



**UNIVERZITET U NOVOM SADU  
FAKULTET TEHNIČKIH NAUKA U  
NOVOM SADU**

---




Dragomir Božoki

# **Automatska transkripcija govora na osnovu video-snimaka korišćenjem neuronskih mreža**

**DIPLOMSKI RAD**  
-osnovne akademske studije-

Septembar 2024.

Novi Sad

	УНИВЕРЗИТЕТ У НОВОМ САДУ • ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА 21000 НОВИ САД, Трг Доситеја Обрадовића 6	Број:
	ЗАДАТАК ЗА ДИПЛОМСКИ (Bachelor) РАД	Датум:

Врсте студија:	Основне академске студије
Студијски програм:	Биомедицинско инжењерство
Руководилац студијског програма:	проф. др Никола Јорговановић

Студент:	Драгомир Божоки	Број индекса:	БИ 55/2020
Област:	Машинско учење I		
Ментор:	проф. др Милан Сечујски		
<p>НА ОСНОВУ ПОДНЕТЕ ПРИЈАВЕ, ПРИЛОЖЕНЕ ДОКУМЕНТАЦИЈЕ И ОДРЕДБИ СТАТУТА ФАКУЛТЕТА ИЗДАЈЕ СЕ ЗАДАТАК ЗА ДИПЛОМСКИ (Bachelor) РАД, СА СЛЕДЕЋИМ ЕЛЕМЕНТИМА:</p> <ul style="list-style-type: none"> <li>- проблем – тема рада;</li> <li>- начин решавања проблема и начин практичне провере резултата рада, ако је таква провера неопходна;</li> <li>литература</li> </ul>			

### НАСЛОВ ДИПЛОМСКОГ (Bachelor) РАДА:

Аутоматска транскрипција говора на основу видео-снимака коришћењем неуралних мрежа

### ТЕКСТ ЗАДАТКА:

<ol style="list-style-type: none"> <li>Дати опис постојећих параметарских метода за аутоматску транскрипцију говора на основу видео снимка без аудио компоненте (читање с усана) и дати упоредни преглед њихових резултата</li> <li>Формирати модел на основу неуралне мреже који ће бити обучен на постојећим видео снимцима говорника са познатим фонетским транскрипцијама (јавно доступни ГРИД корпус);</li> <li>Испитати перформансе добијеног система за читање с усана и извести одговарајуће закључке.</li> </ol>
---

Руководилац студијског програма:	Ментор рада:
проф. др Никола Јорговановић	проф. др Милан Сечујски

Примерак за: <input type="checkbox"/> - Студента; <input type="checkbox"/> - Ментора; <input type="checkbox"/> - Студентску службу факултета
--



УНИВЕРЗИТЕТ У НОВОМ САДУ • ФАКУЛТЕТ ТЕХНИЧКИХ НАУКА  
21000 НОВИ САД, Трг Доситеја Обрадовића 6

## КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА

Редни број, <b>РБР:</b>	
Идентификациони број, <b>ИБР:</b>	
Тип документације, <b>ТД:</b>	Монографска документација
Тип записа, <b>ТЗ:</b>	Текстуални штампани материјал
Врста рада, <b>ВР:</b>	Дипломски рад
Аутор, <b>АУ:</b>	Драгомир Божоки
Ментор, <b>МН:</b>	проф. др Милан Сечујски
Наслов рада, <b>НР:</b>	Аутоматска транскрипција говора на основу видео-снимака коришћењем неуралних мрежа
Језик публикације, <b>ЈП:</b>	Српски
Језик извода, <b>ЈИ:</b>	Српски
Земља публикавања, <b>ЗП:</b>	Република Србија
Уже географско подручје, <b>УГП:</b>	Војводина
Година, <b>ГО:</b>	2024.
Издавач, <b>ИЗ:</b>	Ауторски репринт
Место и адреса, <b>МА:</b>	Нови Сад; трг Доситеја Обрадовића 6
Физички опис рада, <b>ФО:</b> (поглавља/страна/ цитата/табела/слика/графика/прилога)	7 поглавља / 34 стране / 37 цитата / 2 табеле / 0 графика / 12 слика / 0 прилога
Научна област, <b>НО:</b>	БИОМЕДИЦИНСКО ИНЖЕЊЕРСТВО
Научна дисциплина, <b>НД:</b>	МАШИНСКО УЧЕЊЕ I
Предметна одредница/Кључне речи, <b>ПО:</b>	ДИГИТАЛНА ОБРАДА ВИДЕО ЗАПИСА, ТРАНСКРИПЦИЈА ГОВОРА НА ОСНОВУ ВИДЕО ЗАПИСА, НЕУРАЛНЕ МРЕЖЕ
<b>УДК</b>	
Чува се, <b>ЧУ:</b>	У библиотеци Факултета техничких наука, Нови Сад
Важна напомена, <b>ВН:</b>	
Извод, <b>ИЗ:</b>	У овом раду приказане су постојеће параметарске методе за аутоматску транскрипцију говора на основу видео снимка без аудио компоненте, познате као читање с усана. На основу анализе доступних метода, формиран је модел заснован на неуралној мрежи, обучен на видео снимцима из јавно доступног ГРИД корпуса. Резултати тестирања показују да модел постиже задовољавајуће резултате на тест скупу истог говорника, али се примећују одређене потешкоће у његовој генерализацији на непознате говорнике, што се приписује недовољном броју епоха током обуке и малом тренинг скупу. У даљем раду планирано је унапређење модела како би се побољшала његова способност генерализације.
Датум прихватања теме, <b>ДП:</b>	20.7.2024.
Датум одбране, <b>ДО:</b>	19.9.2024.
Чланови комисије, <b>КО:</b>	Председник: др Сениша Сузић
	Члан: асис. др Тијана Носек
	Члан, ментор: проф. др Милан Сечујски
	Потпис ментора



## KEY WORDS DOCUMENTATION

Accession number, <b>ANO</b> :		
Identification number, <b>INO</b> :		
Document type, <b>DT</b> :	Monographic publication	
Type of record, <b>TR</b> :	Textual printed material	
Contents code, <b>CC</b> :	Bachelor's thesis	
Author, <b>AU</b> :	Dragomir Božoki	
Mentor, <b>MN</b> :	Milan Sečujski, PhD	
Title, <b>TI</b> :	Automatic speech transcription based on video-recordings using neural networks	
Language of text, <b>LT</b> :	Serbian	
Language of abstract, <b>LA</b> :	Serbian	
Country of publication, <b>CP</b> :	Republic of Serbia	
Locality of publication, <b>LP</b> :	Vojvodina	
Publication year, <b>PY</b> :	2024.	
Publisher, <b>PB</b> :	Author's reprint	
Publication place, <b>PP</b> :	Novi Sad, Dositeja Obradovica sq. 6	
Physical description, <b>PD</b> : (chapters/pages/ref./tables/pictures/graphs/appendixes)	7 chapters / 34 pages / 37 citations / 2 tables / 0 graphs / 12 images / 0 appendices	
Scientific field, <b>SF</b> :	BIOMEDICAL ENGINEERING	
Scientific discipline, <b>SD</b> :	MACHINE LEARNING I	
Subject/Key words, <b>S/KW</b> :	DIGITAL VIDEO PROCESSING, SPEECH TRANSCRIPTION BASED ON VIDEO, NEURAL NETWORKS	
<b>UC</b>		
Holding data, <b>HD</b> :	The Library of the Faculty of Technical Sciences, Novi Sad, Serbia	
Note, <b>N</b> :		
Abstract, <b>AB</b> :	<p>This paper presents existing parametric methods for automatic speech transcription based on video recordings without the audio component, known as lip reading. Based on the analysis of available methods, a model based on a neural network was formed and trained on video recordings from the publicly available GRID corpus. The test results show that the model achieves satisfactory performance on the test set from the same speaker, but certain difficulties in generalization to unseen speakers are observed, which is attributed to an insufficient number of training epochs, as well as the small size of the test set. Further work is planned to improve the model's generalization capabilities.</p>	
Accepted by the Scientific Board on, <b>ASB</b> :	20.7.2024.	
Defended on, <b>DE</b> :	19.9.2024.	
Defended Board, <b>DB</b> :	President:	Siniša Suzić, PhD
	Member:	asist. Tijana Nosek, PhD
	Member, Mentor:	Milan Sečujski, PhD
		Menthor's sign



**UNIVERZITET U NOVOM SADU  
FAKULTET TEHNIČKIH NAUKA U  
NOVOM SADU**

---



**Automatska transkripcija govora na osnovu video-snimaka  
korišćenjem neuronskih mreža**

**-DIPLOMSKI RAD-**

Student:  
Dragomir Božoki

Mentor:  
prof. dr Milan Sečujski

Septembar 2024.  
Novi Sad

## **Zahvalnost**

*Ovaj diplomski rad posvećujem svojoj porodici – ocu Laslu, majci Kseniji, baki Jagodi i dedi Laslu. Takođe, izražavam duboku zahvalnost prijateljima, profesorima i asistentima koji su me pratili i motivisali kroz ceo proces školovanja.*

*- Dragomir*

APSTRAKT .....	4
ABSTRACT .....	4
1 Uvod .....	6
2 Metode i principi rada neuronskih mreža u VPG .....	8
2.1 Definicija veštačkog neurona .....	8
2.2 Biranje funkcije aktivacije za višeslojne neuronske mreže .....	10
2.3 Konvergencija u neuronskim mrežama .....	11
2.4 Konvolucione i prostorno-vremenske konvolucione neuralne mreže .....	12
2.5 Rekurentne neuronske mreže .....	13
2.6 Funkcija gubitka u VPG .....	17
3 Pregled postojeće literature .....	19
3.1 Automatsko VPG .....	20
3.2 Duboko učenje u VPG .....	20
4 LipReading - principi rada .....	24
4.1 Programsko okruženje i korišćene biblioteke .....	24
4.2 Baza podataka – GRID Korpus .....	24
4.3 Procesiranje podataka i parametri obuke modela .....	25
4.3 LipReadingWebCam – princip rada .....	27
5 Rezultati .....	28
5.1 Perfomanse LipReading modela .....	28
5.2 Perfomansa LipReadingWebCam algoritma .....	29
6 Unapređivanje i dalji razvoj .....	30
7 Zaključak .....	31
Reference .....	32
Biografija .....	34

## APSTRAKT

Čitanje s usana je proces dobijanja govornih informacija posmatranjem pokreta usana govornika u situacijama kada je zvuk odsutan ili nejasan. Iako nije dominantan modalitet poput audio kanala, vizuelna informacija o govoru igra značajnu ulogu u svakodnevnoj komunikaciji. Veština čitanja s usana je posebno korisna osobama sa oštećenim sluhom, jer im omogućava da prate razgovore i razumeju govor putem vizuelnih znakova. Vizuelno prepoznavanje govora (VPG) je multidisciplinarna metoda, koja kombinuje kompjutersku viziju i nauku o obradi govora, sa ciljem da interpretira izgovoreni tekst iz video podataka na osnovu pozicije usana ispitanika. U poslednje vreme automatsko vizuelno prepoznavanje govora se postepeno poboljšava, najviše zahvaljujući sve složenijim i moćnijim modelima i podacima za obuku.

Duboko čitanje sa usana (engl. *deep lip-reading*) je proces izdvajanja govornih informacija iz video snimaka na kojima se prikazuje lice koje govori, ali bez prisustva zvuka, uz pomoć dubokih neuronskih mreža. Ova metoda se u literaturi može sresti pod različitim nazivima, kao što su vizuelno prepoznavanje govora (engl. *Visual Speech Recognition* - VSR), mašinsko učenje za čitanje sa usana, automatsko čitanje sa usana, i dr. Cilj diplomskog rada je pregled već postojećih parametarskih metoda koje se bave ovom temom, ali i formiranje novog modela koji će biti obučen na već postojećoj bazi podataka, i testiran na podacima koji dolaze sa kamere u realnom vremenu. U većini metoda opisanih u ovom radu, kao i u praktičnom delu, koriste se neuralne mreže kao najefikasniji izbor parametarskih metoda. Iako su neuralne mreže algoritam koji je odavno poznat, tek je u skorije vreme sa razvojem softvera i hardvera, našao primenu u multidisciplinarnim problemima. Među značajnijim izazovima koji se javljaju u vizuelnom prepoznavanju govora su postojanje *homofona* – vizuelno slične gestikulacije usana koja odgovara različitim fonemima, uticaj položaja objekta i osvetljenja na kameri takođe utiče na efikasnost modela, kao i ručno označavanje labela (anotiranje) podataka. U ovom radu su opisani i testirani neki od metoda i algoritama za prevazilaženje pomenutih problema.

## ABSTRACT

Lip reading is the process of obtaining speech information by observing the speaker's lip movements in situations where sound is absent or unclear. Although it is not as dominant as the audio channel, visual speech information plays a significant role in everyday communication. The skill of lip reading is especially useful for individuals with hearing impairments, as it enables them to follow conversations and understand speech through visual cues. Visual Speech Recognition (VSR) is a multidisciplinary method that combines computer vision and speech processing science, with the goal of interpreting speech from video data based on the position of the subject's lips. Recently, automatic and visual speech recognition has been gradually improving, mainly due to the increasing size of models and training data.

Deep lip reading is the process of extracting speech information from videos where a speaking face is shown, but without the presence of sound, using deep neural networks. This method is referred to by various names in the literature, such as Visual Speech Recognition (VSR), Machine Learning for Lip Reading, Automatic Lip Reading, etc. The aim of this thesis is to describe existing parametric methods that address this topic, as well as to develop a new model that will be trained on an existing database and tested on data coming live from a camera in real-time. In most of the methods described in this paper, as well as in the practical part, neural networks are used as the most efficient choice of parametric methods. Although neural networks are an algorithm that has been known for a long time, only recently, with the development of software and hardware, have they found application in multidisciplinary



problems. One of the major challenges in visual speech recognition is the presence of homophones — visually similar lip gestures that represent different phonemes, the impact of object position and lighting on the camera on the model's efficiency, as well as manual data labeling. This paper also describes and tests some of the methods and algorithms for overcoming these challenges.

## 1 Uvod

Prema podacima Svetske zdravstvene organizacije (WHO), više od 1.5 milijardi ljudi, odnosno približno 20% svetske populacije, živi sa oštećenim sluhom. Od tog broja, čak 430 miliona osoba ima ozbiljno ili potpuno oštećenje sluha [1]. Gubitak sluha se pretežno pripisuje dvama osnovnim uzrocima: starenju i izloženosti prekomernoj buci. Uz ubrzani razvoj tehnologije u poslednjih nekoliko godina, gubitak sluha usled kontinuirane izloženosti buci postaje sve učestaliji problem [2].

