Kloniranje glasa u realnom vremenu

Corentin Jemine

Rezimirao Lazar Jelić

Apstrakt

Opisan je *text-to-speech* (TTS) sistem za sintezu govora na osnovu teksta. U osnovi ovog sistema su tri nezavisne pretrenirane duboke neuralne mreže koje su u stanju da generišu govor u glasu ciljnog govornika, čiji se uzorci govora ne nalaze u skupu podataka za obučavanje. U [2] je opisan gorenavedeni sistem, koji će se u nastavku označavati sa SV2TTS, dok je u [1] data matematička osnova i implementacija koja podržava rad ovog sistema u realnom vremenu.

1 Uvod

Cilj konstruisanja sistema SV2TTS je generisanje prirodnog govora za raznolike govornike na osnovu relativno malog skupa podataka. Rad se fokusira na *zero-shot* učenje, u kome je nekoliko sekundi govora ciljnog govornika dovoljno da se sintetiše novi govor u glasu ciljnog govorika na osnovu teksta, bez ažuriranja parametara modela ovog sistema.

Ovakvi sistemi imaju razne primene, kao što je mogućnost komunikacije prirodnim govorom kod ljudi koji su izgubili glas i stoga ne mogu da obezbede nove uzorke za obučavanje modela ovog sistema. Pored ove primene, moguće je iskorstiti ovakve sisteme za potrebe prevođenja jezika. Bitno je napomenuti da postoji opasnost od potencijalne zloupotrebe ovakvih sistema, kao što je npr. upotreba nečijeg glasa bez dozvole radi impersonacije te osobe.

Zbog ove opasnosti autori dokazuju da se generisani glas može lako razlikovati od originalnog.

2 Definicija problema

Razmotrimo skup podataka uzoraka grupisanih po govorniku. Označimo j-ti uzorak i-tog govornika sa $\boldsymbol{u}_{ij}$. Uzorci su talasnog oblika vremenskog domena. Označimo sa $\boldsymbol{x}_{ij}$ Log-Mel spektrogram uzorka $\boldsymbol{u}_{ij}$. Log-Mel spektrogram je deterministička neinverzibilna funkcija koja ekstrahuje obeležja glasa na osnovu njegovog talasnog oblika.

Koder $\mathcal{E}$ računa *embedding* $\boldsymbol{e}_{ij} = \mathcal{E}(\boldsymbol{x}_{ij}, \boldsymbol{w}_{\mathcal{E}})$, gde $\boldsymbol{w}_{\mathcal{E}}$ predstavlja parametre kodera. Autori definišu *embedding* vektor govornika kao centroida *embedding* vektora uzoraka tog govornika:

$$\boldsymbol{c}_i = \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{e}_{ij} \tag{1}$$

Sintisajzer $\mathcal{S}$ parametrizovan sa $\boldsymbol{w}_{\mathcal{S}}$ ima zadatak da aproksimira $\boldsymbol{x}_{ij}$ na osnovu centroide $\boldsymbol{c}_i$ i transkript $\boldsymbol{t}_{ij}$ uzorka $\boldsymbol{u}_{ij}$. Umesto $\hat{\boldsymbol{x}}_{ij} = \mathcal{S}(\boldsymbol{c}_i, \boldsymbol{t}_{ij}, \boldsymbol{w}_{\mathcal{S}})$ može se koristiti $\hat{\boldsymbol{x}}_{ij} = \mathcal{S}(\boldsymbol{u}_{ij}, \boldsymbol{t}_{ij}, \boldsymbol{w}_{\mathcal{S}})$ kako bi se omogućilo obučavanje modela korišćenjem manjeg skupa podataka, generalizovala obeležja različitih govornika i podržao rad sistema u realnom vremenu.

Konačno, vokoder $\mathcal{V}$, parametrizovan sa $\boldsymbol{w}_{\mathcal{V}}$, ima zadatak da aproksimira $\boldsymbol{u}_{ij}$ na osnovu $\hat{\boldsymbol{x}}_{ij}$. Imamo $\hat{\boldsymbol{u}}_{ij} = \mathcal{V}(\hat{\boldsymbol{x}}_{ij}, \boldsymbol{w}_{\mathcal{V}})$.

