Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 7.05:62
DOI: https://doi.org/10.24867/11BE18Vujovic

POREĐENJE SISTEMA ZA SINTEZU EKSPRESIVNOG GOVORA SA MOGUĆNOŠĆU KONTROLE JAČINE EMOCIJE

COMPARISON OF EXPRESSIVE SPEECH SYNTHESIS SYSTEMS WITH THE POSSIBILITY OF EMOTION-STRENGTH ADJUSTMENT

Mia Vujović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U sintezi ekspresivnog govora važno je generisati emocionalno obojen govor koji odražava kompleksnost emocionalnih stanja. Brojni TTS sistemi emocije u sintetizovanom govoru modeluju u vidu diskretnih skupova, ali tek kada se uzmu u obzir i varijacije koje postoje unutar emotivnih stanja, generisani govor može biti nalik ljudskom. Ovaj rad obuhvata teorijsku analizu i poređenje dva inovativna sistema za sintezu ekspresivnog govora koji kompleksnost emocija modeluju u vidu kontinualnih vektora kojima je moguće manipulisati. Rezultati pokazuju da je pristup zasnovan na t-SNE embedding vektorima primjenljiv samo u slučaju specifičnih baza podataka, dok je drugi pristup, zasnovan na interpolaciji tačaka u embedding prostoru multi-speaker, multi-style modela, opštiji, ali zahtijeva dodatnu analizu.*

Ključne riječi: *ekspresivna sinteza govora, modelovanje emocija, embedding vektori, duboke neuronske mreže*

Abstract – *In expressive speech synthesis, it is important to generate emotional speech that reflects the complexity of emotional states. Many TTS systems model emotions in discrete codes, but modeling variations within emotional states is crucial for generating human-like speech. The paper presents a theoretical analysis and comparison of two innovative expressive TTS systems that model the complexity of emotion in the form of a continuous vector which can be manipulated. The results show that the approach based on continuous t-SNE embedding vectors is applicable only in the case of specific data bases, while the other approach, based on interpolation of points in the embedding space of a multi-speaker, multi-style model, is more general, but requires additional analysis.*

Keywords: *expressive speech synthesis, emotion modeling, embedding vectors, deep neural netowrks*

1. UVOD

Analiza i modelovanje emocija od velikog je značaja za moderne govorne tehnologije među kojima je i sinteza govora iz teksta (eng. *Text-To-Speech* - TTS). TTS sistem je jedna od ključnih komponenti u ostvarivanju efikasne komunikacije između čovjeka i mašine. Da bi ova komunikacija bila nalik ljudskoj, potrebno je u sintetizovani govor unijeti razne govorne stilove, emocije i druge informacije izvan samog tekstualnog sadržaja.

Ovaj rad se bavi problemom ekspresivne sinteze govora gdje je cilj generisati govor u različitim emotivnim stilovima. Brojni su primjeri sistema koji teže da emocije klasifikuju u diskretne skupove [1-4]. Međutim, kako bi govor generisan TTS sistemom bio u potpunosti prirodan, potrebno je uzeti u obzir i varijacije koje postoje unutar emotivnih stanja i omogućiti intuitivnu kontrolu jačine emocije izražene pri sintetizovanom govoru. Cilj rada je teorijska analiza i poređenje dvije inovativne metode koje teže da modeluju kompleksnost emocija u vidu kontinualnih vektora kojima je moguće manipulisati. U prvom pristupu [5], korišćenjem kontrolnog *embedding* vektora kao dodatnog ulaza LSTM (eng. *Long Short-Term Memory* - LSTM) duboke neuronske mreže (eng. *Deep Neural Network* - DNN), omogućena je kontrola nivoa ekspresivnosti. Predloženi model se sastoji od modula za analizu emocija i modula za sintezu govora.

U okviru ovog rada praktično je realizovan modul za emotivnu analizu. Teorijski opis predloženog modela i praktična realizacija dati su u poglavlju 2. Drugi TTS pristup, proširenje je rada [6]. Doziranje emocije izražene pri govoru ostvareno je interpolacijom tačaka u *embedding* prostoru kreiranom obukom *multi-speaker*, *multi-style* modela. Metod je opisan u poglavlju 3. U poglavlju 4 predstavljeni su rezultati subjektivnog testa na osnovu kojeg su pristupi upoređeni. Nakon toga slijede zaključak i literatura.