Ljudske sposobnosti čitanja sa usana su ograničene i ono predstavlja izuzetno težak zadatak za ljude, pogotovo kada postoji odsustvo konteksta izgovorenih reči. Osobe sa oštećenim sluhom postižu tačnost od samo  $17 \pm 12\%$  čak i za ograničeni skup od 30 jednosložnih reči i  $21 \pm 11\%$  za 30 višesložnih reči [3], te se javlja potreba za algoritmima za automatsko čitanje sa usana. Pored osoba sa oštećenim sluhom, automatsko vizuelno prepoznavanje govora pronalazi svoju primenu i u interfejsima između ljudi i mašina, reparaciji oštećenih ili nekompletnih video zapisa, pri diskretnoj komunikaciji ili komunikaciji u ekstremnim uslovima gde je prisutan značajan nivo buke, kao i u mnogim drugim oblastima.

Vizuelno prepoznavanje govora (VPG) predstavlja kompleksan multidisciplinarni problem, koji je u velikoj meri jezički specifičan i zahteva prilagođavanje za svaki jezik. Rešavanje ovog problema obuhvata znanja iz teorije verovatnoće, lingvistike, prepoznavanja oblika i drugih naučnih oblasti.

VPG je oblast koja se intenzivno istražuje već decenijama. Mnoge osnovne tehnologije razvijene su još u prethodnom milenijumu. Ove tehnike su značajno unapredile stanje u oblasti VPG i srodnim poljima. U poređenju sa tim ranijim dostignućima, napredak u istraživanju i primeni VPG u deceniji pre 2010. bio je relativno spor i manje uzbudljiv, iako su mnoge važne tehnike i paradigme usvojene još tada.

U poslednjih nekoliko godina primećen je novi porast interesovanja za automatsko prepoznavanje govora. Ovu promenu je pokrenula povećana potreba za takvim sistemima u mobilnim uređajima i uspeh novih aplikacija za prepoznavanje govora u mobilnom svetu, kao što su pretraga glasa (eng. *Voice Search* – VS), diktiranje kratkih poruka (SMD) i virtuelni govorni asistenti (npr. *Apple*-ova Siri, *Google Now* i *Microsoft*-ova Cortana). Jednako važan je i razvoj tehnika dubokog učenja u prepoznavanju govora uz pomoć obimnijih baza podataka i značajno povećanih računarskih kapaciteta. Kombinacija tehnika dubokog učenja dovela je do smanjenja stope grešaka za više od 1/3 u odnosu na konvencionalne modele u mnogim praktičnim zadacima i pomogla da se pređe prag usvajanja za mnoge korisnike. Na primer, tačnost prepoznavanja reči na engleskom jeziku ili tačnost prepoznavanja karaktera na kineskom u većini sistema sada prelaze 90%, a u nekim sistemima čak i 95%.

Ključni pokazatelj uspešnosti metode vizuelnog prepoznavanja govora jeste sposobnost identifikacije *homofona* – pojave izražene dvosmislenosti pri analizi pozicije lica i usana tokom izgovora određenih fonema. Različite metode su istražene kako bi se ovaj problem adresirao, od kojih su neke bile više uspešne u rešavanju ovog problema dok su druge zaostajale da daju приметne rezultate. Prema Fišeru [4], postoji pet kategorija vizuelnih fonema, poznatih kao *vizemi*, koji su često pogrešno klasifikovani od strane ljudi kada posmatraju usta govornika. Upravo na ovom mestu automatsko prepoznavanje govora sa usana treba da pokaže svoju superiornost u odnosu na čoveka, budući da algoritmi mašinskog i dubokog učenja mogu mnogo preciznije da identifikuju i pravilno klasifikuju različite foneme koji odgovaraju istom vizemu.

Međutim, vizuelno prepoznavanje govora ima i svoje prednosti u odnosu na audio klasifikaciju podataka. Na primer, izgovori glasova /m/ i /n/ vizuelno su lako razlučivi, dok se pri oslanjanju

isključivo na audio signale često dešavaju greške u klasifikaciji. Ova vizuelna informacija može biti posebno korisna u situacijama gde audio signali nisu dovoljno jasni ili su podložni šumu.

Čitanje sa usana u realnom vremenu sa video podataka koji dolaze direktno sa kamere suočava se sa dodatnim izazovima, kao što su varijabilnost u položaju subjekta, razdaljina između subjekta i kamere, uslovi osvetljenja, kao i potreba za visoko optimizovanim algoritmom [5]. Svaki dodatni korak može značajno usporiti rad sistema, što je posebno izraženo u kontekstu obrade u realnom vremenu. U ovom radu su opisani i testirani neki od načina za prevazilaženje pomenutih problema, uz date sugestije za dalja unapređenja i optimizaciju sistema.

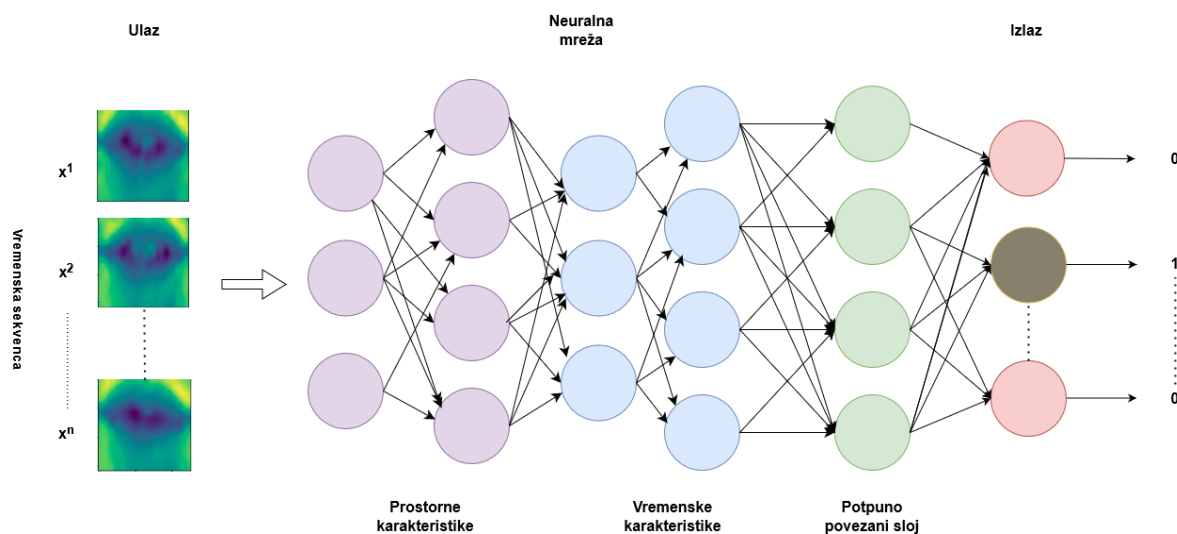
Rad je organizovan na sledeći način. Nakon uvoda, drugo poglavlje se bavi metodama i principima rada neuronskih mreža u vizuelnom prepoznavanju govora. Treće poglavlje govori o pregledu postojeće literature na ovu temu. Četvrto poglavlje detaljno opisuje algoritam korišćen u ovom radu. Peto poglavlje prikazuje dobijene rezultate, dok se u šestom poglavlju predlažu pravci za buduća istraživanja, a u sedmom se iznose zaključci.

## 2 Metode i principi rada neuronskih mreža u VPG

Vizuelno prepoznavanje govora predstavlja metodu nadgledanog učenja, gde se računaru kao ulaz dostavlja video snimak, a od njega se očekuje da kao izlaz klasifikuje svaki frejm video sekvence u odgovarajući fonem. Nizovi fonema se potom spajaju u smislenu rečenicu koja treba da odgovara stvarnoj govornoj sekvenci prikazanoj u tom video snimku. U savremenim pristupima, većina metoda oslanja se na primenu veštačkih neuronskih mreža, tačnije na metode dubokog učenja. Algoritam za vizuelno prepoznavanje govora sastoji se od više slojeva neuronskih mreža, pri čemu svaki sloj ima specifičan zadatak u obradi različitih aspekata informacija koje prenosi ulazna promenljiva.

Primarna metodologija za rešavanje problema uključuje dve faze: (1) izdvajanje vizuelnih i vremenskih karakteristika iz sekvenci frejmova u video zapisima bez zvuka, i (2) procesiranje tih sekvenci karakteristika u jedinice teksta, kao što su karakteri, reči, fraze itd. [5]

Da bi se video snimak uspešno analizirao i frejmovi klasifikovali u odgovarajuće foneme, neophodna je neuronska mreža koja se sastoji iz dva ključna dela. Prvi deo mreže obrađuje prostorne karakteristike, identifikujući koje delove slike treba smatrati relevantnim za kategorizaciju i određivanje odgovarajućeg fonema, a koji delovi nisu od značaja. Drugi deo mreže fokusiran je na vremenske karakteristike. S obzirom na to da je u analizi govora pozicija usana u prethodnim i trenutnim frejmovima ključna za tačnu predikciju, potreban je poseban tip neuronskih mreža koje su sposobne da prate vremenski promenljive atribute. Uopštena arhitektura neuralne mreže je prikazana na slici 2.1:



Slika 2.1. Arhitektura predložene neuralne mreže u VPG

### 2.1 Definicija veštačkog neurona

Veštački neuron je osnovni gradivni blok neuronske mreže inspirisan biološkim neuronima. Njegova funkcija je da prima jedan ili više ulaznih signala, izvrši obradu tih signala i generiše izlaz. Klasičan tip veštačkog neurona – perceptron [6], možemo da postavimo u kontekst zadatka binarne klasifikacije sa dve klase: 0 i 1. Zatim se može definisati funkcija odlučivanja (engl. *decision function*),  $\sigma(z)$ , koja koristi linearnu kombinaciju određenih ulaznih vrednosti  $x$  i odgovarajućeg vektora težine  $w$ , gde je  $z$ , takozvani, mrežni ulaz [6].

$$\mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \quad 2.2.1.$$

$$z = w_1 x_1 + w_2 x_2 + \dots + w_m x_m = \mathbf{w}^T \mathbf{x} \quad 2.2.2.$$

Ako je mrežni ulaz određenog primerka veći od definisane granice  $\theta$ , predviđamo klasu 1, ili u suprotnom, klasu 0. U algoritmu perceptrona funkcija odlučivanja  $\sigma(\cdot)$  je varijanta jedinične step funkcije (engl. *Unit step function*)

$$\sigma(z) = \begin{cases} 1, z \geq \theta \\ 0, z < \theta \end{cases} \quad 2.2.3.$$

Zatim, radi pojednostavljenja implementacije, pomera se granica na levu stranu jednačine:

$$z \geq \theta \quad 2.2.4.$$

$$z - \theta \geq 0 \quad 2.2.5.$$

Definiše se slobodan član (engl. *bias unit*) kao  $b = -\theta$  i uključuje se u mrežni ulaz  $z$ :

$$w_1x_1 + w_2x_2 + \dots + w_mx_m + b = \mathbf{w}^T \mathbf{x} + b \quad 2.2.6.$$

Redefinisana funkcija odlučivanja sada izgleda ovako:

$$\sigma(z) = \begin{cases} 1, z \geq 0 \\ 0, z < 0 \end{cases} \quad 2.2.7.$$

Ovo je način funkcionisanja jednog neurona. Neuronska mreža se sastoji od više slojeva neurona, od kojih svaki sadrži proizvoljan broj neurona. Drugi sloj neurona donosi odluke na osnovu izlaza iz prvog sloja, i ovaj postupak se ponavlja do izlaznog sloja, što stvara kompleksnu strukturu donošenja odluka. Zbog ove složenosti, klasični veštački neuron je danas gotovo u potpunosti zamenjen neuronima sa nelinearnim aktivacionim funkcijama, koje omogućavaju mreži da se prilagodi složenijim obrascima podataka. [7] Prednosti i mane različitih aktivacionih funkcija opisane su u poglavlju 2.2.

Jedan od ključnih sastojaka algoritma mašinskog učenja i obuke neuronskih mreža je definisana ciljna funkcija (engl. *objective function*) koja bi trebalo da bude optimizovana u toku procesa učenja. Ova ciljna funkcija je često funkcija gubitka (engl. *loss function*), koju želimo da minimizujemo. Može se definisati funkcija gubitka  $L$  da bi se naučili parametri modela kao srednja kvadratna greška (engl. *Mean Square Error – MSE*) između izračunatog rezultata i tačne oznake klase:

$$L(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^n (\mathbf{y}^{(i)} - \sigma(\mathbf{z}^{(i)})) \quad 2.2.8.$$

Gde je  $\mathbf{y}^{(i)}$  i-ti stvarni izlaz a član  $\sigma(\mathbf{z}^{(i)})$  predstavlja proizvod mrežnog ulaza i aktivacione funkcije i predstavlja predviđen i-ti izlaz  $\hat{\mathbf{y}}^{(i)}$ , koji se u opštem slučaju razlikuje od  $\mathbf{y}^{(i)}$ . Glavna prednost ovako definisane funkcije cene je ta što je ona diferencijalna i konveksna, prema tome, može se upotrebiti efikasan algoritam optimizacije pod nazivom gradijenti spust (engl. *gradient descent*) za pronalaženje vektora težina koji minimalizuju funkciju cene gubitka  $L$ . Glavna ideja radijantnog spusta je da u svakoj iteraciji izvršavamo korak u suprotnom smeru gradijenta, gde je veličina koraka određena vrednošću brzine učenja i nagibom gradijenta dok se ne dostigne lokalni ili globalni minimum gubitka. Da bismo izračunali gradijent funkcije gubitka, potrebno je da izračunamo parcijalni izvod funkcije gubitka u odnosu na svaku težinu  $w_j$  i u odnosu na pristrasnost  $b$  kao:

$$\frac{\partial L}{\partial w_j} = -\frac{1}{n} \sum_{i=1}^n (\mathbf{y}^{(i)} - \sigma(\mathbf{z}^{(i)})) \mathbf{x}^{(i)} \quad 2.2.9.$$

$$\frac{\partial L}{\partial b} = -\frac{1}{n} \sum_{i=1}^n (\mathbf{y}^{(i)} - \sigma(\mathbf{z}^{(i)})) \quad 2.2.10.$$

Ažurirane težine i pristrasnost možemo da napišemo kao:

$$\Delta w_j = -\eta \frac{\partial L}{\partial w_j} \quad 2.2.11.$$

$$\Delta b = -\eta \frac{\partial L}{\partial b} \quad 2.2.12.$$

gde  $\eta$  predstavlja brzinu učenja (engl. *learning rate*) i predstavlja broj između 0 i 1.  $\Delta w_j$  predstavlja promenu vektora težine dok  $\Delta b$  predstavlja promenu pristrasnosti. [8]

