz koga se, kao izlaz, očekuje sekvenca koja nam govori o postojanju određenog zvuka u tom audio snimku.

Tokom eksperimenata isprobane su brojne varijacije modela u cilju pronalaska najbolje arhitekture. Izvršeno je dodavanje dodatnih slojeva, povećavan je i smanjivan broj GRU ćelija i menjani su ostali hiperparametri. Isprobano je i korišćenje LSTM modela na mestu GRU slojeva što se pokazalo nepogodno sa stanovišta rezultata i vremena obučavanja. Ovi postupci nisu doveli do drastičnih povećavanja performansi te je posmatrano kako se model ponaša ukoliko se za obučavanje koriste kraće i duže sekvence. U tabeli 2 su prikazani rezultati iste arhitekture modela nad različitim dužinama sekvenci.

Dužina sekvence	Mera	Segment based			Event based		
		Precission	Recall	F1	Precission	Recall	F1
10s		78.69	88.62	83.36	46.42	69.31	55.60
5s		**91.09**	**94.63**	**92.70**	**72.51**	**88.21**	**79.59**
1s		91.91	91.28	91.60	68.55	87.12	76.73
0.2s		87.77	96.65	92.00	63.27	91.30	74.74

Tabela 2 Prikaz rezultata *sequence-to-sequence* modela treniranog nad sekvencama različite dužine trajanja.

Očigledno je da model generalno daje značajno bolje rezultate za kraće sekvence, a pri korišćenju značajno kraćih sekvenci (kraćih od sekund) primetno je povećanje odziva modela uz pogoršanje dobijene preciznosti.

Pretpostavka je da bi najbolji model iz prethodne grupe, u realnom sistemu, služio kao filter te određivao koji se zvuk nalazi na snimku. Takvi snimci bi dalje bili prosleđeni odgovarajućem modelu za određivanje tačnih intervala pojavljivanja odgovarajućeg zvuka. Poznavajući rezultate „filter“ modela, zaključeno je da bi velika većina prosleđenih snimaka zaista sadržala konkretan zvuk, te su za obučavanje krajnjeg modela korišćeni prvenstveno podaci u kojima se zasigurno nalazi ciljni zvuk.

Radi poređenja, najbolji dobijeni model obučen je nad kompletnim skupom podataka, ali sa istim ciljem određivanja vremenskog intervala trajanja samo jednog tipa zvuka. Dobijeni rezultati su lošiji, ali ne značajno, što se može pripisati nebalansiranosti podataka u kojima se zvuk nalazi u odnosu na ostale (pr. „Blender“ se nalazi na 436 snimaka, a na 2942 ne). Iz razloga kompleksnosti nije izvršeno odgovarajuće balansiranje, kao što je oversampling ili undersampling, ali je isprobana primena stratifikacije podataka prilikom podele na obučavajuće i test podatke, kao i korišćenje callback funkcije za određivanje najboljeg modela tokom obučavanja, a na osnovu vrednosti mere preciznosti nad validacionim skupom podataka. Rezultati nakon primenjenih modifikacija ostali su približno jednaki, što bi moglo da znači da je prvobitna podela podataka rezultovala srazmernim brojem snimaka sa zvukom u obučavajućem i test skupu. Takođe se može zaključiti da nije došlo do značajnog overfitovanja modela nad obučavajućim podacima. Iz rezultata prikazanih u tabeli 3 može se primetiti da je odziv taj koji čini rezultate lošijim što je i za očekivati jer postoji više sekvenci nad kojima se obučava a dobar deo njih ne sadrži traženi zvuk.

Ansambl model	Mera	Segment based			Event based		
		Precission	Recall	F1	Precission	Recall	F1
Obučavan nad sekvencama gde se zvuk sigurno nalazi.		**91.09**	**94.63**	**92.70**	**72.51**	**88.21**	**79.59**
Obučavan nad svim sekvencama iz podataka.		94.44	84.01	88.92	78.38	76.06	77.20

Tabela 3 Poređenje rezultata modela treniranog nad svim sekvencama audio snimaka i modela treniranog nad sekvencama koji sigurno sadrže traženi zvuk

Druga podgrupa eksperimenata inspirisana je dobrim rezultatima konvolutivnih neuronskih mreža iz prve grupe eksperimenata gde se pokazalo da dobro uočavaju osobine zvukova i na taj način daju dobru pretpostavku da se zvuk pojavljuje u audio snimku. Za razliku od tih eksperimenata, ovde je rešavan problem pronalaska tačnog intervala trajanja zvuka. Kako konvolutivne neuronske mreže nisu namenjene za učenje iz sekvenci i uzimanje temporalnih zavisnosti u obzir, razmatrane su isključivo kratke sekvence a samim tim i jednostavniji model, u odnosu na onaj iz prethodne grupe eksperimenata, sa dva konvolutivna sloja. U tabeli 4 mogu se videti rezultati ovog pristupa i iz njih ponovo možemo primetiti dobru detekciju postojanja zvuka u određenom segmentu (*Segment based* metrika) ali jasno se vide lošije performanse u određivanju tačnog perioda trajanja zvuka u odnosu na pristup koji koristi rekurentnu neuronsku mrežu.