Obučavanje modela se može posmatrati kao minimizacije funkcije troška:

$$min_{\boldsymbol{w}_{\mathcal{E}}, \boldsymbol{w}_{\mathcal{S}}, \boldsymbol{w}_{\mathcal{V}}} L_{\mathcal{V}}(\boldsymbol{u}_{ij}, \mathcal{V}(\mathcal{S}(\mathcal{E}(\boldsymbol{x}_{ij}, \boldsymbol{w}_{\mathcal{E}}), \boldsymbol{t}_{ij}, \boldsymbol{w}_{\mathcal{S}}), \boldsymbol{w}_{\mathcal{V}})) \tag{2}$$

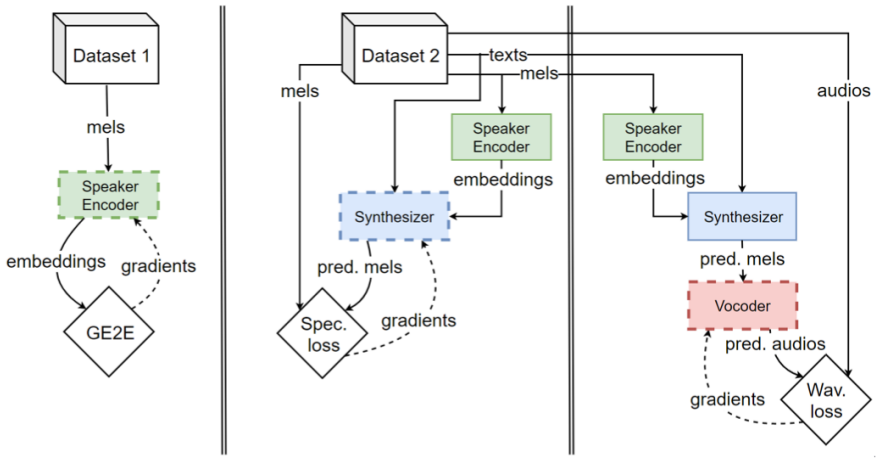
Ovakav pristup ima par nedostataka:

- Neophodno je obučavanje sva tri modela nad istim skupom podataka. To je u ovom slučaju problem zato što se koriste modeli pretrenirani na različitim skupovima podataka.
- Konvergencija modela se teško postiže. Za razliku od drugih komponenti ovog sistema, sintisajzeru je potrebno značajno vreme pre generisanja ispravnih poravnanja Mel spektrograma.

Očigledan način za rešavanje drugog problema je odvojeno obučavanje sintisajzera i vokodera. Pretpostavimo da se koristi pretreniran koder. Sintisajzer se može obučiti da poravna Mel spektrograme ciljnog uzorka:

$$min_{\boldsymbol{w}_{\mathcal{S}}} L_{\mathcal{S}}(\boldsymbol{x}_{ij}, \mathcal{S}(\boldsymbol{e}_{ij}, \boldsymbol{t}_{ij}, \boldsymbol{w}_{\mathcal{S}})) \tag{3}$$

Vokoder se tada može obučavati direktno nad Mel spektrogramima.

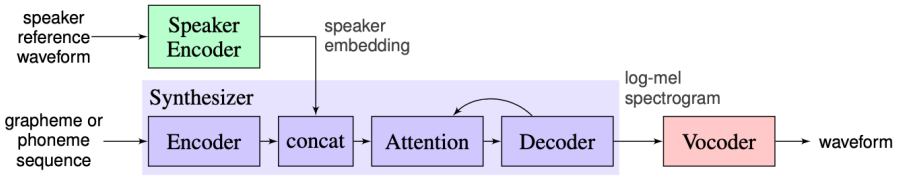


Slika 1: Detaljni pregled obučavanja SV2TTS modela

3 Arhitektura modela

Sistem se sastoji od tri nezavisno obučene komponente:

- **Koder govornika** obučena nad hiljadama govornika za zadatak verifikacije govornika pomoću skupa podataka koji je sačinjen od govora sa pozadinskim šumom bez transkripata, da bi generisala *embedding* vektor fikse dimenzije na osnovu samo nekoliko sekundi referentnog govora ciljnog govornika;
- **Sintisajzer** baziran na modelu *Tacotron 2* [4] koji generiše Mel spektrogram na osnovu teksta i *embedding* vektora govornog signala;
- **Neuralni vokoder** zasnovan na dubokoj neuralnoj mreži *WaveNet* [3] koji pretvara generisani Mel spektrogram u uzorke talasnog oblika vremenskog domena.



Slika 2: Pregled arhitekture sistema SV2TTS

3.1 Koder govornika

Koder govornika se koristi za uslovljavanje sintisajzera da generiše Mel spektrogram koji odgovara referentnom glasu ciljnog govornika. Korišćenje reprezentacije koja zahvata karakteristike različitih govorinika je kritično za dobru generalizaciju, kao i mogućnost identifikacije tih karakteristika samo pomoću kratkog signala, nezavisnog od njegovih fonetskih sadržaja i prisustva pozadinskog šuma.

Mreža kodera mapira sekvencu prozora Log-Mel spektrograma izračunatih na osnovu uzoraka proizvoljne dužine trajanja na *embedding* vektor fiksne dimenzije. Mreža je obučena da optimizuje generalizovanu funkciju troška koja opisuje identifikaciju govornika tako da *embedding* vektori uzoraka istog govornika imaju veliku kosinusnu sličnost, dok se *embedding* vektori uzoraka različitih govornika nalaze udaljeni u vektorskom prostoru. Skup podataka se sastoji od uzoraka podeljenih na segmente dužine trajanja od 1.6 sekunde (skup obeležja) sa pridruženim oznakama govornika (skup klasa obeležja).

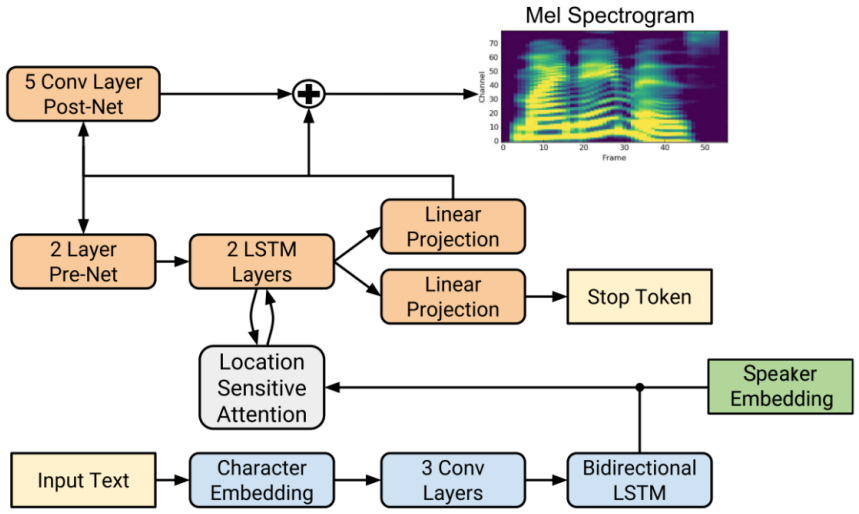
Uzorci predstavljeni kao Log-Mel spektrogrami se prosleđuju neuralnoj mreži koja se sastoji od 3 LSTM sloja sa 768 ćelija. Nakon svakog LSTM sloje sledi projekcioni sloj sa 256 neurona. Finalni *embedding* vektor se kreira L_2 normalizacijom izlaza mreže. Tokom faze zaključivanja, uzorak proizvoljne dužine trajanja je podeljen na prozore dužine trajanja 800 milisekundi sa preklapanjem od 50%. Mreža se nezavisno pokreže za svaki prozor, a izlazi se usrednjuju i normalizuju kako bi se kreirao finalni *embedding* vektor.

