Univerzitet u Nišu
Elektronski fakulet Niš

Prevođenje teksta u glas (sinteza govora)

Mentor: Leonid Stoimenov

Studenti:
16080, Nenad Djordjević
17561, Ivan Bogosavljević
17907, Nikola Rašić

Sadržaj

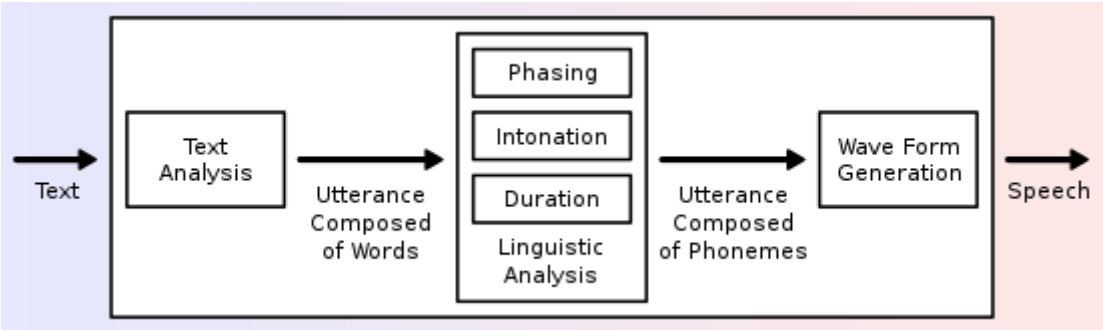
1. Uvod

Sinteza govora predstavlja proizvodnju ljudskog govora putem računarskog sistema poznatog kao sintisajzer govora. Ovaj sistem može biti implementiran u softverske ili hardverske proizvode i koristi se za pretvaranje teksta u govor. Postoji i obrnuti proces, poznat kao prepoznavanje govora. Sintetizovani govor se može stvoriti spajanjem delova snimljenog govora iz baza podataka. Sistemi se razlikuju po veličini memorisanih govornih jedinica. Kvalitet generisanog sintetičkog glasa se poredi sa ljudskim glasom, što je kriterijum za određivanje kvaliteta sistema.

Sistem za pretvaranje teksta u govor se sastoji od dva dela: **Prednjeg i Zadnjeg dela**.

Prednji deo ima dva glavna zadatka. Prvo, sirovi tekst koji sadrži simbole kao što su brojevi i skraćenice pretvara u ekvivalent ispisanih reči. Ovaj proces se često naziva normalizacija teksta, prethodna obrada ili tokenizacija. On zatim svakoj reči dodeljuje fonetske transkripcije i deli i obeležava tekst u prozodijske jedinice, kao što su fraze, klauzule i rečenice. Proces dodeljivanja fonetskih transkripcija rečima naziva se konverzija teksta u fonem ili konverzija grafema u fonem. Fonetske transkripcije i informacije o prozodiji zajedno čine simboličku lingvističku reprezentaciju koju daje front-end.

Pozadinski deo - koji se često naziva sintisajzer - zatim pretvara simboličku lingvističku reprezentaciju u zvuk. U određenim sistemima, ovaj deo uključuje izračunavanje ciljne prozodije (kontura tona, trajanje fonema), kojim se onda stvara izlazni govor.