Dužina sekvence	Mera	Segment based			Event based		
		Precission	Recall	F1	Precission	Recall	F1
5s		**87.28**	**82.09**	**84.60**	50.35	59.53	54.56
1s		84.67	83.83	84.25	**52.33**	**72.16**	**60.67**
0.2s		75.90	93.88	83.94	38.69	78.93	51.93

Tabela 4 Prikaz rezultata korišćenja konvolutivne neuronske mreže obučavane nad sekvencama različite (male) dužine trajanja.

Rešavanje problema određivanja vremenskog intervala trajanja zvuka na audio snimku se može rešiti i pronalaskom

tačnog trenutka njegovog pojavljivanja i trenutka prestanka. Upravo na taj način su vršeni eksperimenti iz treće podgrupe. Arhitektura modela koji rešava ovaj problem se sastoji od jednog sloja rekurentne neuronske mreže čiji se izlazi vezuju na dve potpuno povezane neuronske mreže koje predviđaju sekvence u kojima se nalaze indikatori početka, odnosno završetka trajanja zvuka. Pored potpuno povezanih neuronskih mreža na mestu dekodera isprobani su i slojevi sa GRU ćelijama. Ipak ovaj pristup se pokazao kao inferioran u poređenju sa prethodna dva te nije izvršeno kreiranje ansambla, već je model testiran na samo jednom tipu zvuka (događaju). Pokazalo se da je dekoder zadužen za pronalazak početka zvučnog događaja davao značajno bolje rezultate. Nakon evaluacije nad test skupom i dobijanja loših rezultata, izvršena je evaluacija nad trening skupom kako bi se proverilo da li je ovaj pristup pogrešan. Performanse modela u tom slučaju su bile bolje, ali uočeno je značajno overfitovanje modela koje očigledno utiče na performanse pri evaluaciji nad test skupom. Iz ovog razloga isprobani su dodatni pristupi obučavanju modela kao što je stratifikacija podataka za obučavanje i testiranje i obučavanje modela sa callback funkcijom koja prati preciznost modela nad validacionim skupom i pamti (kreira model checkpoint) najbolji dobijeni model, a ne krajnji model koji može biti overfitovan. Nakon ovih modifikacija, iako bolji, rezultati se nisu pokazali kao zadovoljavajući, naročito u poređenju sa ostalim isprobanim pristupima. U prvom redu tabele 5 prikazani su najbolji rezultati prepoznavanja zvuka blendera iz prethodnih eksperimenata, a u redovima ispod, performanse ovog pristupa.

Mera / Pristup	Segment based			Event based		
	Precission	Recall	F1	Precission	Recall	F1
Seq-to-seq	96.92	98.91	97.58	76.03	91.09	82.88
Seq-to-x2_seq 10s	64.34	25.07	36.08	21.82	11.88	15.38
Seq-to-x2_seq 0.2s	66.96	40.87	50.87	18.18	17.82	18.00

Tabela 5 Poređenje rezultata najboljeg modela i modela iz treće podgrupe, obučavanih isključivo nad snimcima sa zvukom blendera.

Kao varijacija ovog pristupa, na osnovu MFCC vrednosti, kreirane su sekvence koje su obuhvatale promene MFCC vrednosti iz jednog vremenskog koraka u naredni korak. Postojala je pretpostavka da bi se konkretan model lakše obučio za prepoznavanje razlika vrednosti osobina zvuka u odnosu na njihove konkretne vrednosti. Na žalost taj pristup nije dao značajna poboljšanja, a zaključeno je da bi ovaj pristup mogao biti primenjen samo na jedan konkretan unapred poznat zvuk (pošto se koriste samo promene MFCC vrednosti koje su često vrlo slične kada su u pitanju različiti zvukovi). Takođe, postojali bi i veliki problemi ukoliko bi snimak sadržao značajnu količinu šuma.

DISKUSIJA

Nakon izvršenja opisanih eksperimenata uočene su znatne razlike u dobijenim rezultatima u zavisnosti od primenjenog pristupa i u zavisnosti od rešavanog zadatka.