## 2.2 Biranje aktivacione funkcije za višeslojne neuronske mreže

U višeslojnoj neuronskoj mreži može se, u principu, upotrebiti bilo koja diferencijalna funkcija aktivacije. Možemo da upotrebimo linearne funkcije aktivacije. Međutim, u praksi nije preporučljivo koristiti linearne funkcije aktivacije za skriveni i izlazni sloj, jer želimo da predstavimo nelinearnost u tipičnoj veštačkoj neuronskoj mreži da bismo mogli da rešimo složene probleme. Logistička (sigmoidna) funkcija aktivacije verovatno najbolje oponaša koncept neurona u mozgu – možemo to shvatiti kao verovatnoću da će se neuron aktivirati. Međutim, ta funkcija može da bude veoma problematična ako imamo izrazito negativan ulaz, jer će u ovom slučaju izlaz sigmoidne funkcije biti blizu nule. Ako sigmoidna funkcija vraća izlaz koji je blizu nule, neuronska mreža će da uči vrlo sporo i najverovatnije će biti “zarobljena” u lokalnom minimumu. Zbog toga se često preferira hiperbolička tangensna funkcija kao funkcija aktivacije u skrivenom sloju [9]. Formula logistične aktivacione funkcije je data sledećim izrazom:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad 2.2.1.$$

Procena verovatnoće klase u višeklasnoj klasifikaciji se često vrši primenom *softmax* funkcije u poslednjem potpuno povezanom sloju. Funkcija *softmax* je blaža forma *argmax* funkcije, umesto da daje jedan indeks klase, obezbeđuje verovatnoću svake klase. Prema tome, omogućava da izračunamo smislenu verovatnoću klase u višeklasnoj postavci. U softmax funkciji verovatnoća da će određeni uzorak sa zbirnim mrežnim ulazom  $z$  pripadati  $i$ -toj klasi može se izračunati pomoću člana normalizacije u imeniocu, odnosno zbirom eksponencijalno ponderisanih linearnih funkcija, gde  $j$  ide od 1 do  $M$ , a  $M$  je broj klasa:

$$p(z_i) = \frac{e^{z_i}}{\sum_{j=1}^M e^{z_j}} \quad 2.2.2.$$

Predviđen zbir svih ovih vrednosti po svim klasama ima vrednost 1, rezultat funkcije softmax može se zamisliti kao normalizovani izlaz koji je koristan za dobijanje smislenih predviđanja članstva klase u višeklasnom podešavanju.

Još jedna funkcija koja se često koristi u skrivenom sloju neuralne mreže je tangens hiperbolički ( $\tanh$ ), koji može da se tumači kao reskalirana verzija logističke funkcije:

$$\sigma(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad 2.2.3.$$

prednost hiperboličkog tangensa je u tome što ima širi spektar izlaza u rasponu u otvorenom intervalu  $(-1, 1)$ , što može da poboljša konvergenciju algoritma propagacije unazad.

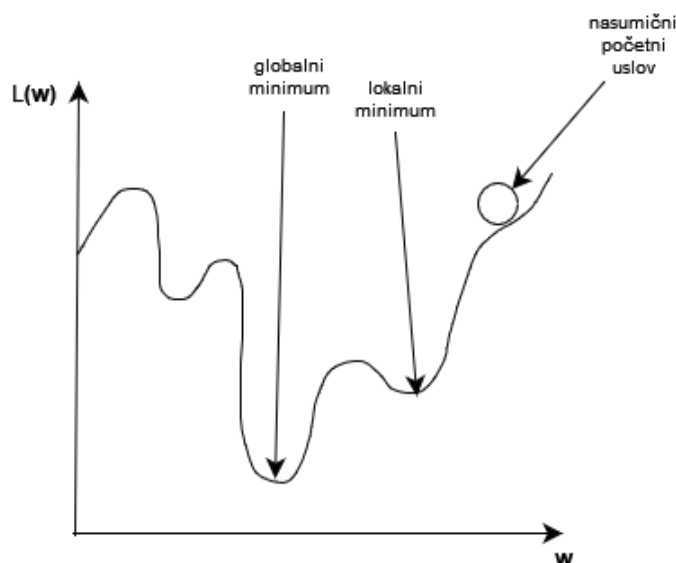
Ispravljачka linearna jedinica (engl. *Rectified Linear Unit* - ReLU) je funkcija aktivacije koja se često koristi u dubokim neuronska mrežama. ReLU je nelinearna funkcija koja je dobra za učenje kompleksnih funkcija na neuronskim mrežama, osim toga izvod ReLU funkcije je uvek 1 za pozitivne

vrednosti ulaza. ReLU rešava problem nestajućeg gradijenta, čineći je pogodnom za duboke neuralne mreže. [10]

$$\sigma(z) = \max(0, z) \quad 2.2.4.$$

## 2.3 Konvergencija u neuronskim mrežama

Višeslojne neuronske mreže su mnogo teže za obučavanje od jednostavnijih algoritama, kao što su logistička regresija ili metod potpunih vektora. U višeslojnim neuronskim mrežama obično imamo stotine, hiljade pa čak i milijarde (transformatori, GPT modeli) težina koje je potrebno optimizovati. Izlazna funkcija ima grubu površinu i algoritam optimizacije može lako da ostane “zarobljen” u lokalnim minimumima, kao što je prikazano na slici 2.3.1:



Slika 2.3.1. Algoritam za optimizaciju mogu da postanu zarobljeni u lokalnim minimumima

uvećanjem brzine učenja  $\eta$  možemo da izbegnemo lokalni minimum. Sa druge strane, takođe uvećavamo mogućnost preskakanja globalnog minimuma.

Poboljšanje gradijentnog spusta je moguće ostvariti pomoću skaliranja atributa. Ova procedura normalizacije pomaže učenju gradijentnog spusta da brže konvergira, ali ne čini originalni skup podataka normalno raspoređenim. Standardizacija pomera srednju vrednost svakog atributa tako da je on centriran u nuli, svaki atribut ima standardno odstupanje od 1 (jedinična varijansa). Jedan od razloga zašto standardizacija pomaže pri učenju gradijentnog spusta je to što je jednostavnije pronaći brzinu učenja koja dobro funkcioniše za sve težine i pristrasnosti. Ako su atributi znatno različitih redova veličina, brzina učenja koja dobro funkcioniše za ažuriranje jedne težine biće prevelika ili premala za ažuriranje druge težine. [11]

Kvalitetna alternativa za algoritam gradijentnog spusta je stohastički gradijentni spust (engl. *Stochastic gradient descent – SGD*) koji se naziva i iterativni gradijentni spust. Umesto da ažuriramo težine na osnovu akumuliranih grešaka na svim primerima za obučavanje, ažuriramo parametre inkrementalno za svaki primer za obučavanje:

$$\Delta w_j = \eta (y^{(i)} - \sigma(z^{(i)})) x_j^{(i)} \quad 2.3.1.$$

$$\Delta b = \eta(\mathbf{y}^{(i)} - \sigma(\mathbf{z}^{(i)})) \quad 2.3.2.$$

Pošto se svaki gradijent izračunava na osnovu jednog primera za obučavanje, oblast greške ima više šuma nego u gradijentom spustu, što omogućava SGD algoritmu da izbegne “plitke” lokalne minimume [12].

## 2.4 Konvolucione i prostorno-vremenske konvolucione neuralne mreže

Za prostornu analizu koriste se dobro poznate konvolucione neuronske mreže (engl. *Convolutional Neural Networks* - *CNN*). CNN predstavljaju familiju modela koja je prvobitno inspirisana načinom na koji vizuelni korteks ljudskog mozga prepoznaje objekte. Razvoj CNN seže do 1990-ih godina, kada je prvi put predstavljena nova arhitektura neuronskih mreža za klasifikaciju ručno pisanih cifara iz slika [13]. Uspešno izdvajanje istaknutih (relevantnih) atributa ključno je za performanse bilo kog algoritma mašinskog učenja. Određeni tipovi neuronskih mreža, poput CNN, imaju sposobnost da automatski uče attribute iz neobrađenih podataka koji su najkorisniji za određeni zadatak. Zbog toga se slojevi CNN često posmatraju kao izdavači atributa. Rani slojevi, koji su odmah iza ulaznog sloja, izdvajaju attribute niskog nivoa iz neobrađenih podataka, dok kasniji slojevi prepoznaju sve kompleksnije karakteristike slike. Ove karakteristike su ključne za prepoznavanje lokalnih atributa koji utiču na tačnu klasifikaciju fonema na osnovu slike. CNN omogućavaju identifikaciju relevantnih delova slike koji su od presudnog značaja za pravilno prepoznavanje vizuelnih informacija.

Pretpostavimo da imamo ulazni tenzor  $\mathbf{X}$  sa dimenzijama  $(H, W, C)$  gde je  $H$  visina slike,  $W$  širina slike a  $C$  broj kanala (npr. 3 za RGB slike) i primenjujemo 2D konvolucioni filter sa dimenzijama  $(H_f, W_f, C, K)$  gde je  $H_f$  visina filtra,  $W_f$  širina filtra,  $C$  broj kanala u ulazu i  $K$  broj izlaznih kanala. Tada rezultat konvolucije  $\mathbf{Y}$  ima dimenzije  $(H', W', K)$ , gde  $H'$  i  $W'$  zavise od dimenzija ulaza, filtera i koraka. Formula za 2D konvoluciju se može definisati na sledeći način:

$$\mathbf{Y}(h, w, k) = \sum_{h'=0}^{H_f-1} \sum_{w'=0}^{W_f-1} \sum_{c=0}^{C-1} \mathbf{X}(h + i, w + j, c) \cdot \mathbf{F}(i, j, c, k) + B(k) \quad 2.4.1.$$

$\mathbf{X}(h + i, w + j, c)$  je vrednost piksela u ulazu u poziciji  $(h + i, w + j)$  za kanal  $C$ ,  $\mathbf{F}(i, j, c, k)$  je težina filtera u poziciji  $(i, j)$  za kanal  $C$  i izlazni kanal  $k$ , a  $B(k)$  je pristrasnost za izlazni kanal  $k$ .

Kod klasičnih CNN postoji ograničenje u radu sa vremenski promenljivim podacima jer ne uzimaju u obzir vremensku dimenziju, što se može i primetiti u jednačini 2.4.1. Kako je vizuelno prepoznavanje govora zasnovano na video snimcima, koji su po svojoj prirodi vremenski promenljivi, potrebna je konvoluciona neuronska mreža koja može da obradi i prostorne i vremenske promene u podacima.

U video snimcima svaki frejm u datom trenutku je u visokoj korelaciji sa frejmovima koji su se pojavili neposredno pre njega. Samim tim, tačna klasifikacija često zahteva ne samo informacije sadržane u trenutnom frejmu, već i one koje su se pojavile u prethodnim frejmovima. Za vizuelno prepoznavanje govora, gde je prisutna kontinuirana promena slike kroz vreme, neophodna je posebna vrsta CNN poznata kao prostorno-vremenske konvolucione neuralne mreže (engl. *spatial-Temporal Convolutional Neural Networks* - *STCNN*). STCNN vrši konvoluciju ne samo kroz prostornu dimenziju već i kroz vremensku dimenziju, omogućavajući modelu da detektuje i analizira promene koje se dešavaju između frejmova. Osnovna ideja je da se koristi 3D konvolucija. To ovaj model čini posebno pogodnim za zadatke kao što je vizuelno prepoznavanje govora, gde je važno razumeti kako se položaj usana menja tokom vremena [14].

Ako imamo ulazni tenzor  $\mathbf{X}$  sa dimenzijama  $(T, H, W, C)$  gde je,  $T$  broj vremenskih frejmova (engl. *temporal depth*),  $H$  visina slike (engl. *height*),  $W$  širina slike (engl. *width*),  $C$  broj kanala, i primenjujemo 3D konvolucioni filter  $\mathbf{F}$  sa dimenzijama  $(T_f, H_f, W_f, C, K)$  gde je  $T_f$  dubina filtera,  $H_f$  visina filtera



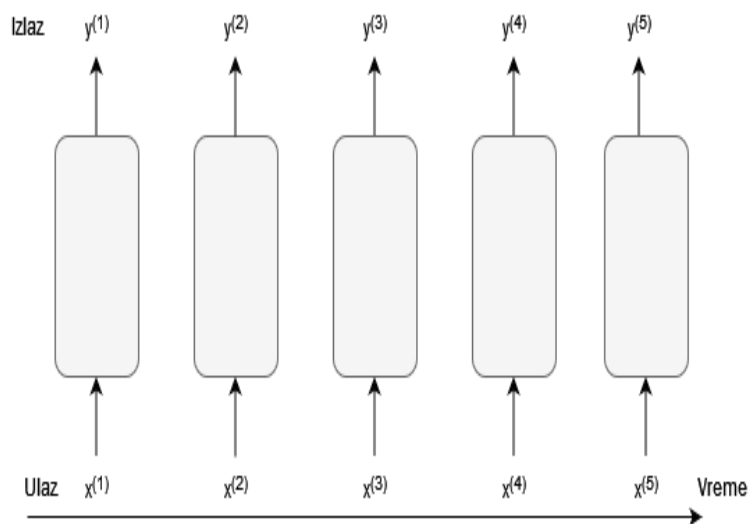
$W_f$  širina filtera,  $K$  broj izlaznih kanala (engl. *output channels*), onda je rezultat konvolucije  $\mathbf{Y}$  sa dimenzijama  $(T', H', W', K)$ , gde  $T', H', W'$  zavise od dimenzija ulaza, filtera, i koraka koji se koriste. Tada se 3D konvolucija može izraziti kao:

$$\mathbf{Y}(t, h, w, k) = \sum_{t'=0}^{T_f-1} \sum_{h'=0}^{H_f-1} \sum_{w'=0}^{W_f-1} \sum_{c=0}^{C_f-1} \mathbf{X}(t+t', h+h', w+w', c) \cdot \mathbf{F}(t', h', w', c, k) + B(k) \quad 2.4.2.$$

gde je  $\mathbf{X}(t+t', h+h', w+w', c)$  vrednost piksela u ulazu,  $\mathbf{F}(t', h', w', c, k)$  težina filtera,  $B(k)$  pristrasnost za izlazni kanal  $k$ .