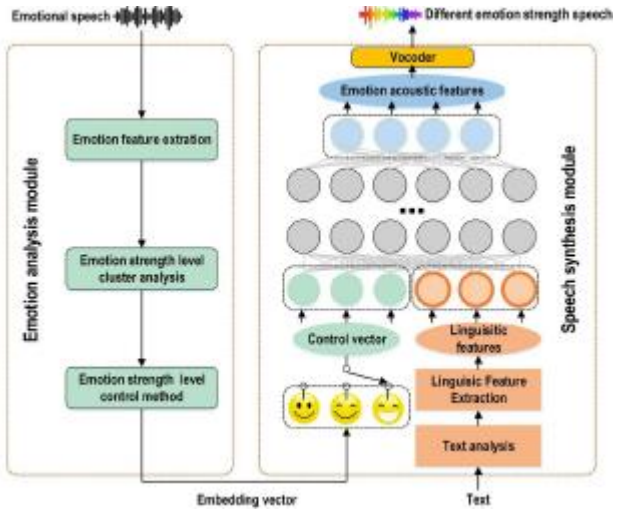
2. T-SNE *EMBEDDING* VEKTORI ZA KONTROLU JAČINE EMOCIJE

Na Slici 1 prikazana je arhitektura TTS modela. Modul za sintezu govora zasniva se na parametarskoj DNN sintezi. Iz ulaznog teksta izdvojena su jezička obilježja koja se preslikavaju u akustička korišćenjem neuronske mreže. Na osnovu izlaznih akustičkih obilježja, WORLD vokoder generiše govorni signal. Neuronska mreža je imala hibridnu arhitekturu sa tri nerekurzivna skrivena sloja praćenu sa 2 LSTM sloja. Kako bi se mogao sintetizovati govor sa različitim nivoom emocije, kao dodatni ulaz LSTM mreže korišćen je kontrolni 2D vektor dobijen na osnovu modula za emotivnu analizu.

Modul za analizu emocija kao ulaz prima bazu podataka sa govorom u željenom emotivnom stilu. Iz baze se izdvajaju akustička obilježja koja se obično koriste u zadacima prepoznavanja emocija. Za to je upotrijebljen *OpenSMILE* (eng. *open-Source Media Interpretation by*

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Vlado Delić, red. prof.

Large feature-space Extraction) softverski alat [7] pomoću kojeg je izdvojen 384-dimenzionalni vektor obilježja korišćen u *INTER-SPEECH 2009* emotivnom izazovu. Statističkom analizom obilježja, prostor uzoraka je podijeljen na klastere korišćenjem *k-means* algoritma klasterizacije. Izabrana su tri klastera za podjelu uzoraka u tri emotivna nivoa, a subjektivnim testovima slušanja određeno je koji klaster predstavlja koji nivo emocije (najmanje izražena, srednje izražena i najizraženija). Kako bi se dobio kontrolni 2D vektor, dimenzionalnost prostora obilježja je smanjena korišćenjem t-SNE algoritma (eng. *t-Distributed Stochastic Neighbor Embedding* – t-SNE) [8]. Klasterizaciona podjela je prikazana na Slici 2a.



Slika 1. *Arhitektura ekspresivnog TTS sistema [5].*

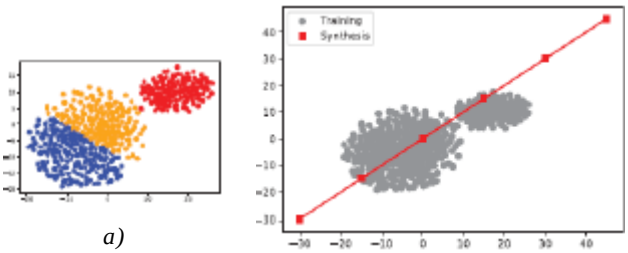
Primijećena je praktično linearna zavisnost t-SNE obilježja koja opisuju uzorke odgovarajućih klastera, tj. visoke vrijednosti obilježja odgovarale su klasteru sa najizraženijom emocijom, a niske vrijednosti klasteru sa najmanje izraženom emocijom.

Imajući to u vidu, uvedena je pretpostavka da se manipulacijom t-SNE vrijednosti može ostvariti kontinualna kontrola jačine emocije. Kako bi se navedena prepostavka ispitala, izabrano je 6 tačaka iz t-SNE prostora na jednakom međusobnom rastojanju duž prave $x = y$ (Slika 2b). Eksperimentalni rezultati dobijeni korišćenjem izabranih vrijednosti kontrolnog vektora kao proširenja ulaza LSTM mreže, pokazali su da povećanjem vrijednosti obilježja, raste i jačina emocije izražene pri sintetizovanom govoru. Za navedeni pristup su korišćene emotivne baze sreće i ljutnje za jednog govornika. Baze su bile na kineskom jeziku i nijesu javno dostupne.