3.2 Sintisajzer

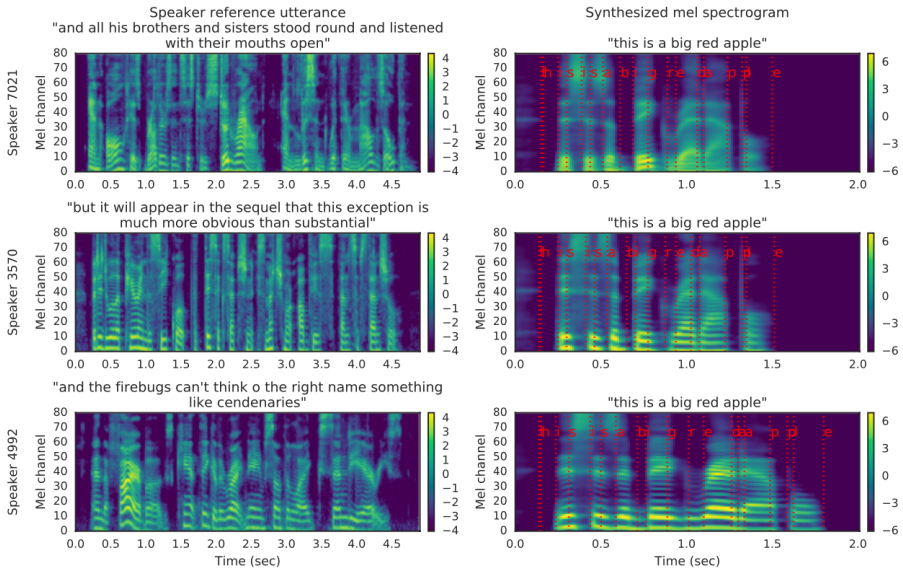
Proširuje se rekurentna neuralna mreža *Tacotron 2* kako bi se podržao veći broj govornika. *Embedding* vektor ciljnog govornika se konkatenira sa izlazom kodera govornika.

Sintisajzer je obučen nad uređenim parovima $(\boldsymbol{u}_i, \boldsymbol{t}_j)$, gde $\boldsymbol{u}_i$ predstavlja uzorak i-tog govornika, dok $\boldsymbol{t}_j$ predstavlja j-ti transkript. Na ulazu u mrežu, vrši se mapiranje transkripta na sekvencu fonema, koje doprinose bržoj konvergenciji i poboljšanom izgovoru reči koje se retko pojavljuju u svakodnevnom jeziku. U osnovi mreže se nalazi pretreniran model *Tacotron 2* na čiji se ulaz dovodi izlaz kodera govornika sa fiksiranim parametrima kako bi se ekstrahovao *embedding* vektor govornika. U ovoj fazi ne postoji informacija o identitetu govornika čiji se uzorak govora obrađuje.

Ciljni Mel spektrogram se računa nad prozorima dužine trajanja 50 milisekundi uz korak dužine trajanja 12.5 milisekundi. Za ove potrebe se koristi banka sa 80 filtera Mel skale koji su praćeni sa Log kompresijom dinamičkog opsega (eng. *dynamic range compression*). Na izlaznom sloju mreže se primenjuju L_2 i L_1 normalizacije, respektivno.