Za zadatak detekcije postojanja zvuka na snimku, kao najbolji, izdvaja se pristup u kome se koristi konvolutivna neuronska mreža ukombinovana sa potpuno povezanom neuronskom mrežom. U kontrastu, za zadatak određivanja vremenskog intervala trajanja zvuka pokazalo se da je najbolje primeniti model čija je jedna od namena rad sa vremenskim sekvencama. Pokazalo se da, iako slični, zadaci rešavani u ovom radu ne mogu biti savladani na jedinstven, ili pak sličan način. Iz vrednosti prikazanih u tabeli 1 primetni su drastično lošiji rezultati modela koji koristi rekurentne neuronske mreže, a koji za rešavanje drugog zadatka daje najbolje rezultate. Ovo se može objasniti razlikom u prirodi rešavanih problema i to time što problem detekcije postojanja zvuka ne zavisi od tačnog vremenskog trenutka pojavljivanja zvuka na snimku, te se stoga ne može ni opravdati korišćenje pristupa zasnovanog na rekurentnim neuronskim mrežama (odnosno upotrebi GRU jedinica).

Ukoliko se uporede rezultati eksperimenata i pristupa iz druge grupe, odnosno onih za određivanje vremenskog intervala trajanja zvuka uviđa se superiornost sequence-to-sequence modela u odnosu na ostale. Ovo znači da su vremenske sekvence, odnosno istorijski događaji vrlo bitni za konkretan problem, te modeli koji koriste konvolutivne neuronske mreže, a koji nemaju mehanizam za čuvanje informacija o više sukcesivnih vremenskih događaja, daju dobre, ali lošije rezultate. Ukoliko posmatramo pristup u kome se na osnovu podataka kodiranih pomoću GRU jedinica dobijaju sekvence sa indikatorima početka i kraja trajanja zvuka, zaključuje se da primenjeni model ne odgovara prirodi problema. Naime, arhitektura ovakvog modela se generalno može opisati kao sequence-to-sequence, što podrazumeva postojanje kompletne sekvence za obučavanje, dok se u isprobanom pristupu od modela očekuje prepoznavanje posebno početka, a posebno kraja sekvence (dela zvuka u kom počinje i u kom se završava traženi zvuk). Priroda i namena modela su prepoznavanje i prediktovanje sekvence, te je očekivano da ne daje odlične rezultate za drugačiji zadatak. Ovu pretpostavku podržava i činjenica da model obučavan za prepoznavanje čitave sekvence zvuka (vremenskog intervala postojanja zvuka), a koji ima jedan dekoder sloj, daje vrlo dobre rezultate.

Značajno je napomenuti da su modeli iz druge grupe obučavani u režimu ansambla, odnosno kreiran je poseban model za svaku klasu (mogući zvuk) pojedinačno. S tim u vezi korisno je uporediti rezultate dobijene primenom najboljeg pristupa za različite događaje tj. klase. U tabeli 6 prikazani su rezultati dobijeni nakon obučavanja 10 najboljih modela za 10 mogućih zvukova.

Mera / Klasa	Segment based				Event based			
	Preci-ssion	Recall	F1	Cnt	Preci-ssion	Recall	F1	Cnt
Alarm	95.5	90.7	93.1	304	81.7	90.2	85.7	153
Blender	96.3	98.9	97.6	367	76.0	91.1	82.9	101
Mačka	92.1	94.4	93.3	198	78.2	92.1	84.5	101
Posuđe	80.6	80.9	80.7	241	68.6	78.2	73.1	165
Pas	87.9	93.3	87.9	179	67.2	91.1	77.3	101
Električni brijač	98.1	99.6	98.1	234	**81.8**	**97.8**	**89.1**	46
Prženje	97.8	98.1	97.8	160	67.6	82.1	74.2	28
Voda iz slavine	92.9	96.6	92.9	148	51.0	78.8	61.9	33
Govor	90.7	94.0	90.7	861	71.3	89.3	79.3	428
Usisivač	**98.5**	**98.7**	**98.5**	233	77.8	87.5	82.4	40

Tabela 6 Poređenje rezultata najboljeg modela obučenog nad pojedinačnim klasama.

Uočljivo je da modeli za klase čiji je broj primeraka veći nemaju bolje rezultate u odnosu na one čiji podaci su manje zastupljeni, što pomalo iznenađuje. Ovakvo ponašanje može se objasniti činjenicom da su zvukovi kao što su blender, alarm i električni brijač, iako manje zastupljeni, vrlo karakteristični i konzistentni tokom vremena, te je potrebno manje primeraka za uspešno obučavanje modela. Ljudski govor ili posuđe, sa druge strane, (njihova amplituda i frekvencija) mogu se drastično razlikovati kako od snimka do snimka tako i tokom samog trajanja zvuka.