2.1. Praktična realizacija modula za emotivnu analizu

Kako bi se ispitalo da li opisani pristup važi u opštem slučaju, na proizvoljnim bazama ekspresivnog govora, modul za analizu emocija je praktično realizovan i primijenjen na bazama podataka na engleskom jeziku, dva ženska i jednog muškog govornika u tri emotivna stila: srećni, tužni i promotivni.

Iz baza podataka su izdvajena ista obilježja kao i u originalnom radu korišćenjem *OpenSMILE* alata. Za analizu podataka i podjelu uzoraka na emotivne nivoe, korišćena su dva pristupa klasterizacije.



a) *b)*

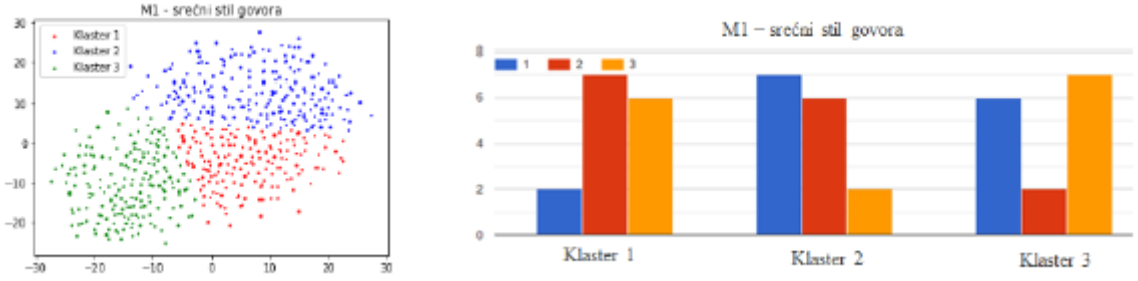
Slika 2. *a) K-means klasterizacija uzoraka emotivne baze srećnog govora; b) Šest embedding vektora korišćenih za testiranje fleksibilnosti modela [5].*

U prvom pristupu, sva informacija o podacima je u potpunosti sačuvana, tj. klasterizacija podataka je rađena u 384-dimenzionalnom prostoru obilježja. Takva raspodjela uzoraka je zatim prikazana u 2D prostoru, jer 384-dimenzionalni prostor nije moguće vizualizovati, a i dodatno, bilo je potrebno dobiti *embedding* vektor za reprezentaciju kontinualne snage emocije. U drugom pristupu, polazilo se od hipoteze da je redukcijom dimenzionalnosti moguće poboljšati performanse sistema usljed kompresije irelevantninh podataka, jer mnoga obilježja mogu biti korelisana. Zbog toga je redukcija dimenzionalnosti t-SNE pristupom izvršena prije postupka klasterizacije. U oba slučaja cilj je bio uočiti pravilnosti između klastera, pa manipulisati vrijednostima t-SNE vektora i time omogućiti kontrolu jačine emocije.

Da bi se odredilo koji od formiranih klastera predstavlja koji nivo emocije, formirani su subjektivni testovi slušanja u kojima je učestvovalo 15 slušalaca. Posmatrane su tri emotivne baze: ženski govornik – emocija sreće, muški govonik - emocija sreće i emocija tuge. Iz svakog klastera posmatranih baza je izdvojeno po 15 audio signala i klastere je trebalo poređati od 1 do 3, tako da 1 označava klaster sa najmanje izraženom emocijom, a 3 klaster sa najizraženijom emocijom.

Rezultati testova su bili prilično ujednačeni, što je i očekivano, jer baze podataka nijesu snimane planski tako da reflektuju različite emotivne nivoe. Ipak, i na osnovu minimalnih razlika definisano je koji klaster predstavlja koji nivo, kako bi se rezultati mogli porediti sa onima iz originalnog rada.