Slika 3: Arhitektura modifikovanog modela *Tacotron 2*

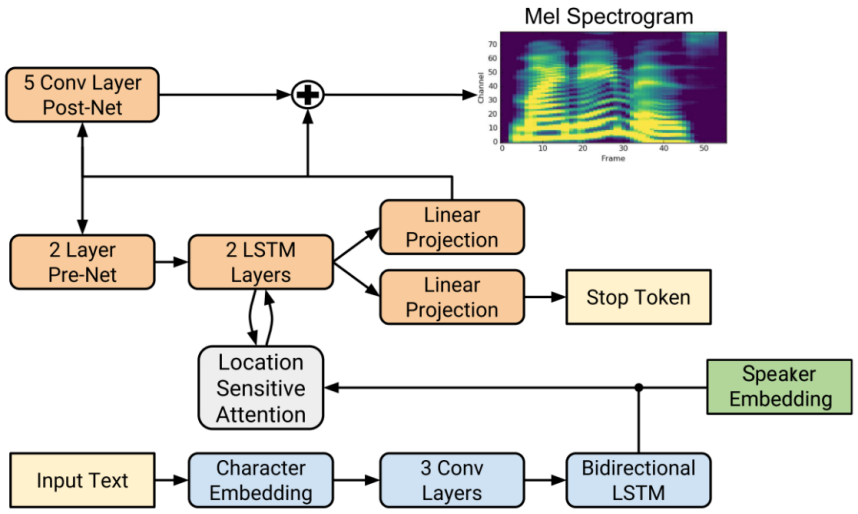


Slika 4: Primer sinteze rečenice drugačijim glasovima

3.3 Neuralni vokoder

Koristi se uzorak-po-uzorak autoregresivna duboka mreža *WaveNet* kao vokoder za inverziju generisanog Mel spektograma. Mreža se sastoji od 30 proširenih konvolucionih slojeva. Činenjica da sintisajzer generiše Mel spektrogram koji

obuhvata sve relevantne detalje potrebne za visokokvalitetnu sitezu različitih glasova dopušta relativno jednostavno obučavanje vokodera tako što se koriste skupovi podataka sa govorima velikog broja govornika.



Slika 5: Arhitektura modela *WaveNet*

4 Implementacija

Prvi korak u korišćenju autorove implementacije ideje opisane u radu jeste instaliranje neophodne biblioteke *tensorflow* za rad sa neuralnim mrežama, biblioteka poput *librosa* i *webrtcvad* za obradu audio signala, kao i biblioteke *multiprocess* za paralelizaciju obučavanja. Pored ovih biblioteka koriste se i one za konstrukciju korisničkog interfejsa *toolbox* modula, ali sa obzirom na to da će se na lokalnoj mašini koristiti samo *CLI* ovog modela, nije ih potrebno navoditi.

Drugi korak je pisanje jednostavnih *Bash* skripte *install.sh* za kloniranje repozitorijuma, instalaciju biblioteka iz prethodnog koraka i preuzimanje skupa podataka i pretreniranih modela. Pored ove skripte napisana je i *train.sh* skripta za automatizaciju pokretanja obučavanja sve tri komponente modela na lokalnoj mašini.

Treći korak je pisanje *Python* skripte *main.py* za testiranje pretreniranog modela ili modela obučenog na lokalnoj mašini. Očekuje se od korisnika da obezbedi uzorke govora dužih od 5 sekundi za potrebe testiranja. Ove uzorke je potrebno postaviti u direktorijum *data* i promeniti konstantu *SAMPLE_IN* u *main.py* skripti.

Kako bi *main.py* uspešno snimila generisani uzorak govora u *WAV* formatu, neophodno je promeniti samu implementaciju. Izmena podrazumeva da funkcija *preprocess_wav* modula *encoder.audio* pored samog audio signala vrati i frekvenciju uzorkovanja kao povratnu vrednost. Zbog propagacije povratnih vrednosti kroz module višeg nivoa, potrebno je promeniti one module koji pozivaju ovu funkciju. Nakon ovoga, model je spreman za obučavanje i testiranje.

5 Skup podataka

Autori koriste dva javna skupa podataka koji se sastoje od uzoraka govora različitih govornika.

Prvi skup, VCTK sadrži 44 sata čistog govora 109 govornika. Tišina i pozadinski šum prisutan u svakom uzorku su uklonjeni što medijanu dužine trajanja uzorka smanjuje sa 3.3 sekunde na 1.8 sekundi. Frekvencija uzorkovanja je smanjena na 24 kHz. Skup podataka je podeljen na tri podskupa: obučavajući skup, validacioni skup, testni skup.