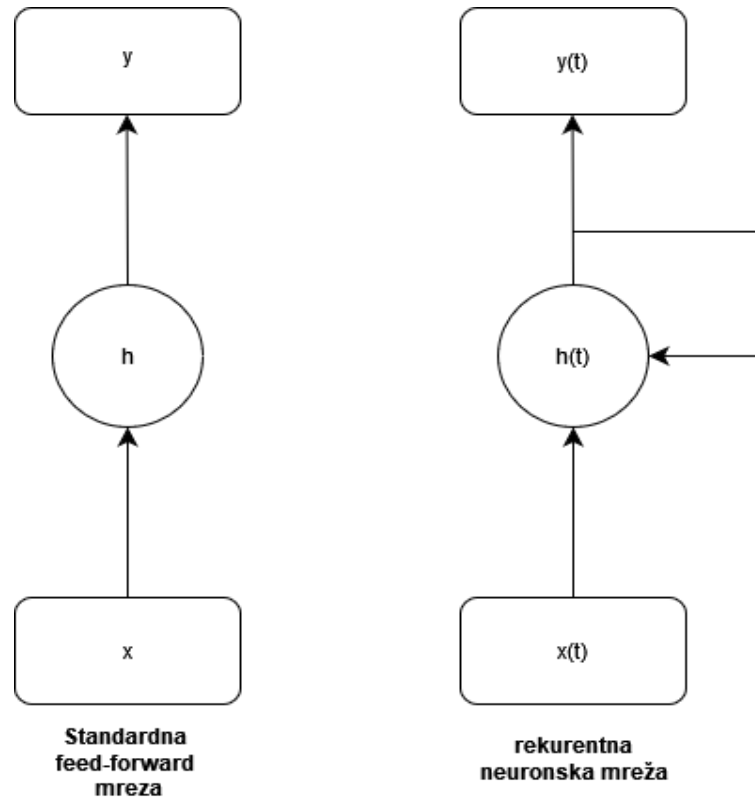
## 2.5 Rekurentne neuronske mreže

Za analizu sekvencijalnih podataka koriste se rekurentne neuronske mreže (RNN), ono što sekvencijalne podatke tj. sekvence čini jedinstvenim u poređenju sa drugim tipovima podataka jeste činjenica da se elementi u sekvenci prikazuju u određenom redosledu i nisu nezavisni jedan od drugog. Tipični algoritmi mašinskog učenja za nadgledano učenje pretpostavljaju da ulaz čine nezavisni i identično distribuirani (IID) podaci, što znači da su primeri za obučavanje međusobno nezavisni i da imaju istu osnovnu gustinu raspodele verovatnoće. Prema tome, na osnovu pretpostavke međusobne nezavisnosti, redosled u kojem su primeri za obučavanje dati modelu nije važan. Međutim, ova pretpostavka nije validna kada koristimo vremenske sekvence – prema definiciji, redosled je važan. Kako je zadatak da predvidimo fonem izgovoren na slici, neophodno je razmotriti prethodne pozicije usana po redosledu sortiranom prema vremenu prikazivanja, umesto da ove primere za obučavanje upotrebimo po nasumičnom redosledu. Indeksi ukazuju na redosled instanci, a dužina sekvence je  $T$ . U podacima vremenske sekvence, u kojima svaka tačka primera  $x(t)$  pripada određenom vremenskom trenutku  $t$ . Na slici 2.5.1 prikazan je primer podataka vremenske sekvence u kojima ulazni atributi  $x$  i ciljne oznake  $y$  prirodno prate redosled u skladu sa njihovim vremenskim osama. [15]



Slika 2.5.1. Primer podataka vremenske sekvence

U standardnoj nerekurentnoj (engl. *Feedforward*) mreži informacije teku iz ulaza do skrivenog sloja, a zatim iz skrivenog sloja do izlaznog sloja. Sa druge strane, u RNN skriveni sloj prima ulaz iz aktuelnog vremenskog koraka i skrivenog sloja iz prethodnog vremenskog koraka. Tok informacija u uzastopnim vremenskim koracima u skrivenom sloju omogućava mreži da pamti prošle događaje. Ovaj tok je, obično, prikazan kao petlja koja je poznata kao povratna grana u grafovskoj terminologiji, po čemu je i osnovna RNN arhitektura dobila ime. Način obrade podataka u nerekurentnoj i rekurentnoj neuronskoj mreži prikazan je na slici 2.5.2:



Slika 2.5.2. Tok podataka standardne feedforward i rekurentne neuralne mreže

Algoritam propagacije unazad kroz vreme (engl. *Back-Propagation Through Time* – BPTT) je čest i efikasan način obuke RNN-a, predstavljen 1990. godine [16]. Izvod gradijenta može biti komplikovan, ali je osnovna ideja da je ukupan gubitak  $L$  zbir svih funkcija gubitka u trenucima  $t = 1$  do  $t = T$ :

$$L = \sum_{t=1}^T L(t) \quad 2.5.3.$$

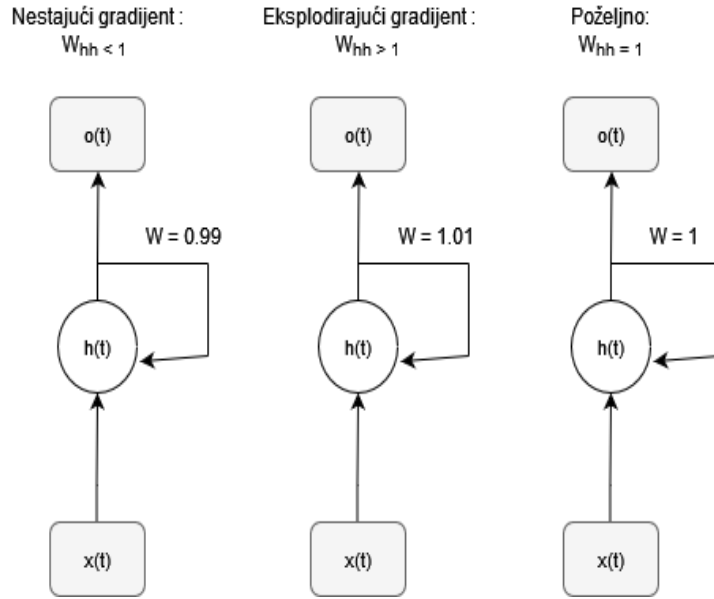
gubitak u trenutku  $t$  zavisen je od skrivenih jedinica u svim prethodnim vremenskim koracima od 1 do  $t$ , gradijent će biti izračunat na sledeći način:

$$\frac{\partial L^{(t)}}{\partial \mathbf{w}} = \frac{\partial L^{(t)}}{\partial \mathbf{o}} \cdot \frac{\partial \mathbf{o}}{\partial \mathbf{h}} \cdot \left( \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} \cdot \frac{\partial \mathbf{h}^{(k)}}{\partial \mathbf{w}} \right) \quad 2.5.4.$$

Ovde je  $\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}}$  izračunat kao proizvod po uzastopnim vremenskim koracima:

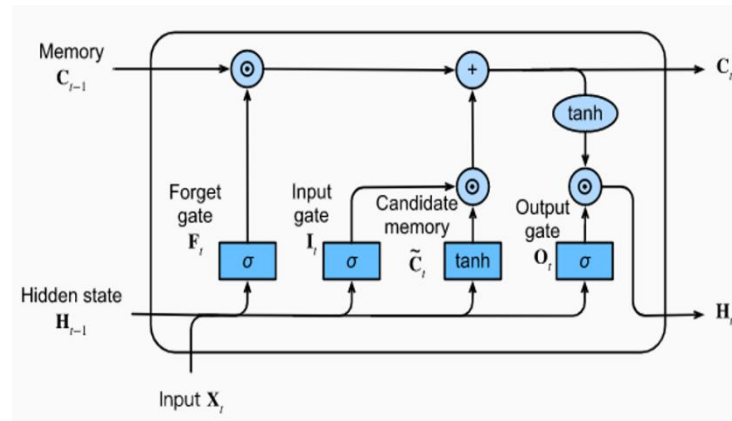
$$\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}} = \prod_{i=k+1}^t \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}} \quad 2.5.5.$$

BPTT uvodi neke nove izazove. Zbog faktora  $\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}}$  u izračunavanju gradijenta funkcije gubitka javljaju se problemi takozvanog nestajućeg i eksplodirajućeg gradijenta [17]. U suštini,  $\frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(k)}}$  ima  $t - k$  množenja, prema tome, množenje težine  $w$  samom sobom  $t - k$  puta rezultira faktorom  $w^{t-k}$ . Kao rezultat, ako je  $|W| < 1$ , ovaj faktor postaje veoma mali kada je  $t - k$  malo. Sa druge strane, ako je težina povratne grane,  $|W| > 1$ ,  $w^{t-k}$  postaje veoma velik kada je  $t - k$  veliko. Jedno od rešenja bi bilo osigurati da je  $|W| = 1$ , kao što je prikazano na slici 2.5.3.:



Slika 2.5.3 Problemi u izračunavanju gradijenta

*Long Short-Term Memory* - LSTM je memorijska ćelija koja zamenjuje skriveni sloj RNN-a, a prvi put predstavljena za rešavanje problema nestajućeg gradijenta 1997. [18], njena blok šema je prikazana na slici 2.2.4. U svakoj memorijskoj ćeliji postoji povratna grana koja ima poželjnu težinu  $|W| = 1$ , kao što je pomenuto, za rešavanje problema nestajućeg i eksplodirajućeg gradijenta. Vrednosti povezane sa ovom povratnom granom zajednički se nazivaju stanjem ćelije.



Slika 2.5.4. blok šema LSTM ćelije 1

Stanje iz prethodnog vremenskog koraka  $C_{t-1}$  je modifikovano da bi se dobilo stanje ćelije u aktuelnom vremenskom trenutku  $C_t$ , bez množenja direktno sa bilo kojim faktorom težine. Tok informacija u ovoj memorijskoj ćeliji kontroliše nekoliko računarskih jedinica, koje često nazivamo kapije (engl. *gates*). Na prethodnoj slici  $\oplus$  predstavlja elementarno sabiranje, a  $\odot$  predstavlja elementarno množenje.  $\mathbf{X}_t$  se odnosi na ulazne podatke u trenutku  $t$ , a  $\mathbf{H}_t$  ukazuje na skrivene jedinice u trenutku  $t - 1$ . Četiri polja su označena aktivacionom funkcijom, ili sigmoidnom funkcijom ili tangensom hiperboličnim, kao i skupom težina. Polja primenjuju linearnu kombinaciju izvršavanjem množenja matrice ili vektora na njihovim ulazima, a to su  $\mathbf{h}(t - 1)$  i  $\mathbf{X}_t$ . Ove jedinice izračunavaju se pomoću sigmoidne funkcije aktivacije, čije su izlazne jedinice prosledjene kroz  $\odot$ , nazvane kapije. U LSTM postoje tri različita tipa kapija – kapija zaborava, kapija prolaza i izlazna kapija.

Kapija zaborava,  $f_t$ , omogućava memorijskoj ćeliji da resetuje stanje ćelije bez beskonačnog povećanja. Ona odlučuje kojim informacijama je dozvoljeno da prolaze i koje informacije će biti zaustavljene. Ona

nije bila deo originalne LSTM ćelije, dodata je nekoliko godina kasnije da bi bio poboljšan polazni model [11]. Jednačina je sledeća:

$$f_t = \sigma(\mathbf{w}_{x_f} \mathbf{x}_t + \mathbf{w}_{hf} \mathbf{h}(t-1) + \mathbf{b}_f) \quad 2.5.6.$$

ulazna kapija,  $\mathbf{i}_t$ , i kandidatska vrednost  $\mathbf{C}_t$  odgovorne su za ažuriranje stanja ćelije. Računaju se na sledeći način:

$$\mathbf{i}_t = \sigma(\mathbf{w}_{x_i} \mathbf{x}_t + \mathbf{w}_{hi} \mathbf{h}(t-1) + \mathbf{b}_i) \quad 2.5.7.$$

$$\tilde{\mathbf{C}}_t = \tanh(\mathbf{w}_{x_c} \mathbf{x}_t + \mathbf{w}_{hc} \mathbf{h}(t-1) + \mathbf{b}_c) \quad 2.5.8.$$

stanje ćelije u trenutku  $t$  se računa na sledeći način:

$$\mathbf{C}_t = (\mathbf{C}_{t-1} \odot \mathbf{f}_t) \oplus (\mathbf{i}_t \odot \tilde{\mathbf{C}}_t) \quad 2.5.9.$$

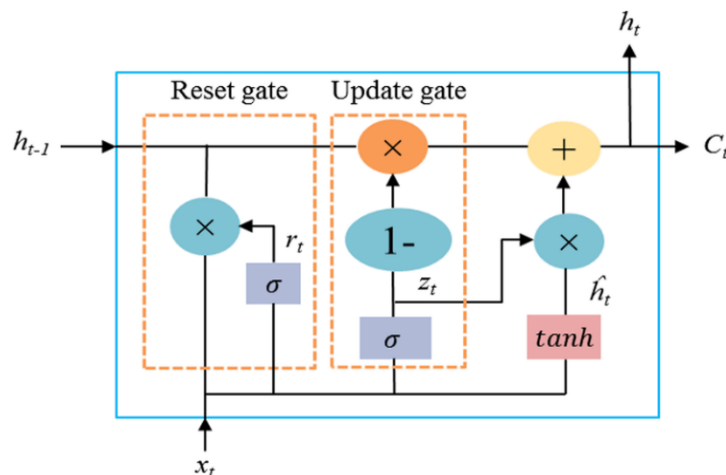
izlazna kapija,  $\mathbf{O}_t$ , odlučuje kako će biti ažurirane vrednosti skrivenih jedinica:

$$\mathbf{O}_t = \sigma(\mathbf{w}_{x_o} \mathbf{x}_t + \mathbf{w}_{ho} \mathbf{h}_{t-1} + \mathbf{b}_o) \quad 2.5.10.$$

skrivenne jedinice u aktuelnom vremenskom koraku izračunavamo na sledeći način:

$$\mathbf{h}^{(t)} = \mathbf{O}_t \odot \tanh(\mathbf{C}_t) \quad 2.5.11.$$

*Gated Recurrent Unit* - GRU je tip rekurentne neuronske mreže koji ima jednostavniju arhitekturu nego LSTM, prema tome, računarski su mnogo efikasnije, i njihov način rada je uporediv sa LSTM mrežama [19]. GRU Posедуje dve kapije, *update gate* i *reset gate*. Istraživači i inženjeri često testiraju obe vrste neuralne mreže i empirijski utvrđuju da li LSTM ili GRU daje bolje rezultate [20][21]. Blok šema GRU neuronske mreže je prikazan na slici 2.5.5.:



Slika 2.5.5. blok šema ćelije GRU Neuralne

$$[\mathbf{u}_t \mathbf{r}_t]^T = \sigma(\mathbf{w}_z + \mathbf{w}_h \mathbf{h}_{t-1} + \mathbf{b}_g) \quad 2.5.12$$

$$\tilde{\mathbf{h}} = (\mathbf{v}_z \mathbf{z}_t + \mathbf{U}_h (\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad 2.5.13$$

$$\mathbf{h} = (1 - \mathbf{U}_t) \odot \mathbf{h}_{t-1} + \mathbf{u}_t \odot \tilde{\mathbf{h}}$$

2.5.14

U rekurentnim neuronskim mrežama sekvencijalni podaci predstavljeni su ulaznom sekvencom  $[Z_1, \dots, Z_T]$ . U velikom broju radova navedenim u ovom diplomskom radu, kao i u samom projektu, koristi se **bi-direkionalna GRU** (engl. *bidirectional l gated reccurent network* - Bi-GRU) mreža. Bi-GRU mreža se sastoji od dva RNN: jedna obrađuje ulazne podatke od početka do kraja  $[Z_1, \dots, Z_T]$ , dok druga obrađuje ulazne podatke od kraja do početka  $[Z_T, \dots, Z_1]$ . Izlazi obe neuronske mreže obično se nadovezuju u svakom vremenskom koraku, što omogućava mreži da čuva informacije iz prošlosti i budućnosti. Ovakva arhitektura omogućava bolji uvid u kontekst sekvencijalnih podataka, što je ključno za tačnije prepoznavanje i razumevanje složenih obrazaca u vremenskim sekvencama.