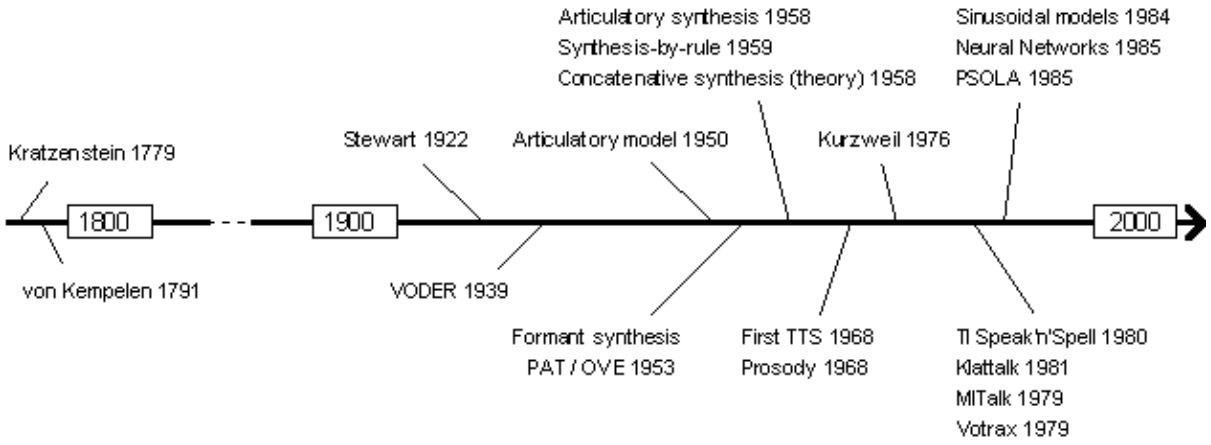


Slika 1. Generalni prikaz sistema pretvaranja teksta u govor

2. Istorija

Iako je tehnologija pretvaranja teksta u govor relativno nova disciplina tehnologije, još od 12. veka ljudi su pokušali da “imitiraju” i oponašaju ljudski govor. Pretečom svih sinteza govora se smatra Rodžer Bejkon. Bronzana ili mesingana glava bila je legendarni automat u ranom modernom periodu čije je vlasništvo pripisivano srednjevekovnim naučnicima, kao što je Rodžer Bejkon, koji je stekao reputaciju čarobnjaka zbog ove sprave. Ona je bila izrađena od bronze ili mesinga, predstavljala je mehaničku iluziju i imala je mogućnost da “odgovara” na pitanja, na čije je odgovore uglavnom odgovarala sa da ili ne.

Prvu konkretnu mašinu 1779. godine napravio je nemačko-danski naučnik Kristijan Gotlib Kracenštajn i za nju osvojio prvu nagradu na konkursu koji je raspisala Ruska carska akademija nauka i umetnosti, za modele koje je napravio po ugledu na ljudski vokalni trakt i koji su bili u stanju da proizvedu pet samoglasnika (a, e, i, o, u).



Slika 2. Istorija sinteze govora

2.1. Prvi elektronski uređaji

Prvi kompjuterski zasnovani sistemi za sintezu govora nastali su kasnih 1950-ih godina. Noriko Umeda razvio je prvi opšti engleski sistem za pretvaranje teksta u govor 1968. godine, u Elektrotehničkoj laboratoriji u Japanu. Godine 1961., fizičar Džon Leri Keli mlađi i njegov kolega Luis Gerstman koristili su računar IBM 704 da sintetišu govor, što je jedan od najistaknutijih događaja u istoriji Bell Laboratorije. Dominantni sistemi 1980-ih i 1990-ih bili su DECtalk sistem, zasnovan uglavnom na radu Denisa Klatta sa MIT-a, i Bell Labs sistema. Ovaj poslednji je bio jedan od prvih višejezičnih sistema nezavisnih od jezika, koji je u velikoj meri koristio metode obrade prirodnog jezika.

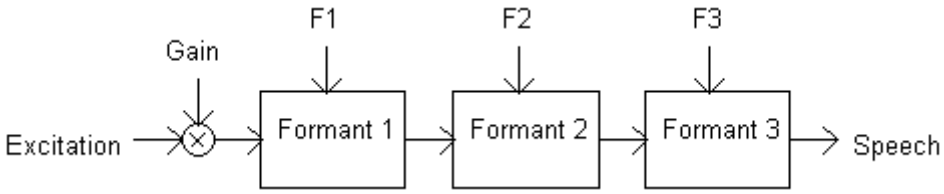
3. Tehnologije sintisajzera

Najvažniji kvaliteti sistema za sintezu govora su prirodnost i razumljivost. Prirodnost opisuje koliko blisko izlaz zvuči kao ljudski govor, dok je razumljivost lakoća sa kojom se izlazni govor razume. Idealan sintisajzer govora je i prirodan i razumljiv. Sistemi za sintezu govora obično pokušavaju da maksimiziraju obe karakteristike.