Nijedan od dva klasterizaciona pristupa nije doveo do linearne zavisnosti t-SNE obilježja i jačine emocije kakva je ostvarena u originalnom radu. Tako npr., ako bi opisani pristup važio u opštem slučaju, kod muškog govornika i srećnog stila govora, očekivano bi bilo da klaster 3 predstavlja najmanje izraženu emociju, klaster 1 srednje izraženu, a klaster 2 najizraženiju emociju (Slika 3a). Međutim, ako bi pristup bio vođen testovima slušanja (Slika 3b), zavisnost bi se modelovala suprotno od one opisane u radu, tj. veće t-SNE vrijednosti u ovom slučaju bi značile manje izraženu emociju (klaster 2), a manje vrijednosti izraženiju emociju (klaster 3). I to je primijećeno samo u slučaju ove baze. U drugim primjerima, neko univerzalno pravilo o tome kako se mijenjaju t-SNE vrijednosti u zavisnosti od klastera (nivoa emocije) nije bilo moguće izvesti. Svakako, čak i ako bi pri svakoj govornoj bazi postojala različita, ali jasno uočljiva zavisnost između nivoa emocije i t-SNE vrijednosti, pošto



a) *b)*

Slika 3. *a) K-means klasterizacija uzoraka emotivne baze srećnog govora muškog govornika; b) Rezultati subjektivnih testova za emotivnu bazu srećnog govora muškog govornika.*

ona ne važi u opštem slučaju, uvijek bi je trebalo prethodno testirati, pa tek onda upotrijebiti za sintezu što bi pristup učinilo nepraktičnim.

Uzimajući u obzir činjenicu da subjektivni testovi u radu [5] pokazuju dosljedne rezultate (svi slušaoci su bili usaglašeni oko jačine emocije koju klasteri predstavljaju), pretpostavka je da su baze namjenski snimane tako da se osjete različiti emotivni nivoi, pa su stoga i ostvareni rezultati drugačiji. Svakako, pristup ne važi u opštem slučaju i linearna zavisnost 2D obilježja potrebna za kontinualnu manipluaciju emocijama nije ostvariva u bazama korišćenim u ovom radu.

3. KONTROLA NIVOA IZRAŽENE EMOCIJE INTERPOLACIJOM TAČAKA U *EMBEDDING PROSTORU*

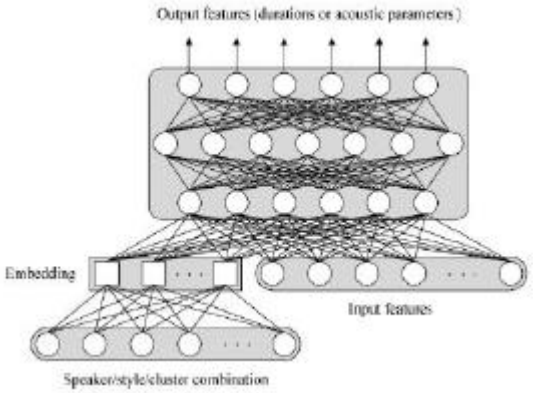
Sinteza govora u određenom govornom stilu i sa određenim glasom, na osnovu male količine ciljnih podataka stila/govornika, predstavljena je u radu [6].

Pristup se zasniva na standaradnoj DNN sintezi koja obuhvata dvije neuronske mreže – mrežu za predviđanje trajanja fonema i mrežu za predviđanje akustičkih obilježja. Da bi se modelovale različite kombinacije govornika i stilova kao dodatni ulaz obje mreže koristio se *N*-dimenzionalni vektor koji je obuhvatao ID govornika, ID stila i ID klastera (dio govornog korpusa dosljedan u pogledu akustičkog i prozodijskog kvaliteta), a svi oni su prethodno predstavljeni u obliku združenog 67-dimenzionalnog *one-hot* vektora, što je ukupan broj kombinacija govornika, stila i klastera (eng. *Speaker-Style-Cluster* - SSC) u govornoj bazi. Na slici 4 prikazana je arhitektura mreže. Mreža je mapirala svaki SSC u prostor niže dimenzije ($N = 15$) tako da je svaka kombinacija predstavljena sa dvije tačke u odgovarajućem *embedding* prostoru – jedna je u vezi sa trajanjem fonema, a druga sa akustičkim obilježjima. Za oba *embedding* prostora, bliskost tačka označavala je njihovu perceptivnu sličnost.