Potpuno povezani slojevi (engl. *fully-connected layers*) su slojevi gde je svaki neuron jednog sloja povezan sa svakim neuronom narednog sloja. svaki čvor u sloju  $n$  je povezan sa svim čvorovima u sloju  $n + 1$  pomoću koeficijenta težine i ovakva vrsta neuralne mreže nam dozvoljava višeklasnu klasifikaciju pomoću generalizacije tehnike jedan protiv svih (engl. *one-versus-all* - OvA). [22]

## 2.6 Funkcija gubitka u VPG

Connectionist Temporal Classification - CTC [23] je funkcija gubitka koja je često korišćena u zadacima sekvencijalnog predviđanja, kao što su vizuelno prepoznavanje govora. Ova funkcija omogućava da se dekodira sekvenca promenljive dužine iz ulaznog niza sa fiksnom dužinom, čak i kada su oznake poravnate s ulazima na nelinearan način.

CTC funkcija gubitka omogućava modelu da poravna ulazne sekvence na izlazne oznake bez potrebe za eksplicitnim poravnavanjem između ulaza i izlaza. U zadatku prepoznavanja govora, ulazna sekvenca može imati mnogo više vremenskih koraka nego što ima fonema u reči koju treba prepoznati. CTC koristi koncept poravnanja koji uključuje umetanje praznog (engl. *blank*) simbola između izlaza modela kako bi se omogućila fleksibilnost u vremenskom poravnavanju.

Neka je:

$\mathbf{x} = (x_1, x_2, \dots, x_T)$  sekvenca predikcija modela dužine  $T$ , gde svaka  $x_t$  predstavlja verovatnoću za svaki simbol iz skupa  $L$ , uključujući i prazan simbol (engl. *blank*).

$\mathbf{y} = (y_1, y_2, \dots, y_U)$  sekvenca ciljnih oznaka dužine  $U$ , gde je  $U \leq T$

$\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_U)$  predstavlja jednu od mogućih sekvenci iz  $L'$  koja može proizvesti  $\mathbf{y}$

CTC računa verovatnoću svih mogućih poravnanja između predikcija i ciljne sekvence. Verovatnoća sekvence  $\mathbf{y}$  data ulaznom sekvencom  $\mathbf{x}$  je suma verovatnoća svih mogućih poravnanja  $\boldsymbol{\pi}$  koja odgovaraju  $\mathbf{y}$ :

$$P(\mathbf{y}|\mathbf{x}) = \sum_{\boldsymbol{\pi} \in B^{-1}(\mathbf{y})} P(\boldsymbol{\pi}|\mathbf{x}) \quad 2.6.1.$$

Ovde  $B^{-1}(\mathbf{y})$  označava skup svih mogućih sekvenci  $\boldsymbol{\pi}$  koje se, nakon uklanjanja praznog simbola i spajanja uzastopnih ponavljanja, preslikaju u  $\mathbf{y}$ . Verovatnoća sekvence  $\boldsymbol{\pi}$  se računa kao proizvod verovatnoća po uzastopnim vremenskim koracima:

$$P(\boldsymbol{\pi}|\mathbf{x}) = \prod_{t=1}^T P(\pi_t|x_t) \quad 2.6.2.$$

Gubitak se zatim definiše kao negativni logaritam ukupne verovatnoće ciljne funkcije  $\mathbf{y}$ :

$$L_{\text{CTC}}(\mathbf{y}, \mathbf{x}) = -\log P(\mathbf{y}|\mathbf{x})$$

2.6.3.

Intuicija rada CTC funkcije gubitka je sledeća:

1. Dodati prazan simbol u set mogućih karaktera
2. Dizajnirati mrežu da klasifikuje svaku ulaznu vrednost
3. Kolabirati ponavljajuće izlaze i prazan simbol
  - AAAAAA → A
  - AAABBB → AB
  - AAA\_\_BB → AB
  - AAA\_AA → AA
4. Uporediti kolabiran izlaz sa ciljnom sekvencom

Zbog eksponencijalnog broja mogućih poravnanja, direktno izračunavanje ove verovatnoće bi bilo neefikasno. Umesto toga, koristi se napred-nazad (engl. *forward-backward*) algoritam, koji efikasno računa verovatnoću ciljane sekvence sumiranjem preko mogućih poravnanja koristeći dinamičko programiranje.

### 3 Pregled postojeće literature

U diplomskom radu analizirano je 7 različitih radova koji se bave temom vizuelnog prepoznavanja govora. Istraživanje na temu automatskog čitanja sa usana ima dugu istoriju. Prvi radovi iz ove oblasti bave se unapređivanjem tačnosti i uspešnosti klasifikacije fonema, dok kasniji modeli počinju uspešno da klasifikuju i prepoznaju cele reči i rečenice. Svaki rad donosi doprinos ovoj oblasti, bilo rešavanjem problema koji se javljaju tokom obuke, treniranja i primene modela, ili uvođenjem novih metoda i ideja za prevazilaženje izazova u ovoj oblasti.

Trenutno postoji nekoliko baza podataka koje su posebno pogodne za vizuelno prepoznavanje govora [24]. Sa stanjem iz septembra 2024. godine, tabela 3.1 prikazuje ključne karakteristike ovih baza podataka, uključujući veličinu svake baze, broj izgovorenih rečenica, kao i uslove u kojima su rečenice izgovorene, bilo da se radi o prirodnom okruženju (*P*) ili kontrolisanim uslovima (*K*).

Naziv baze	veličina	Broj rečenica	P/K
<a href="#">LRW</a>	~70 GB	~500 000 R	P
<a href="#">LRS2</a>	~50 GB	~200 000 R	P
<a href="#">LRS3-TED</a>	~70 GB	~500 000 R	P
<a href="#">CAS-VSR-W1k</a>	~65 GB	~1 000 000 R	P
<a href="#">GRID</a>	6.5 - 40GB *	~ 33 000 R	K
<a href="#">CMLR</a>	~60 GB	~10 000 R	K

Tabela 3.1 prikaz dostupnih baza za VPG

\* veličina *GRID* baze zavisi od veličine video snimaka koji se preuzimaju, koji može da varira između 190KB i 1.2GB po govorniku, ukupno ima 32 govornika. Važno je napomenuti da za rezoluciju 1.2GB i 420 MB nedostaju podaci za jednog govornika

Da bi se izmerila uspešnost nekog modela vizuelnog prepoznavanja govora, u oblast automatskog prepoznavanja govora uvedene su metrike *Word Error Rate* - *WER* i *Characher Error Rate* - *CER*, koje predstavljaju standardizovan način procene uspešnosti modela [25]. *WER* metrika može da ima vrednost između 0 i 1, gde 0 znači da su predviđen i originalan tekst identični a 1 znači da su sasvim različiti.

$$WER = (S + D + I) / N \quad 3.1.$$

gde je *S* broj reči koje su pogrešno prepoznate, *D* je broj reči koje su izostavljene, *I* broj reči koje su dodatno uključene i *N* je broj reči u referentom tekstu.

*CER* metrika je metrika koja nam daje procenat karaktera koji su netačno predviđeni. Što je niža vrednost, bolja je perfomansa VPG modela. *CER* može da ima vrednost između 0 i 1, gde 0 znači da su predviđen i originalan tekst identični a 1 znači da su kompletno različiti.

$$CER = (S + D + I) / N \quad 3.2.$$

gde je *S* broj karaktera koji su pogrešno prepoznati, *D* je broj karaktera koji su izostavljeni, *I* su dodatno uključeni i *N* je broj karaktera u referentnom tekstu.

WER i CER metrike se često množe koeficijentom 100 kako bi se uspešnost modela prikazala u procentualnom obliku, što olakšava interpretaciju rezultata i omogućava jednostavnije poređenje različitih modela.

### 3.1 Automatsko VPG

Istraživači sa George Washington univerziteta [26] su uradili jedan od najzanimljivijih i prvih radova na temu čitanja sa usana koji ne koristi neuronske mreže. Razlog za to leži u činjenici da ovakav pristup rešavanju problema zahteva značajno procesiranje frejmova kako bi se izdvojili relevantni atributi slike, a hardverski i softverski resursi u to vreme još uvek nisu bili dovoljno razvijeni. Sistem koristi skrivene Markovljeve modele (engl. *HMM*) obučene da razlikuju optičke informacije i postiže tačnost prepoznavanja od 25,3 procenata na 150 test rečenica. Ovo je prvi sistem koji omogućava kontinuirano optičko automatsko prepoznavanje govora (engl. *OASR*). Ovaj nivo performansi, bez korišćenja sintaktičkih, semantičkih ili bilo kojih drugih kontekstualnih vodiča u procesu prepoznavanja, ukazuje da OASR može biti značajan dodatak za robusno više-modalno prepoznavanje u bučnim okruženjima. Pored toga, otkrivene su nove karakteristike važne za OASR, a korišćeni su novi pristupi vektorskoj kvantizaciji, obuci i klasterovanju. Ova studija sadrži tri glavne komponente.

Prvo, postavljena je hipoteza o 35 statičnih i dinamičkih optičkih karakteristika koje karakterišu senku usne šupljine govornika. Koristeći odgovarajuću matricu korelacije i analizu glavnih komponenti, studija je odbacila 22 karakteristike usne šupljine. Preostalih 13 karakteristika usne šupljine su uglavnom dinamičke, za razliku od statičnih karakteristika koje su koristili prethodni istraživači.

Drugo, studija je spojila foneme koji su optički slični u oblasti usne šupljine govornika u vizeme. Vizemi su objektivno analizirani i diskriminirani koristeći HMM i algoritme klasterovanja. Što je najvažnije, vizemi za govornika, dobijeni računanjem, u skladu su sa mapiranjem fonema na vizeme o kojem govori većina stručnjaka za čitanje s usana. Ova sličnost, u određenom smislu, potvrđuje izbor karakteristika usne šupljine.

Treće, studija je obučila HMM-ove da prepoznaju, bez gramatike, set rečenica od 150 reči, koristeći vizeme, trizeme (tri vizema) i generalizovane trizeme (klasterizovane trizeme). Sistem je postigao stope prepoznavanja od 2 %, 12,7 procenata i 25,3 procenata koristeći, redom, HMM-ove za vizeme, HMM-ove za trizeme i HMM-ove za generalizovane trizeme. Studija zaključuje da metodologije korišćene u ovoj istrazi pokazuju potrebu za daljim istraživanjem kontinuiranog OASR i integracije optičkih informacija sa drugim metodama prepoznavanja. Dok se ova studija fokusira na izvodljivost, validnost i segregirani doprinos isključivo kontinuiranom OASR, budući visoko robusni sistemi prepoznavanja trebalo bi da kombinuju optičke i akustičke informacije sa sintaktičkim, semantičkim i pragmatičkim informacijama.

Autori rada [27] su uspeli da unaprede prepoznavanje govora na nivou reči i rečenica u bučnim sredinama kombinovanjem vizuelnih atributa sa audio atributima. Ponovo su koristili skrivene Markovljeve modele (HMM), u kombinaciji sa manuelnim izmenama, kako bi unapredili algoritam. Baza podataka korišćena u ovom istraživanju je *IBM ViaVoice*, koja sadrži 17.111 rečenica od 261 različitog govornika za trening (oko 35 sati). Ova baza podataka nije javno dostupna. Iako se njihov rad ne može smatrati isključivo vizuelnim prepoznavanjem govora, predstavlja značajan napredak u oblasti prepoznavanja govora. WER sa samo vizuelnim informacijama je 51.08% dok je WER na audio snimcima 48.10%

### 3.2 Duboko učenje u VPG

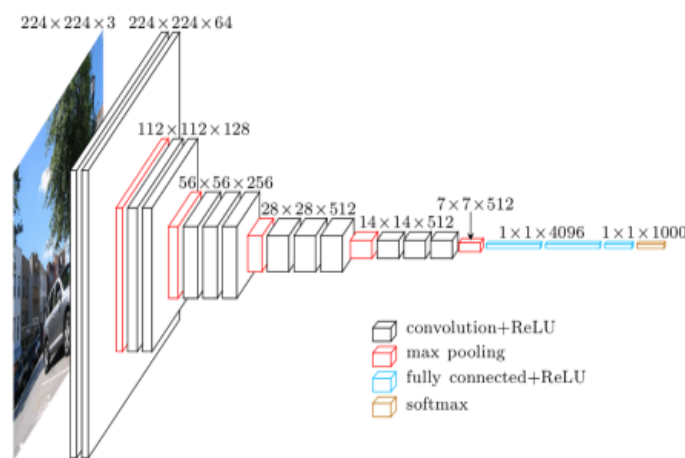
Oblast automatskog prepoznavanja govora (engl. *Automatic speech recognition* - ASR) ne bi bio u stanju u kom je danas bez modernih dostignuća u oblasti, od kojih se mnoga prvi put pojavljuju u konekstu ASR u radovima. Pojavom CTC funkcije gubitka duboko učenje prerasta iz komponente vizuelnog prepoznavanja govora u sisteme koji koriste isključivo duboko učenje od početka procesa do kraja procesa, takozvani end-to-end sistemi. , (engl. *Deep Learning ASR*)



End-to-End model u ovom slučaju znači da se ceo proces, od obrade video snimaka (npr. detekcija usana, izdvajanje vizuelnih karakteristika) do prepoznavanja i transkripcije izgovorenog teksta, odvija unutar jedne arhitekture neuralne mreže. To znači da nije potrebno ručno definisati ili programirati različite faze obrade, kao što su segmentacija usana ili ekstrakcija atributa – sve to automatski obavlja model.

*Lip Reading In the Wild* [28] je rad sa univerziteta Oksford. Baza podataka LRW je značajno složenija u odnosu na baze u prethodno pomenutim radovima. Obradom je obuhvaćen izuzetno veliki skup podataka sa video snimcima, prikupljenim pretežno sa britanskih televizijskih programa (BBC). Iz baze je uspešno izdvojeno preko 1000 sati govornog materijala sa vokabularom od preko 500 000 različitih rečenica i više od 1000 različitih govornika.