Drugi skup, LibriSpeech sadrži uniju dva skupa podataka čistog govora. Prvi se sastoji od ukupno 436 sati govora 1172 govornika sa frekvencijom uzorkovanja 16 kHz. Tišina i pozadinski šum prisutan u svakom uzorku su uklonjeni što medijanu dužine trajanja uzorka smanjuje sa 14 sekundi na 5 sekundi. Skup podataka je podeljen na tri podskupa: obučavajući skup, validacioni skup, testni skup.

Vrši se nezavisno testiranje sintisajzera i vokodera nad oba skupa podataka opisana u prethodnoj sekciji. Koder govornika se testira nad skupom podataka koji se sastoji od 36 miliona uzoraka sa medijanom dužine trajanja od 3.9 sekundi snimljenih 18 anonimnih hiljada govornika.

6 Rezultati

Autori se oslanjaju na subjektivne evaluacije gomile ljudi pomoću MOS (eng. *Mean Opinion Score*) sistema. Svi generisani uzorci govora se ocenjuju na skali od 1 do 5 za kriterijume prirodnosti i sličnosti sa ciljnim govornikom.

Obučavajući skup	**Testni skup**	**Prirodnost**	**Sličnost**
VCTK	LibriSpeech	4.28 ± 0.05	1.82 ± 0.08
LibriSpeech	VCTK	4.01 ± 0.06	2.77 ± 0.08

Tabela 1: Evaluacija nepoznatih govornika

Sa druge strane, rezultati dobijeni na lokalnoj mašini se ne mogu porediti sa rezultatima dobijenim od strane autora opisanog rada. Generisani uzorci ne sadrže izgovorene reči, već se mogu čuti samo različite vrste isprekidanog šuma. Fiksiranjem vokodera (izlaz modela) mogu se oceniti enkoder govornika i sintisajzer (ulaz modela). Na osnovu ove dve činjenice može se zaključiti da je upravo sintisajzer slaba tačka modela obučavanog na lokalnoj mašini zato što enkoder sadrži informaciju samo o boji glasa, a ne o sadržaju uzorka govora. Čitalac može dati subjektivnu ocenu na osnovu generisanih uzoraka koji se nalaze u repozitorijumu projekta.

Postoje dva osvnovna razloga zbog kojih obučavani model ima loše performanse. Prvi razlog je nedostatak procesorske moći za obučavanje modela u razumnom vremenu. Zbog ovoga se model obučavao relativno kratko i nije stigao da konvergira ka rezultatima koji su autori opisali. Drugi razlog je nedostatak prostora na disku za skladištenje skupa podataka u svrhe obučavanja i testiranja modela. Zbog ovoga model nije imao prilike da obučava nad uzorcima govora od strane različitih govornika što smanjuje njegovu mogućnost da generalizuje.

CPU	Intel Core i5 Dual-Core 1.6 GHz
GPU	Intel HD Graphics 6000 1536 MB
RAM	8 GB 1600 MHz DDR3
Storage	SSD 128 GB

Tabela 2: Specifikacije lokalne mašine

7 Zaključak

Pokazan je TTS sistem koji kombinuje nezavisno obučavan koder govornika, sintisajzer i neuralni vokoder. Upotrebom *embedding* vektora koji sadrže jedinstvena obeležja govornika, sintisajzer je u mogućnosti da generiše Mel spektrograme koji odgovaraju visokokvalitetnom govoru ne samo za glasove govornike čiji su uzorci prisutni u skupu podataka, već i za glasove govornika nad čijim uzorcima se ovaj model nikada ranije nije obučavao. Način obučavanja modela poznat kao *transfer learning* je neophodan za uspeh celog sistema. Razdvajanjem obučavanja kodera govornika i sintisajzera, sistem znatno smanjuje zahteve za obim skupa podataka.

8 Resursi

Demo: `https://github.com/jelic98/raf_pg/blob/main/project`

Reference

[1] Corentin Jemine. Master thesis: Real-time voice cloning, 2019.

[2] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis, 2019.

[3] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis, 2018.

[4] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions, 2018.