Dve primarne tehnologije koje generišu sintetičke govorne talase su konkatenativne sinteze i formantne sinteze. Svaka tehnologija ima prednosti i slabosti, a predviđena upotreba sistema sinteze cobično određuje koji pristup se koristi.

Konkatenativna sinteza se zasniva na spajanju (nizanju) segmenata snimljenog govora. Generalno, konkatenativnom sintezom se dobija sintetizovani govor koji najprirodnije zvuči.

Formantna sinteza ne koristi uzorke ljudskog govora tokom izvršavanja. Umesto toga, sintetizovani govorni izlaz se kreira korišćenjem aditivne sinteze (spajanje sinusnih talasa) i akustičkog modela (sinteza fizičkog modeliranja).



Slika 3. Osnovna struktura kaskadnog formantnog sintisajzera

4. Problemi sinteze govora

Najčešći problemi u sintezi govora odnose se na normalizaciju teksta, pretvaranja teksta u foneme (glasove), kao i intonaciju i emociju tokom govora. Proces normalizacije teksta nije jednostavan. Tekstovi su puni heteronima, brojeva i skraćenica koje sve zahtevaju proširenje u fonetsku predstavu. Na engleskom jeziku postoji mnogo pravopisnih pravila koji se izgovaraju različito u zavisnosti od konteksta.

Odlučivanje o tome kako prevesti brojeve je još jedan problem koji TTS sistemi moraju da reše. Jednostavan je programski izazov pretvoriti broj u reči (barem na engleskom), kao što je „1325“ da postane „hiljadu trista dvadeset pet“. Međutim, brojevi se javljaju u mnogo različitih konteksta; „1325“ se takođe može čitati kao „jedan tri dva pet“, „trinaest dvadeset pet“ ili „trinaest stotina dvadeset pet“.

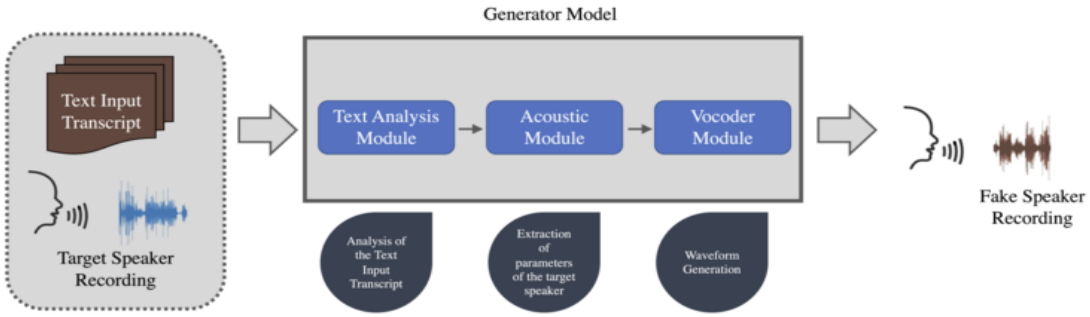
Sistemi za sintezu govora koriste dva osnovna pristupa za određivanje izgovora reči na osnovu njenog pravopisa, procesa koji se često naziva konverzija teksta u fonemu ili grafema u fonemu (fonema je termin koji lingvisti koriste za opisivanje karakterističnih zvukova u jeziku). Najjednostavniji pristup konverziji teksta u foneme je pristup zasnovan na rečniku, gde se rečnik reči i izgovora čuva u programu, tj. bazi podataka.

5. Audio deepfakes

Audio deepfake je tehnologija veštačke inteligencije koja se koristi za generisanje govora koji zvuči kao da je izrečen od strane određene osobe, čak i ako ta osoba nikada nije izrekla te reči. Ova tehnologija je prvobitno razvijena za korisne primene kao što su pravljenje audio knjiga i pomoć ljudima koji su izgubili glas. Komercijalno, otvara se niz mogućnosti, uključujući personalizovane digitalne asistente i usluge prevođenja govora sa prirodnim zvukom.