Prednosti ovog rada uključuju izgradnju velike baze podataka koji su prikupljeni iz stvarnog sveta. Tehnološki napreci koji su ostvareni su integracija prostorno-vremenskih karakteristika, korišćenjem VGG-M arhitekture[26] za analizu vizuelnih karakteristika lica što je korak ka boljim klasifikacijama.



Slika 3.1 Arhitektura VGG Neuralne mreže

Osmišljena su četiri različita modela, svaki s naglaskom na obradu  $T$  ulaznih frejmova ( $T = 25$ , što odgovara intervalu od 1s). Među njima, model "3D Convolution with Multiple Towers (MT-3)" pokazuje najbolje rezultate. Ovaj model obrađuje svaki frejm nezavisno pre nego što podaci dospeju do konvolucionog sloja, što se razlikuje od ostalih modela koji povezuju frejmove ranije u arhitekturi. Izlazi *pool* sloja svakog frejma se zatim nadovezuju, što omogućava modelu da integriše informacije iz različitih vremenskih tačaka pre nego što se primeni 3D konvolucija. Ovaj pristup omogućava modelu da efikasnije prepozna prostorno-vremenska obeležja.

Mane ovog rada su te što iako koriste prostorno-vremenske modele (STCNN) oni nisu postigli značajno bolje rezultate od prostornog modela (CNN), što sugerise da njihova metoda nije u potpunosti iskoristila informacije iz vremenskih sekvenci. Model nije mogao da obradi sekvence promenljive dužine, što predstavlja značaj nedostak za primenu u realnim scenarijima gde dužine rečenice variraju, takođe rad se fokusirao samo na klasifikaciju na nivou reči, bez pokušaja da razviju model za predikciju kompletnih rečenica, što ograničava praktičnu primenu.

M. Wand, J. Koutnik i J. Schmidhuber u svom radu [29] prvi uvode i eksperimentišu sa LSTM rekurentnim neuronskim mrežama u oblasti čitanja sa usana ali se ne bave ni podacima na nivou rečenica ni mogućnosti da se klasifikuje govor bez obzira na to ko je govornik. Drugim rečima njihov model nije sposoban da pravilno radi sa govorom različitih govornika, bez potrebe da bude specifično treniran za svakog pojedinačnog govornika.

U radu *Lip Reading using CNN and LSTM* [30] korišćenjem VGG neuralne mreže koja je prethodno istrenirana na slikama lica poznatih ličnosti (IMDB i Google Images), autori su uspeali da razviju model za klasifikaciju reči i fraza koristeći MIRACL-VC1 bazu podataka, koja sadrži samo 10 reči i 10 fraza. Najbolji model u njihovom radu postignut je metodom "zamrzavanja" parametara VGG mreže, a

treniran je samo RNN deo modela. Uprkos tome, njihov najbolji model je dostigao CER od samo 0.56 u klasifikaciji reči i WER 0.44 iako su oba seta podataka sadržala samo 10 mogućih klasa za klasifikaciju.

LipNet [31] model je predložen kao prvi end-to-end model koji koristi kombinaciju konvolucionih neuronskih mreža i rekurentnih neuronskih mreža, zajedno s povezanim vremenskim skupovima (CTC), kako bi prepoznao celokupne rečenice bez potrebe za prethodnom segmentacijom na nivou fonema ili reči. Model uči direktno iz video snimaka bez ručnog anotiranja ili korišćenja audio signala, čime eliminiira potrebu za dodatnom obradom podataka. Pored toga, CNN se koristi za ekstrakciju prostornih karakteristika iz video sekvenci usana, dok RNN obrađuje vremenske informacije, algoritam koristi CTC funkciju gubitka što omogućava modelu da predviđa sekvence različitih dužina bez preciznog mapiranja između ulaznih i izlaznih podataka.

End-to-end modeli eliminišu potrebu za segmentacijom video snimka na pojedinačne reči pre predikcije cele rečenice. LipNet takođe eliminiše potrebu za ručnim podešavanjem prostorno-vremenskih vizuelnih karakteristika i za odvojenim treniranjem modela za rečenice. Ovaj model predstavlja veliki empirijski uspeh i veruje se da će se performanse dalje unapređivati kako bude dostupno više podataka.

LipNet je testiran na GRID korpus skupu podataka. Rezultati su pokazali značajno poboljšanje u odnosu na prethodne modele, postižući visoku tačnost u prepoznavanju rečenica. Ovaj model je postigao bolju preciznost u poređenju sa metodama koje se oslanjaju na ručnu segmentaciju ili na kombinaciju zvučnih i vizuelnih signala. Glavni doprinos ovog rada je što je pokazao kako je moguće efikasno primeniti duboke neuronske mreže za prepoznavanje govora na nivou rečenice isključivo na osnovu vizuelnih informacija

*SyncVSR* [32] rad predstavlja značajan napredak u oblasti vizuelnog prepoznavanja govora, naročito u klasifikaciji vizema. U osnovi, metoda se oslanja na naprednu tehniku sinhronizacije audio tokena sa video snimcima, poznatu kao *Crossmodal Audio Token Synchronization*.

Na nivou reči, korišćena je funkcija gubitka specifična za klasifikaciju reči (engl. *Word Classification Loss*), dok se na nivou rečenica koristi kombinovana CTC-Attention funkcija gubitka (engl. *Character Classification Loss*). Konkretno, međuentropijska funkcija gubitka meri razliku između predviđenih i stvarnih oznaka klasa na nivou reči. Na nivou rečenica, koristi se kombinacija CTC funkcije gubitka za enkoder i funkcije gubitka zasnovane na jezičkom modelovanju za dekode, što je poznato kao spojena CTC-Attention funkcija gubitka. Ova metoda omogućava preciznije nadgledanje na nivou frejma, što znatno poboljšava tačnost modela i omogućava mu da postigne vrhunske rezultate. Za trening modela korišćena je baza podataka koja je već korišćena u radu "*Lip Reading in the Wild*" za engleski jezik, kao i CAS-VSR-W1K baza podataka za kineski jezik. Ove baze podataka uključuju preko 600 sati snimaka, što omogućava temeljnu obuku modela na nivou reči. Arhitektura modela za vizuelno prepoznavanje govora na nivou reči (*word-level VSR*) obuhvata kombinaciju 3D CNN-a, ResNet18 i transformera.

Za analizu na nivou rečenica (*sentence-level VSR*), koristi se *Conformer* model, koji spaja karakteristike konvolucionih i transformacionih mreža, pružajući bogatiji kontekst i efikasnije procesiranje sekvencijalnih podataka. Ova arhitektura se oslanja na prethodne radove u oblasti mašinskog učenja, koji su omogućili razvoj naprednih modela za vizuelno prepoznavanje govora.

*SyncVSR* koristi i spoljni model koji se naziva *informacija o granicama reči*, odnosi se na uključivanje eksplicitnih podataka o tome kada svaka reč počinje i završava u audio ili video sekvenci. Ova informacija se često dobija pomoću tehnika nametnutog poravnanja, koje mapiraju audio na tekst na nivou fonema.

*SyncVSR* takođe koristi i *spoljni jezički model* je dodatni model koji se koristi kako bi poboljšao prepoznavanje govora, posebno na nivou rečenica. Dok osnovni model prepoznaje zvuke ili slike reči na osnovu video ili audio podataka, spoljni jezički model (obično treniran na velikim korpusima teksta) pomaže da prepoznati niz reči ima smisla u jezičkom kontekstu. Na primer, spoljni jezički model može ispraviti greške u prepoznavanju reči na osnovu verovatnoće njihove pojave u određenim gramatičkim

strukturama ili kontekstima. U prepoznavanju govora, to znači da model može "pretpostaviti" koje reči slede u rečenici na osnovu jezičkih pravila i statistike, što može značajno smanjiti stopu greške na nivou reči (WER).

SyncVSR postiže vrhunske rezultate na nivou reči za engleski i kineski jezik, predstavljajući trenutno „*state of the art*“ u ovoj oblasti. Na *LRW* bazi podataka, model ostvaruje od 93.2 %, a dodatno poboljšava rezultat na 95 % kada koristi informacije o granicama reči. Na *CAS-VSR-W1k* bazi podataka, koja je sa kineskog govornog područja testira model, dostiže tačnost od 58.2% što nadmašuje prethodne modele.. U zadacima prepoznavanja rečenica, model je testiran na bazama podataka pod nazivima *LRS2* i *LRS3* koje su na engleskom jeziku. Model postiže WER od 22% na *LRS2* i 30.5 % na *LRS3* bez korišćenja spoljnog jezičkog modela. Sa korišćenjem jezičkog modela daje još bolje rezultate - 20% na *LRS2* i 28.1 na *LRS3*.

Pored toga, SyncVSR bolje obrađuje homofone, zahvaljujući nekoliko ključnih faktora. Prvi faktor je potpuna sinhronizacija sekvenci sa audio tokenima. Zatim, korišćenje audio rekonstrukcije koja implementira audio rekonstrukcije kao deo treninga modela. To znači da model uči kako da rekonstruiše zvuk na osnovu vizuelnih podataka. Gubitak tokom rekonstrukcije se koristi kao deo funkcije gubitka kako bi se osigurala tačnija rekonstrukcija zvuka iz vizuelnih podataka. Zahvaljujući ovim faktorima, SyncVR je u stanju da preciznije razlikuje vizuelno slične reči u različitim kontekstima, što značajno poboljšava njegovu klasifikacionu sposobnost. Ovo je posebno važno u okruženjima sa visokim nivoom buke.

## 4 LipReading - principi rada

U četvrtom poglavlju detaljno je predstavljena praktična implementacija diplomskog rada. Razvijena je originalna arhitektura neuronske mreže, inspirisana najnovijim istraživanjima u oblasti vizuelnog prepoznavanja govora. Ovaj deo rada uključuje precizno definisane parametre mreže, uključujući veličinu ulaznog sloja, vrste korišćenih neuronskih mreža, broj neurona u svakom sloju, kao i tehnike regularizacije koje su implementirane radi sprečavanja prekomernog prilagođavanja modela. Obrada podataka za ulaz u mrežu i izbor funkcije gubitaka nisu samostalno razvijeni u okviru ovog rada.

Pored toga, razvijen je i implementiran algoritam za obradu podataka u realnom vremenu. Ovaj algoritam omogućava detekciju lica na video snimcima, izdvajanje regije oko usana, praćenje početka i kraja govora, kao i pokretanje procesa obrade i predikcije na osnovu zabeleženih podataka. Iako postoje slične arhitekture za obradu podataka i prepoznavanje govora u realnom vremenu, u ovom radu je celokupna ideja i implementacija ovog dela projekta izvedena samostalno.

### 4.1 Programsko okruženje i korišćene biblioteke

Ceo sistem razvijen je korišćenjem programskog jezika Python v3.9 u okruženju Julia. Za obradu i procesiranje podataka korišćene su biblioteke opencv-python v4.9.0.80 i numpy v1.26.2, dok je za implementaciju neuronskih mreža i učenje modela korišćen TensorFlow v2.16.1. Vizualizacija rezultata i evaluacija modela obavljene su pomoću biblioteka matplotlib v3.8.2 i seaborn v0.13.8, kao i ugrađenih TensorFlow funkcija za procenu performansi modela.

### 4.2 Baza podataka – GRID Korpus

GRID korpus [33] predstavlja obimnu bazu podataka koja sadrži audio-vizuelne rečenice izgovorene od strane 34 govornika (18 muškaraca i 16 žena) na engleskom jeziku. Svaki govornik izgovara 1.000 rečenica, a baza sadrži ukupno 34.000 audio i video snimaka. Rečenice u GRID korpusu prate specifičnu strukturu, gde svaka rečenica ima sledeći šablon: komanda (4 opcije) + boja (4 opcije) + predlog (4 opcije) + slovo (25 opcija) + broj (10 opcija) + pridev (4 opcije). Ovaj pristup omogućava generisanje do 64.000 jedinstvenih rečenica, kao što su: „set blue by four please“ ili „place red at C zero again“. U ovom radu korišćen je podskup podataka koji obuhvata video zapise prvog govornika (s1), uzeto je prvih 500 pojedinačnih video snimaka. Primer jednog frejma govornika prikazan je na slici 4.2.1:



slika 4.2.1. govornik s1, video 6, frejm 9

### 4.3 Procesiranje podataka i parametri obuke modela

Proces procesiranja podataka započinje analizom svakog pojedinačnog frejma video snimka, gde je potrebno identifikovati lice na snimku. Za detekciju lica korišćena je biblioteka *dlib* [34]. Nakon detekcije lica (u slučaju GRID korpusa, na svakom snimku je prisutno samo jedno lice), određuju se tačke koje definišu usne [35], a zatim se izrezuje regija dimenzija 64x64 piksela oko usana. Svaki frejm se zatim konvertuje u sivu skalu. Potom se grupišu u listu koja se potom transformiše u tenzor, standardizuje, i na kraju priprema za ulaz u neuronsku mrežu. Primer jednog frejma uzorka koji je pripremljen za neuronsku mrežu je prikazan na slici 4.3.1:

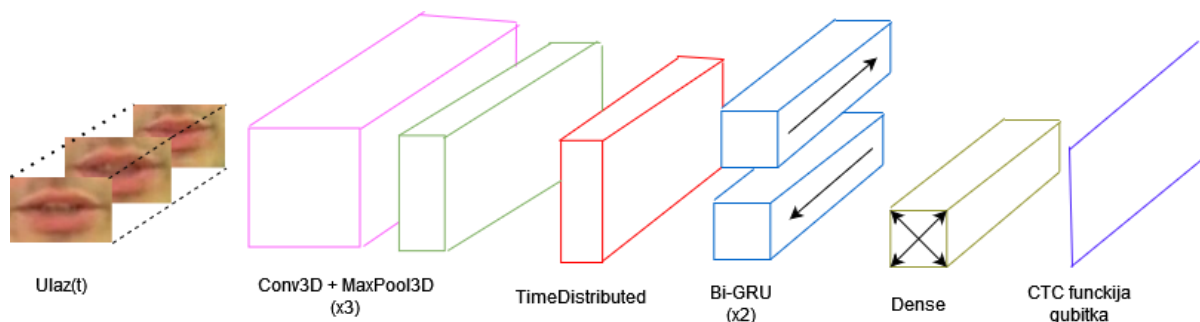


slika 4.3.1. primer frejma pripremljenog za ulaz u neuronsku mrežu

Svaki karakter koji se može pojaviti u video snimku mora biti označen jedinstvenim brojem. U ovom slučaju, ukupno postoji 40 mogućih karaktera, uključujući i prazan simbol. Ovaj zadatak se realizuje korišćenjem odgovarajuće funkcionalnosti iz biblioteke za mašinsko učenje, koja omogućava definisanje vokabulara za razbijanje na tokene, gde je svaki token numerički kodiran prema odgovarajućem slovu i uključivanje specijalnog simbola za nepoznate karaktere. Ovaj korak je ključan jer neuralne mreže operišu sa numeričkim vrednostima i ne mogu direktno koristiti tekstualne podatke. Nakon što neuronska mreža generiše listu brojeva kao predikcije, ti brojevi se dekoduju nazad u tekstualni format.