Jednom obučena, ovakva mreža mogla je sintetizovati bilo koji govor predstavljen SSC kombinacijom, samo odabirom odgovarajućih 15-dimenzionalnih *embedding* vektora. Takođe, izborom proizvoljne tačke u *embedding* prostoru, mreža je mogla generisati novi, do tada nepostojeći glas. Upravo u tome se krila mogućnost kontinualne kontrole jačine emocije izražene pri sintetizovanom govoru. Izborom tačke $\boldsymbol{c}$ u *embedding* prostoru između neutralnog govora $\boldsymbol{x}$ jednog govornika i određenog stila $\boldsymbol{y}$ tog istog govornika, postiže se kontrola jačine izražajnosti, i to samo mijenjanjem faktora uticaja α i β, tako da je:

$$\boldsymbol{c} = \alpha * \boldsymbol{x} + \beta * \boldsymbol{y}, \quad (1)$$

pri čemu važi da je $\alpha + \beta = 1$. Što je β veće, emocija je izraženija i obrnuto.



Slika 4. *Predložena DNN arhitektura [6].*

Za razliku od prethodno predstavljenog modela (poglavlje 2), ovo pravilo je opšte, važi za bilo koji govorni stil ili emociju i ne zahtjeva emotivno-specifične baze podataka.

4. POREĐENJE PRISTUPA I REZULTATI

Kako sinteza govora zasnovana na t-SNE *embedding* vektorima nije mogla biti realizovana usljed nepostojanja zavisnosti između 2D obilježja i jačine emocije, direktno poređenje metoda nije izvršeno. Ipak, potencijalni rezultati prvog pristupa, u vidu rezultata klasterizacije dobijenih u okviru modula za analizu emocija, upoređeni su sa rezultatima sinteze drugog pristupa kroz subjektivni test slušanja u kojem je učestvovalo 14 slušalaca. Test se sastojao od 24 pitanja – 12 sa rečenicama iz orginalne baze i 12 sa sintetizovanim rečenicama. Svako pitanje je sadržalo po 3 audio snimka koje je trebalo poređati od 1 do 3 u zavisnosti od jačine izražene emocije. Vrijednosti parametra β u sintetizovanim rečenicama su iznosile 0.2 za najmanje izraženu emociju, 0.5 za srednje izraženu emociju, dok su vrijednosti 0.8 i 1.0 ravnopravno korišćene u slučaju najizraženije emocije.

Kako nije bilo moguće realizovati sintezu govora pomoću prvog pristupa, u formiranim testovima rečenice nemaju isti sadržaj. Očekivano je da je slušaocu dosta teško da u potpunosti ignoriše sadržaj rečenice i skoncentriše se

samo na način njenog izgovora, pa i to treba uzeti u obzir pri evalucaiji prikazanih rezultata.

Sa prosječnom tačnošću od 17.9% slušaoci su ispravno odgovorili na sva pitanja u vezi sa originalnom bazom podataka, dok je prosječna tačnost porasla na 56.5% u slučaju sintetizovanih rečenica. Zasebnim posmatranjem emotivnih nivoa uočeno je da su svi nivoi prepoznati sa većom tačnošću u slučaju sintetizovanih rečenica (60.0% - 70.0%), dok je tačnost pri prepoznavanju nivoa originalnih rečenica znatno niža (35.0% - 40.0%).

Zanimljiva zapažanja mogla su se izvesti odvojenim posmatranjem rezultata za emocije sreće i tuge (Tabela 1). Visoka tačnost identifikovanja nivoa postignuta je u sintetizovanom srećnom govoru i iznosila je preko 70.0% za najmanje i srednje izraženu emociju, dok je za najizraženiju emociju ona dostigla čak 92.0%. Kod originalnog srećnog govora, tačnost identifikacije nivoa je bila znatno manja, ispod 45.0%. Najveći nivo izražene tuge prepoznat je sa većom tačnošću u slučaju orginalnih rečenica, i dostiže 92.9%. Pretpostavlja se da je uzrok ovakvog rezultata to što u bazi postoji mali broj rečenica koje zaista imaju jako izraženu emociju tuge i kao takve su izdvojene u zaseban klaster i prepoznate u slučaju originalne baze. Međutim, kako je emocija u bazi u prosjeku bez većih varijacija u nivoima, kada se obuči model za tužni i neutralni stil govora, rečenice se ne razlikuju mnogo od ostalih, pa ni interpolacija između neutralnog i tužnog *embedding* vektora ne dovodi do uočljivih razlika. U slučaju najmanje i srednje izražene emocije, neznatno bolji rezultati su postignuti pri sintezi, ali tačnost i dalje ne prelazi 50.0%. Osobe često na različit način ispoljavaju, a samim tim i doživljavaju emociju tuge, pa niska tačnost može biti i posljedica načina doživljaja emocije.