Audio i video deepfakes, nedavno nazvane audio i video manipulacije, postaju široko dostupni pomoću jednostavnih uređaja kao što su mobilni telefoni ili računari. Ovi alati su takođe korišćeni za širenje dezinformacija pomoću zvuka, pa čak i lažnih video zapisa. Ovo je dovelo do zabrinutosti globalne javnosti u vezi sa sajber-bezbednošću u vezi s nuspojavama korišćenja lažnih audio zapisa. Ljudi ih mogu koristiti kao logičku pristupnu tehniku lažiranja glasa, gde se mogu koristiti za manipulisanje javnim mišljenjem radi propagande, klevete ili terorizma. Ogromne količine glasovnih snimaka se svakodnevno prenose preko Interneta, a otkrivanje lažiranja je izazov. Međutim, ljudi koji su iskorišćavali audio-lažnjake ciljali su ne samo pojedince i organizacije, već i političare i vlade. Početkom 2020. godine, prevaranti su koristili softver zasnovan na

veštačkoj inteligenciji da imitiraju glas izvršnog direktora da bi odobrili transfer novca od oko 35 miliona dolara putem telefonskog poziva. Zbog toga je neophodno autentifikovati svaki audio snimak koji se distribuira kako bi se izbeglo širenje dezinformacija.



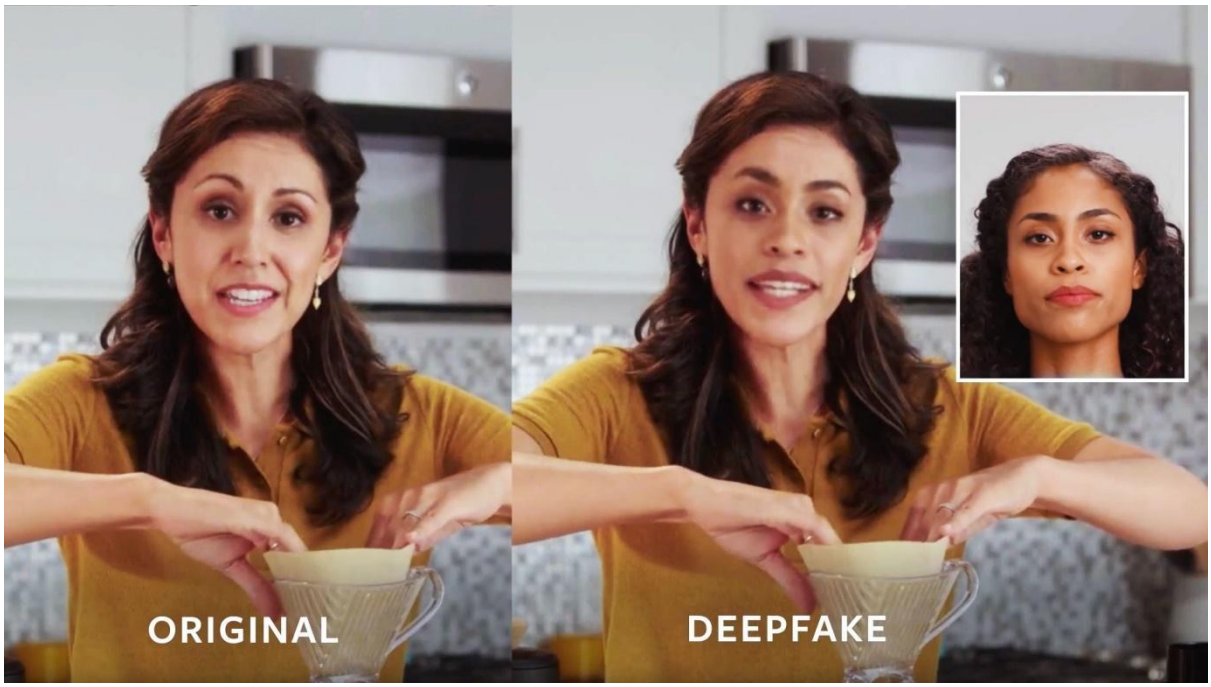
Slika 4. Generisanje lažnog govora na uzorku pravog govora

Postoje tri kategorije audio deepfake zapisa i to su:

Zasnovani na ponavljanju - Deepfakes zasnovani na ponavljanju su zlonamerna dela koja imaju za cilj da reprodukuju snimak glasa sagovornika.