Pre nego što se procesirani podaci proslede modelu, potrebno ih je organizovati u odgovarajući skup podataka. Ovaj proces započinje identifikovanjem putanje do foldera koji sadrži video snimke. Nakon toga, podaci (video snimci) se nasumično mešaju kako bi se smanjila mogućnost prekomernog prilagođavanja modela. Zatim se podaci usklađuju sa odgovarajućim oznakama za svaki frejm.

Podaci se grupišu u batch veličine 2, pri čemu se vrši proširenje (engl. *padding*) svakog *batch*-a na dimenzije ([75, 64, 64], [40]), gde je 75 dužina video snimka, 64x64 se odnosi na širinu i visinu frejmova, dok je 40 broj mogućih karaktera koji može da odgovara svakom frejmu. *Prefetching* se koristi za optimizaciju performansi, tako što se podaci učitavaju unapred asinhrono. Podaci se potom dele na trening i test skupove: prvih 450 video snimaka koristi se za treniranje modela, dok se preostalih 50 video snimaka koristi za validaciju algoritma.



slika 4.3.2 Arhitektura neuralne mreže

Kao što je prikazano na slici 4.3.2, arhitektura modela se sastoji od tri konvoluciona sloja Conv3D, koji služe za ekstrakciju prostorno-vremenskih obeležja iz ulaznog video zapisa. Ulazni sloj ima dimenzije  $75 \times 64 \times 64 \times 1$ , što odgovara dimenzijama svakog video snimka u bazi podataka. Svi konvolucionni slojevi koriste ReLU funkciju aktivacije, detaljno opisanu u poglavlju 2. Na kraj svakog sloja dodaje se i MaxPool3D sloj, čija je uloga da smanji osetljivost mreže na male promene u okolini, čineći je otpornijom na šum.

Nakon konvolucionih slojeva, koristi se sloj koji pretvara svaki trodimenzionalni izlaz iz konvolucionih slojeva u jednodimenzionalni tenzor, što znači da svaki frejm (sa svojim prostornim karakteristikama) postaje jednodimenzionalni tenzor. Omogućena je pomenuta vektorizacija u jednodimenzionalni tenzor na svaki frejm nezavisno, tretirajući svaki frejm kao zaseban uzorak, nakon čega se sekvenca prosleđuje u rekurentni sloj kao niz vektora. Razlog vektorizacije je taj što RNN, GRU ili LSTM slojevi očekuju sekvencijalne podatke kao ulaz, obično u formi jednodimenzionalnih tenzora. Ovo omogućava modelu da uči sekvencijalne odnose između frejmova, što je ključno za zadatke poput vizuelnog prepoznavanja govora.

Nakon sloja za vektorizaciju, u modelu se nalaze dva sloja *Bidirectional GRU* (dvosmerna rekurentna mreža sa *Gated Recurrent Unit* - GRU). Slojevi koriste 128 jedinica, a inicijalizacija težina se vrši pomoću *Orthogonal* inicijalizatora. Ovaj inicijalizator postavlja težine tako da su matrice ortogonalne, što doprinosi stabilnosti tokom treninga, posebno kod rekurentnih mreža poput GRU. Na taj način, *Orthogonal* inicijalizacija pomaže u očuvanju informacija kroz više vremenskih koraka i smanjuje rizik od problema poput eksplodirajućeg ili nestajućeg gradijenta, čime se poboljšava efikasnost učenja modela. Omogućeno je da GRU sloj vrati čitavu sekvencu izlaznih vrednosti za svaki vremenski korak, umesto samo poslednjeg izlaza. Ovo je ključno kada sledeći sloj treba da obradi kompletnu sekvencu podataka, kao što je slučaj u modelima koji koriste slojeve za obradu sekvencijalnih informacija. *Dropout* [36] (sa parametrom 0.5) predstavlja tehniku regularizacije koja smanjuje rizik od nadprilagođavanja (engl. *overfitting*) modela. Postavljanjem vrednosti na 0.5, svaki neuron ima 50% šanse da bude isključen tokom treninga, čime se smanjuje prevelika zavisnost od pojedinih neurona i poboljšava sposobnost modela da generalizuje na novim podacima. Aktivaciona funkcija je takođe ReLU.

Na kraju mreže postavlja se potpuno povezani sloj (engl. *fully connected layer*) koji ima dimenziju jednaku veličini vokabulara (39), plus jedan dodatni izlaz za neviđene podatke. Ovaj sloj koristi *He-Normal* inicijalizaciju za težine, ova tehnika inicijalizacije postavlja početne težine slojeva na vrednosti izvučene iz normalne raspodele sa srednjom vrednošću nula i standardnom devijacijom koja zavisi od broja ulaza u taj sloj, i na taj način se održava kvalitetan gradijent tokom obuke. Kao funkcija aktivacije koristi se *softmax*, koja normalizuje izlazne vrednosti tako da predstavljaju procentualne verovatnoće za svaki mogući karakter. Izlaz s najvećom verovatnoćom označava koji je karakter najverovatniji u transkriptu.

Za obuku neuralne mreže koristi se CTC funkcija gubitka, detaljno opisana u poglavlju 2. Kako bi se proces treninga optimalno pratio, implementirane su različite *callback* funkcije. Među njima je *scheduler*, koji održava konstantnu stopu učenja tokom prvih 30 epoha, nakon čega eksponencijalno opada, čime se mreži omogućava bolja konvergencija.

Za optimizaciju modela koristi se Adam optimizator, poznat po svojoj efikasnosti i robusnosti u učenju na osnovu gradijenata, gradijent u početnom trenutku ima podešenu vrednost 0.0001. Obuka neuralne mreže obavlja se kroz 60 epoha, tokom kojih se mreža postepeno prilagođava kako bi ostvarila što bolje performanse na zadatku bez natprilagođavanja.

### 4.3 LipReadingWebCam – princip rada

Za rukovanje podacima sa kamere koristi se niz funkcija koje zajednički obavljaju ključne zadatke u okviru procesa vizuelnog prepoznavanja govora. Ove funkcije uključuju aktiviranje kamere, detekciju lica u svakom kadru, izolaciju područja usana, praćenje govora, i na kraju, konvertovanje podataka u tenzor nakon završetka govora. Svi ovi podaci se zatim obrađuju za prosleđivanje neuronskoj mreži, koja vrši konačnu predikciju govornog sadržaja.

Prva funkcionalnost koja je implementirana u prikupljanju podataka u realnom vremenu jeste da obrađuje frejmove video zapisa kada je detektovan govor putem kamere i vraća tenzor sa frejmovima spreman za predikciju u neuronskoj mreži. Svaki frejm je dimenzije 64x64 piksela. Ako je niz frejmova kraći od 75, funkcija dodaje (engl. *padding*) frejmove kako bi dostigla dužinu od 75 frejmova. U slučaju da je govor duži od 75 frejmova, funkcija zaustavlja snimanje nakon 75 frejmova i koristi samo prvih 75 za dalju obradu. Time se osigurava konzistentna dužina ulaza za neuronsku mrežu.

Zatim se primenjuje posebna funkcija koja ima ulogu za vizualnu reprezentaciju na kameri, prikazujući da je algoritam uspešno detektovao lice i izdvojio regiju usana. U sebi sadrži algoritam za detekciju govora ili tišine, koji procenjuje da li osoba govori na osnovu razdaljine između gornje i donje usne. Detekcija govora se vrši kada razdaljina između usana premašuje vrednost od 19.5 piksela, što zavisi od različitih faktora kao što su udaljenost između kamere i lica, kao i karakteristike same kamere. Kako bi se detektovao govor, koristi se matematička funkcija koja računa euklidsko rastojanje između tačaka koje predstavljaju gornju i donju usnu.

U funkciji se takođe nalaze i dva flag-a. Prvi flag označava da li osoba trenutno govori, dok drugi flag čuva vremensku oznaku poslednje detekcije govora, kako bi se upravljalo prelazima između govora i tišine. Ukoliko je prošlo manje od 0.5 sekunde od aktivacije prvog flaga i ponovo je detektovan flag za govor zatvaranje usta se ne računa kao prekid govora već kao izgovaranje fonema /m/ /b/ ili /p/ gde je potrebno spojiti usne pri izgovoru.

Glavna funkcija sadrži *while* petlju koja kontinuirano snima frejmove sa kamere u realnom vremenu. Ključni deo funkcije je detekcija opadajuće ivice. Ova promena označava kraj govora. Kada se detektuje kraj govora, funkcija pokreće proces pripreme frejmova, svi frejmovi se pretvaraju u sivu skalu, koji se zatim prosleđuju neuronskoj mreži radi predikcije. Nakon što neuronska mreža predvidi izgovoreni tekst, rezultat se ispisuje u donjem delu prozora kamere, omogućavajući korisniku da vidi prepoznat govor u realnom vremenu.



## 5 Rezultati

Da bi se izračunale performanse LipReading neuralne mreže, računaju se *WER* i *CER* metrike, standardne metrike za izračunavanje performanse VPG modela. Metrike su opisane i uvedene u rad u poglavlju 3.

Model je treniran na 450 video snimaka. Nakon 60 epoha treninga, testiranje modela izvršeno je na podacima iz iste baze podataka, konkretno na 50 snimaka od pet različitih govornika: s1, s7, s9, s11.

Algoritam je neprekidno smanjivao funkciju gubitka kako na trening skupu, tako i na validacionom skupu, što sugerise da nije došlo do natprilagođavanja. To takođe ukazuje na mogućnost da bi algoritam mogao biti dodatno unapređen kroz više epoha obučavanja, ali zbog ograničenog vremena nije bilo moguće sprovesti dodatne epohe.

### 5.1 Performanse LipReading modela

Model za VPG-a evaluiran je na neviđenim test podacima, testiranje je izvršeno na test skupu koji se sastoji od 50 neviđenih uzoraka - video snimaka, na 5 različitih govornika. Rezultati su predstavljeni u tabeli 5.1.1., gde su date prosečne vrednosti *WER* i *CER* metrike za svih 50 uzoraka svakog govornika.

Govornik	Prosečan WER (%)	Prosečan CER (%)
S1	<b>12.10</b>	<b>7.21</b>
S7	71.04	54.28
S9	68.22	51.09
S11	74.68	58.01

Tabela 5.1.1 WER i CER na različitim govornicima

Model je postigao prosečan WER od 12.10 % i prosečan CER od 7.21 % na govorniku s1, što ukazuje na visok nivo tačnosti u prepoznavanju izgovorenog teksta na ispitaniku na kome je baza i obučavana. Kombinovanje prostorno-vremenskih konvolucionih i Bi-GRU slojeva pokazalo se ključnim za dobijanje ovih rezultata, jer omogućava modelu da efikasno obrađuje kako prostorne tako i vremenske karakteristike iz video podataka.

Pored toga, CTC funkcija gubitka omogućila je modelu da obradi ulazne podatke promenjive dužine kao i da proizvede izlazne podatke promenjive dužine, što predstavlja posebno važan korak u prirodnoj obradi govora.

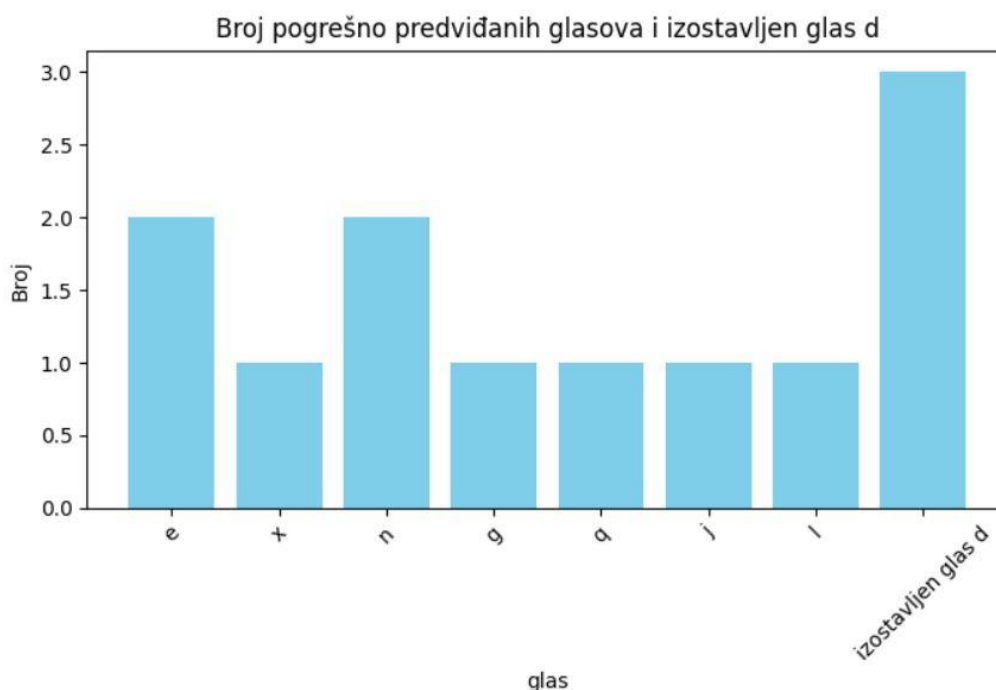
Za ostale govornike, s7, s9, s11 i s16, model pokazuje znatno lošije rezultate, pri čemu nijedan od njih nije ostvario WER manji od 64% niti CER manji od 46%. Ovi rezultati mogu ukazivati na prekomerno prilagođavanje modela podacima prvog govornika (s1). Ipak, verovatniji razlog je to što model nije obučen na dovoljno velikom i raznovrsnom skupu podataka, koji bi uključivao govornike sa različitim karakteristikama usana i različitim brzinama govora. Takođe, 60 epoha treninga verovatno nije bilo dovoljno da model postigne preciznost u predikciji. U daljim istraživanjima na ovu temu, pažnja će biti usmerena na rešavanje ovog problema kroz obuku na većem, raznovrsnijem i robusnijem skupu podataka, kako bi se unapredila tačnost modela.

Model je pravio greške pri klasifikaciji vizema, posebno kod alveolarnih suglasnika. Najviše konfuzije primećeno je kod alveolarnih suglasnika /d/, /n/ /t/ /s/ i /l/, koji se izgovaraju kontaktom jezika i



nadzubne zone. Ova konfuzija je razumljiva jer su ti suglasnici vizuelno gotovo identični kada se lice posmatra спреда.