V. ZAKLJUČAK

U ovom radu predložena su rešenja zadataka detekcije postojanja zvuka i predikcije vremenskog intervala trajanja zvuka na snimku. Korišćeni su sintetički podaci generisani *Scaper* alatom, koji su zatim obrađeni i prevedeni u niz MFCC vrednosti kao pogodnih računarskih reprezentacija. Generisane reprezentacije podataka korišćene su za obučavanje više različitih modela mašinskog učenja za oba rešavana zadatka.

Sistem koji uspešno prepoznaje, a zatim izdvaja tačno vreme trajanja željenog zvuka iz audio snimka bi značajno unapredio tehnologije za opremanje pametnih kuća, sigurnosnih sistema, sistema za prepoznavanje govora, sistema za forenzičku analizu, kao i u raznih IoT (*internet of things*) uređaja.

Rezultati metodologije kojom se došlo do najboljih rezultata, ali i ostalih, dokazuju upotrebnu vrednost primenjenih metoda mašinskog učenja za konkretne zadatke. Potvrđena je upotrebljivost MFCC vrednosti kao pogodne reprezentacije zvučnih signala za mašinsku obradu.

Isprobani modeli, pak, nisu upotrebljeni nad podacima iz realnog okruženja, te se dobijeni rezultati moraju uzeti sa dozom rezerve, ali mogu predstavljati dobru osnovu za dalji razvoj, eksperimente i istraživanja u radu sa realnim podacima i kao izvor pokazatelja uspešnosti prikazanih metodologija. S tim u vezi mogu se dati predlozi za unapređenja.

Prvi korak u budućem istraživanju bio bi prikupljanje stvarnih podataka i to u mnogo većim količinama kako bi se iskazala verodostojnost rezultata. U slučaju realnih podataka, predlaže se detaljnija analiza i predprocesiranje podataka radi uklanjanja eventualnog šuma, obradom samog zvučnog signala ili modifikacijama procesa dobijanja MFCC vrednosti.

Ukoliko se obučavanje modela vrši nad svim podacima, predlaže se upotreba mehanizama za njihovo balansiranje, a u cilju približnog izjednačavanja broja primeraka svih postojećih zvukova, što bi potencijalno dovelo do boljih rezultata.

Korisno bi bilo, takođe, izgraditi i evaluirati jedan hibridni model koji bi se sastojao od ulaznog "filter" sloja za određivanje postojanja nekog zvuka na snimku i od sloja za određivanje tačnog vremenskog intervala trajanja pronađenog zvuka. Za prvi sloj predlaže se upotreba modela zasnovanog na konvolutivnim neuronskim mrežama sa potpuno povezanim slojem, dok bi se za drugi sloj, na osnovu dobijenih rezultata u ovom radu najbolje pokazao ansambl sequence-to-sequence modela. Za uspešnu evaluaciju ovakvog pristupa potreban je dovoljno veliki skup podataka iz kog bi se moglo odvojiti dovoljno snimaka koji nisu korišćeni za obučavanje ni jednog modela u sistemu, što tokom ovog istraživanja nije bilo moguće u dovoljnoj meri za donošenje validnih zaključaka.

Kako se pokazalo da dužina sekvence audio snimka značajno utiče na performanse svih razmatranih modela, potrebno je odrediti optimalnu dužinu nad kojom bi model radio pri čemu bi željena upotreba ovakvog sistema imala najviše uticaja. Pritom, bilo bi potrebno evaluirati ponašanje modela za detekciju zvuka nad kraćim sekvencama.

Naposletku, iako u ovom radu neuspešan, pristup u kome se od modela očekuje prepoznavanje početka i kraja trajanja zvuka mogao bi biti modifikovan na taj način da se od modela ne očekuje prepoznavanje sekvence u kojoj se nalaze indikatori početka ili kraja, već predikcija indikatora za svaki uzorak iz ulazne sekvence posebno.

LITERATURA

[1] IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events, http://dcase.community/challenge2020/

[2] Salamon, Justin & Macconnell, Duncan & Cartwright, Mark & Li, Peter & Bello, Juan. (2017). Scaper: A Library for Soundscape Synthesis and Augmentation. 10.1109/WASPAA.2017.8170052.

[3] S. B. Davis and P. Mermelstein: "Comparison of para-metric representations for monosyllabicwordrecognitionin continuously spoken sentences", IEEE Transactionson Acoustic, Speech, and Signal Processing, Vol. 28,No. 4, August 1980, pp. 357–366.

[4] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "Metrics for polyphonic sound event detection", Applied Sciences, 6(6):162, 2016