Zasnovani na sintezi - Kategorija zasnovana na sintezi govora odnosi se na veštačku proizvodnju ljudskog govora, korišćenjem softverskih ili hardverskih sistemskih programa. Sinteza govora uključuje pretvaranje teksta u govor, i ima za cilj da u realnom vremenu stvori veštački govor.

Zasnovani na imitaciji - Audio deepfake zasnovan na imitaciji je način transformacije originalnog govora iz jednog izvora (originala) - tako da zvuči kao da je to izgovorio neki drugi govornik (izvor).



Slika 5. Primer deepfake videa

6. Detekcija lažnih audio zapisa

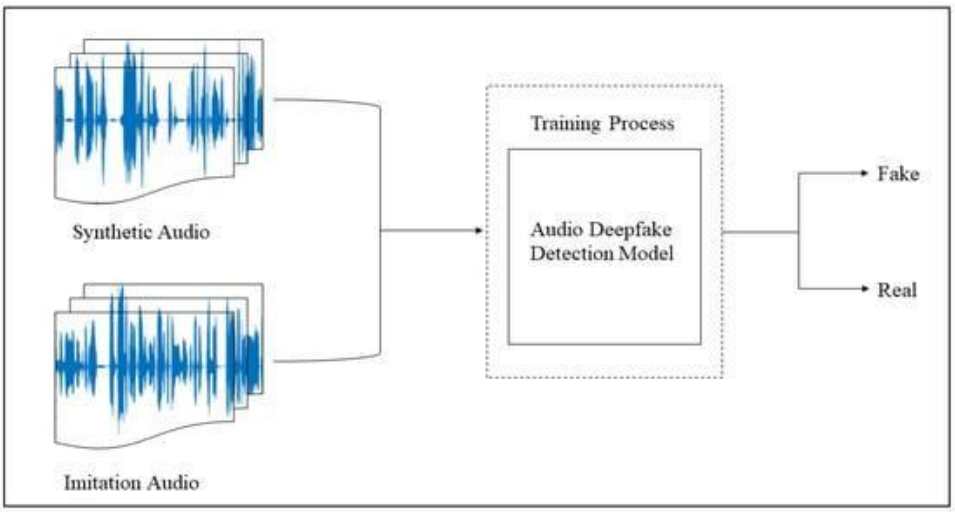
Detekcija lažnog zvuka je zadatak koji se bavi određivanjem da li je dati zvuk govora stvaran ili lažan. Ova tema postaje sve važnija u forenzičkoj istraživačkoj zajednici, kako bi se pratila brza evolucija tehnika falsifikovanja.

Metode otkrivanja lažnog zvuka se mogu podeliti u dve kategorije na osnovu aspekta koji se koriste za obavljanje ovog zadatka. Prva kategorija se fokusira na aspekte niskog nivoa, tražeći stvari koje su uveli generatori na nivou uzorka. Druga kategorija se, umesto toga, fokusira na karakteristike višeg nivoa, koje predstavljaju složenije aspekte kao što je semantički sadržaj govornog audio zapisa.

Što se tiče Deepfake audio zapisa, uvedene su mnoge metode detekcije kako bi se lažne audio datoteke razlikovale od stvarnog govora. Razvijeno je više Machine Learning (mašinsko učenje) i Deep Learning modela koji koriste različite strategije za otkrivanje lažnog zvuka.

Mašinsko učenje je grana veštačke inteligencije (AI) i računarske nauke koja se fokusira na korišćenje podataka i algoritama za imitiranje načina na koji ljudi uče, postepeno poboljšavajući tačnost učenja kod programa i računara.

Prvenstveno, svaki audio snimak treba prethodno obraditi i transformisati u odgovarajuće audio karakteristike, kao što su Mel-spektrogrami, koji se koriste za predstavljanje jačine zvuka na određenoj frekvenciji kroz logaritamsku skalu. Ove karakteristike se unose u model detekcije, koji zatim vrši neophodne operacije, kao što je proces obuke. Izlaz se dovodi u bilo koji potpuno povezan sloj sa funkcijom aktivacije da bi se proizvela verovatnoća predviđanja klase 0 što znači da je audio lažan ili klase 1 kao stvarnog zvuka.