Na slici 5.1.1 prikazan je ukupan broj izostavljanja glasa /d/ i pogrešno predviđanih glasova koji se pojavljuju u test skupu govornika s1 pri predikciji četvrte pozicije u tekstu, gde se po šablonu baze uvek nalazi jedan fonem. Predviđeni glasovi odnose se na one glasove gde je model umesto odgovarajućeg glasa predvideo neki drugi glas. Izostavljanje se dešavalo kada se glas nalazio neposredno posle reči koja se završava glasom /n/ ili /t/, ili pre reči koja počinje istim glasom, primer su rečenice kao što su „lay blue at d six now” i „place white at d three soon“, gde je model predvideo „lay blue at six now” i „place white at three soon”.



slika 5.1.1 broj pogrešno predviđanih glasova i izostavljen glas /d/ u test skupu(s1)

## 5.2 Perfomansa LipReadingWebCam algoritma

Trenutni problemi sa LipReadingWebCam eksperimentom ukazuju na nekoliko ključnih tehničkih izazova koji utiču na uspešnost modela u realnom vremenu:

Model je obučen na sekvencama konstantne dužine od 75 frejmova, dok su podaci sa kamere promenljive dužine, što zahteva proširivanje kako bi sekvenca bila odgovarajuće dužine. Međutim, proširivanje se vrši belim frejmovima (koji ne sadrže informacije o govoru) i nije nešto što je model naučio tokom obuke. To dovodi do lošijih rezultata, jer mreža ne može pravilno da interpretira podatke iz ovakvih frejmova. Loši uslovi snimanja (neodgovarajuće osvetljenje, visok nivo šuma u pozadini) značajno otežavaju mreži da izvuče relevantne informacije iz frejmova. U ovakvim uslovima, algoritam za detekciju usana nije u stanju da precizno odredi ključne karakteristike za prepoznavanje govora.

## 6 Unapređivanje i dalji razvoj

Model je treniran i testiran na podacima jednog govornika iz baze koja sadrži ukupno 33 govornika. Ovaj pristup pruža uvid u sposobnosti modela, ali i značajno ograničava njegovu mogućnost generalizacije, što se vidi pri testiranju na podacima drugih govornika. Kako bi model postigao bolje rezultate i bio primenjiv u širem opsegu situacija, neophodno je proširiti obuku na većem skupu podataka koji bi obuhvatao različite govornike, njihove varijacije u govoru, kao i različite osvetljenosti i pozicije lica.

Pored toga, performanse modela bi mogle biti poboljšane dodavanjem dodatnih slojeva u neuronsku mrežu ili modifikacijom parametara postojećih slojeva. Korišćenjem slojeva sa većim brojem jedinica, efikasnijim metodama regularizacije ili optimizacijom hiperparametara, model bi mogao bolje da generalizuje na neviđene situacije. Uzimajući u obzir da rekurentne neuronske mreže (RNN) pokazuju određena ograničenja u prepoznavanju sekvencijalnih podataka, razmatra se primena modela zasnovanih na transformerima. Ovi modeli se zasnivaju na mehanizmu tzv. samopažnje (*self-attention*), koji omogućava paralelno procesiranje sekvenci i bolju analizu dugoročnih zavisnosti u podacima. Zbog toga bi mogli generisati kvalitetnije izlaze, naročito u složenijim scenarijima sa višestrukim govornicima i bučnim okruženjima.

Proširivanje frejma koji ulazi u neuronsku mrežu kako bi obuhvatila regiju oko očiju i obrva predstavlja dodatni korak koji bi mogao značajno unaprediti performanse modela. Ova regija lica često pokazuje aktivnost pri izgovoru određenih slova i rečenica, što može pružiti korisne informacije za prepoznavanje govora. Uključivanjem ovih dodatnih područja u analizu, model bi mogao bolje uhvatiti suptilne facijalne izraze povezane s fonemima i rečima, što može poboljšati preciznost i efikasnost sistema za vizuelno prepoznavanje govora. [37]

Uvođenje transformera u arhitekturu modela zahteva potpunu rekonstrukciju neuronske mreže, ali bi moglo značajno poboljšati preciznost predikcija. Pored vizuelnih podataka, u budućnosti se može razmatrati i integracija audio signala kao dodatnog ulaza u model. Ovakav multimodalni pristup omogućio bi modelu da obrađuje video snimke u čijoj audio komponenti se javljaju buka, višestruki govornici, ili lica koja nisu optimalno pozicionirana prema kameri. Ovaj pristup mogao bi znatno unaprediti performanse modela u realnim scenarijima kao što su prepoznavanje govora u bučnim okruženjima ili u situacijama sa niskom rezolucijom video signala.

Kombinacija ovih unapređenja učinila bi model robusnijim i efikasnijim za širok spektar primena, pružajući stabilnije rezultate u zahtevnijim uslovima, što bi u velikoj meri olakšalo njegovu implementaciju u sisteme za prepoznavanje govora u realnom vremenu.

U narednom periodu planira se rad na unapređenju algoritma predikcije sa kamere u realnom vremenu kroz sledeće korake:

- Obuka mreže na podacima koji uključuju frejmove bez informacija (npr. sa belim frejmovima), kako bi model bio robusniji na veštačko dopunjavanje sekvencija.
- Poboljšanje kvaliteta video snimaka i detekcije lica kroz korišćenje boljih kamera i uslova snimanja, što će smanjiti šum i obezbediti preciznije frejmove.
- Optimizacija algoritma kako bi se smanjila latencija i ubrzalo procesiranje podataka u realnom vremenu.

## 7 Zaključak

Glavni cilj diplomskog rada bio je da opiše način funkcionisanja neuronskih mreža u oblasti vizuelnog prepoznavanja govora, da pruži pregled postojeće literature koja se odnosi na automatsko učenje bez primene neuronskih mreža, kao i na savremene modele koji se baziraju na neuronskim mrežama i principima dubokog učenja. Nakon toga, osmišljena je i arhitektura neuralne mreže koja je trenirana i evaluirana na javno dostupnoj bazi podataka, kako bi se praktično prikazalo rešavanje problema. Model je pokazao obećavajuće rezultate na snimcima iz kontrolisanih uslova.

U radu je treniran i evaluiran model koji primenjuje duboko učenje za *end-to-end* pristup, mapirajući sekvence slika govornikovih usana u cele rečenice. Ovaj novi pristup u oblasti automatskog prepoznavanja govora eliminiše potrebu za segmentacijom video zapisa na reči pre predikcije cele rečenice. Empirijska evaluacija je ilustrovala značaj ekstrakcije prostorno-vremenskih karakteristika i efikasne vremenske agregacije podataka.

Vizuelno prepoznavanje govora je jedna od oblasti koja privlači veliku pažnju i interesovanje istraživača, smatra se jednim od dugotrajnih problema u oblasti veštačke inteligencije. Sistemi VPG su sve prisutniji na mobilnim uređajima, desktop računarima, kao i u obliku virtuelnih asistenata. Ipak, ova tehnologija se i dalje suočava sa brojnim izazovima i problemima koje istraživači nastoje da reše.

Tačnost vizuelnog prepoznavanja govora u određenim uslovima mogla bi biti poboljšana primenom naprednijih tehnika obrade slike i video signala, što bi smanjilo šum i ometanje u videu i povećalo preciznost prepoznavanja. Povećanje broja podataka za obuku, koji uključuju različite uslove snimanja, takođe bi doprinelo boljim performansama sistema u sličnim okruženjima.

Za postizanje tačnosti sistema koja bi mogla nadmašiti ljudske sposobnosti u svim uslovima neophodne su nove tehnike i paradigme u vizuelnom modelovanju. Budući ASR sistemi će morati da funkcionišu kao dinamički sistemi, integrišući komponente sa povratnim informacijama koje kontinuirano vrše predikcije i prilagođavanja. Ovi sistemi će morati da prepoznaju više govornika u mešovitim okruženjima, razdvoje govor od šuma i prate jednog govornika dok ignorišu druge. Ova kognitivna sposobnost pažnje, koju ljudi prirodno poseduju, trenutno izostaje u modernim sistemima automatske detekcije govora. Budući ASR sistemi, u koje spadaju i VPG sistemi će morati da prepoznaju ključne karakteristike govora iz dostupnih podataka i da se dobro generalizuju na nove govornike, pored toga sistemi će morati da prepoznaju ključne karakteristike govora iz dostupnih podataka i da dobro generalizuju na nove govornike.

## Reference

- [1] „[Deafness and hearing loss](#)”, [World Health Organisation](#)
- [2] „[Deafness and hearing loss problem](#)”, [World Health Organisation](#)
- [3] „*Perceptual dominance during lipreading*”, Randolph D. Easton, Marylu Basala, 1982
- [4] „Confusions among visually perceived consonants”, *Journal of Speech & Hearing Research*, pp. 796-804 11(4), Fisher, 1968 ”
- [5] „[Computer Vision Lip Reading 2.0](#)”, Allen Ye, Sharon Kwong, 2022
- [6] „The Perceptron: A Perceiving and Recognizing Automation”, F Rosenblatt, Cornell Aeronautical Laboratory, 1957.
- [7] „*Deep Learning*“, Goodfellow, I., Bengio, Y., & Courville, A. , MIT Press.,2016
- [8] „*The Neural Network Input - Process - Output Mechanism*“, James McCaffrey, 2013
- [9] „*Activation Functions in Artificial Neural Networks: A Systematic Overview*“, pp 2-3. Johannes Lederer, 2021
- [10] „*Review and Comparison of Commonly Used Activation Functions for Deep Neural Networks*“, Tomasz Szandała.
- [11] "Backpropagation through time: what it does and how to do it," pp. 1550-1560, in *Proceedings of the IEEE*, P.J.Werbos, 1990
- [12] „*Large-Scale Machine Learning with Stochastic Gradient Descent*“, pp 177-186, Leon Bottou, 2010
- [13] „*Mašinsko učenje Scikit-learn i Pytorch*” pp. 451-463, Sebastian Raschka, Yuxi Liu, Vahid Mirjalili, 2022.
- [14] „*Spatio-Temporal Convolution-Attention Video Network*”, Ali Diba, Mohammad M Arzani, Luc Van Gool, ETH Zurich, 2023
- [15] „*Mašinsko učenje Scikit-learn i Pytorch*” pp. 499-508, Sebastian Raschka, Yuxi Liu, Vahid Mirjalili, 2022.
- [17] „*On the difficulty of training recurrent neural networks*”, R. Pascanu, T. Mikolov I Y. Bengio, 2012.
- [18] „*Long Short-Term Memory*” S. Hocheiter, J. Schmidhuber, *Neural Computation* 9(8), 1735 – 1780, 1997
- [19] „*An Empirical Exploration of RNN Architectures*”, Rafal Jozefowicz, Wojciech Zaremba i Ilya Sutskever, *Proceedings of ICM*, 2342 – 2350, 2015
- [20] „*Mašinsko učenje Scikit-learn i Pytorch*” pp. 512, Sebastian Raschka, Yuxi Liu, Vahid Mirjalili, 2022.
- [21] „*Empirical evaluation of gated recurrent neural networks on sequence modeling*”. arXiv preprint arXiv:1412.3555, J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014.
- [22] „[Fully Connected Layer Vs Convolutional Layer](#)” Diego Unzeuta, 2022.

- [23] „*Connectionist temporal classification: labeling unsegmented sequence data with recurrent neural networks*“, A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber, In ICML, pp. 369–376, 2006
- [24] [Papers With Code - Lipreading Databases – \[https://paperswithcode.com/task/lipreading\]](https://paperswithcode.com/task/lipreading)
- [25] „*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*“, pp 45-47, Dan Jurafsky i James H. Martin, 2000
- [26] „*Continuous automatic speech recognition by lipreading. In Motion-Based recognition*” pp. 321–343, A. J. Goldschen, O. N. Garcia, and E. D. Petajan. C Springer, 1997
- [27] „*Audio Visual Speech Recognition*” , Technical report, Neti, G. Potamianos, J. Luetttin, I. Matthews, H. GLotin, D. Vergyri, J. Sison i A. Mashari, IDIAP, 2000.
- [28] „*Lip Reading In the Wild*” , Joon Son Chung and Andrew Zisserman, Visual Geometry Group, Department of Engineering Science, University of Oxford
- [29] „*Lipreading with Long Short-Term Memory*”, pp. 6115–6119, 2016, International Conference on Acoustics, Speech and Signal Processing, M. Wand, J. Koutnik, and J. Schmidhuber
- [30] „*Lip Reaing using CNN and LSTM*” Amit Garg, Jonathan Noyola, Sameep Bagadia
- [31] „*LIPNET: END-TO-END SENTENCE-LEVEL LIPREADING*” Yannis M. Assael1, Brendan Shillingford, Shimon Whiteson, Nando de Freitas, Department of Computer Science, University of Oxford, Oxford, UK 1
- [32] „*SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End crossmodal Audio Token Synchronization*” , Young Jin Ahn1, Jungwoo Park, Sangha Park, Jonghyun Choi, Kee-Eung Kim
- [33] [The Grid AudioVisual sentence corpus – \[https://spandh.dcs.shef.ac.uk/gridcorpus/\]](https://spandh.dcs.shef.ac.uk/gridcorpus/)
- [34] „*Dlib-ml - A machine learning toolkit*”, pp. 1755–1758, D. E. King., JMLR., 2009.
- [35] „*300 faces in-the-wild challenge: The first facial landmark localization challenge*” , pp. 397–403, 2013. In IEEE International Conference on Computer Vision C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic.
- [36] „*Dropout: A simple way to Prevent Neural Networks from Overfitting*”, pp. 1929-1958, Journal of Machine Learning Research 15.1 , N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, 2014
- [37] „*Improving Lip Reading by Integrating Visual and Audio Features Zhang*”, pp. 2633-2641, .In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , Zhang, Q 2017

## Biografija

Božoki Dragomir je rođen 9. juna 2001. godine u Novom Sadu. Osnovno obrazovanje stekao je u Budisavi, dok je srednjoškolsko obrazovanje završio u Medicinskoj školi „7. april“ u Novom Sadu. Nakon završene srednje škole upisao je smer Biomedicinsko inženjerstvo na Fakultetu tehničkih nauka u Novom Sadu, gde je studirao od 2020. do 2024. godine.

Tokom studija je aktivno učestvovao u brojnim Erasmus programima za razmenu studenata, a takođe je i alumnista programa BOLD akademske razmene u Sjedinjenim Američkim Državama. Ovaj program je bio sponzorisan i finansiran od strane američke ambasade u Beogradu i fokusirao se na ekonomski razvoj sa ciljem kreiranja projekata i tech startup ekosistema koji bi doprinosili razvoju ekonomije Republike Srbije.