Tabela 1. *Tačnost prepoznavnaja emotivnih nivoa.*

		Srećni stil govora		
	Emocija	**Slabo izražena**	**Srednje izražena**	**Jako izražena**
Audio signali	***Original***	43.7%	37.5%	42.0%
	Sinteza	76.8%	73.2%	92.0%

		Tužni stil govora		
	Emocija	**Slabo izražena**	**Srednje izražena**	**Jako izražena**
Audio signali	***Original***	34.0%	33.9%	92.9%
	Sinteza	46.4%	39.3%	34.0%

5. ZAKLJUČAK

Praktičnom realizacijom modula za emotivnu analizu u okviru prvog pristupa pokazano je da zavisnost t-SNE obilježja sa jačinom emocije neophodna za kontinualnu manipluaciju nivoom nije ostvariva u opštem slučaju, odnosno zahtijeva specifične baze ekspresivnog govora u kojima su osjetni različitni emotivni nivoi. Sa druge strane, rezultati subjektivnog testa pokazuju da drugi pristup, zasnovan na interpolaciji tačaka u *embedding* prostoru, ne zahtijeva snimanje specifičnih emotivnih baza, već je upravljanje emotivnim stilom moguće i na osnovu podataka koji ne reflektuju varijacije u jačini ekspresivnosti. Naročito dobri rezultati su postignuti za srećni stil govora, međutim, kako je subjektivni test pokazao znatno lošije rezultate za emociju tuge, potrebno je dodatno ispitati pojedinosti po kojima ovaj pristup funkcioniše.

Dalje istraživanje podrazumijeva dodatna testiranja i analizu pojedinačnih dimenzija *embedding* vektora kako bi se utvrdile pravilnosti za njihovu manipulaciju u cilju potpune kontrole i sintetize govora sa željenim karakteristikama i za govornike za koje je dostupna samo neutralna govorna baza.

6. LITERATURA

[1] Iida A., Campbell N., Higuchi F., Yasumura M., "A corpus based speech synthesis system with emotion", *Speech Communication 40*, 161–187. 10, 2003.

[2] Yamagishi J., Onishi K., Masuko T., Kobayashi T., "Acoustic modeling of speaking styles and emotional expressions in HMM based speech synthesis", *IEICE TRANSACTIONS on Information and Systems 88*, 502–509., 2005.

[3] L. Xue, X. Zhu, X. An, L. Xie, "A comparison of expressive speech synthesis approaches based on neural network", *Proc.the Joint Workshop of the 4th Workshop 60 on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pp. 15–20, 2018.

[4] Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, Yusuke Ijima, "An investigation to transplant emotional expressions in DNN-based tts synthesis", *Asia- Pacific Signal and Information Processing Association Summit and Conference*, pages 1253–1258, 2017.

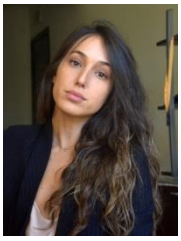
[5] Zhu, X., Xue, L., "Building a Controllable Expressive Speech Synthesis System with Multiple Emotion Strengths", *Cognitive Systems Research*, Volume 59, Pages 151-159 January 2020.

[6] Milan Sečujski, Darko Pekar, Siniša Suzić, Anton Smirnov, Tijana Nosek, "Speaker/Style-Dependent Neural Network Speech Synthesis Based on Speaker/Style Embedding", *Journal of Universal Computer Science*, vol. 26, no. 4, 434-453, 2020.

[7] Florian Eyben, Felix Weninger, Martin Wöllmer, Björn Schuller, "open-Source Media Interpretation by Large feature-space Extraction", *audEERING GmbH*, Version 2.3, November 2016.

[8] Laurens van der Maaten, Geoffrey Hinton, "Visualizing Data using t-SNE", *Journal of Machine Learning Research 9*, 2579-2605, 2008.

Kratka biografija:



Mia Vujović rođena je u Nikšiću 1997. god. Diplomski rad na Fakultetu tehničkih nauka odbranila je 2019. godine čime je stekla zvanje diplomiranog biomedicinskog inženjera. Iste godine je upisala master studije na Fakultetu tehničkih nauka, smjer obrada signala. Ispite je položila 2020. godine sa prosječnom ocjenom 10.0 i time stekla uslov za odbranu master rada.