Slika 6. Ilustracija detekcije lažnog govora

7. Upotreba sinteze govora u današnjoj tehnologiji

Prevođenje teksta u govor u 21. veku, a najviše sa nastankom društvenih mreža i shodno njihovom proširenju na široke narodne mase, ima veliki značaj i uticaj na sve aspekte života. Jedan od primera korišćenja TTS sistema je na društvenim mrežama Instagram, TikTok, YouTube, i to u svrhe kreiranja video sadržaja kratkog formata. Svaka društvena mreža ih drugačije naziva, Reels, TikToks ili Shorts. Moguće je tokom kreiranja ovih video sadržaja ispisati tekst koji će se pojavljivati na ekranu i koji će neki glas veštačke inteligencije čitati. Na ovaj način kreatori ne moraju da pokazuju svoje lice ili glas, i mogu anonimno da postavljaju razne sadržaje, koji mogu da imaju pozitivan ili negativan uticaj na bilo koga ko gleda taj sadržaj.

Druga vrlo korisna primena ove tehnologije je kreiranje i video i audio lažnih zapisa. Može se koristiti u prezentacione svrhe ili u svrhe kreiranja video sadržaja za platforme kao što je YouTube. Primer ovakvog vida korišćenja TTS sistema je sajt https://www.synthesia.io/ na kome se mogu kreirati video i audio zapisi uz pomoć veštačke inteligencije u raznorazne svrhe. Kako se navodi na njihovom sajtu, ovaj sajt koriste i velike kompanije poput Amazon, Reuters, BBC…

Još jedna vrlo zanimljiva stvar u kojoj se koriste ovi sistemi su generisanje glasova iz filmova, serija, crtanih filmova ili igrica. Na sajtovima https://app.uberduck.ai/ i https://fakeyou.com koristi se sinteza govora koja može generisati zvuk omiljenih TV zvezda na primer.

Jedan mnogo poznatiji primer ovakvih sistema koristimo u svakodnevnom životu. Koristan je kada putujemo u zemlje čiji jezik ne znamo ili kada želimo da prevedemo nešto što ne razumemo, a zatim i čujemo izgovor te reči ili fraze. U pitanju je Google Translate. U razgovoru sa nekim ko ne zna naš jezik možemo ukucati reč ili frazu u ovaj program i pustiti isti da sintetiše govor na jeziku koji nam je potreban. Osoba takođe može pričati, a program će sam pretvoriti glas u tekst, tj izvršiti obrnut proces, prepoznavanje govora.

8. Problem koji želimo rešiti

Uočili smo da na tržištu mobilnih aplikacija za Android telefone ne postoji mnogo opcija za aplikacije za dopisavanje za slepe i slabovide ljude. Naš cilj je da kreiramo aplikaciju pomoću koje je moguće tekstualnu poruku pročitati slepoj ili slabovidoj osobi. Ova aplikacija bi bila na principu Facebook Messenger aplikacije i imala bi mogućnost da čita tekstualne poruke, a ujedno i da govor sintetiše u tekst. Ovakva aplikacija bi imala korist i kod gluvih i nagluvih osoba jer bi pomoću videa mogla da prepozna znakovni jezik i zatim ga pretvori u tekst ili govor. Na ovaj način moguća je i komunikacija između npr. slabovide i gluve osobe.

8.1. Postojeća rešenja za ovaj problem

Primena TTS sistema ima veliku upotrebu u mobilnim telefonima ili aplikacijama, međutim ti sistemi se koriste za celokupan mobilni telefon, kao što je Apple Siri, Google assistant i nekad može biti komplikovano poslati poruku osobi koja slabije vidi.

Naša aplikacija bi uklopila sve komponente sinteze govora i teksta, kao i prepoznavanja teksta i znakovnog jezika i bila bi vrlo jednostavna i intuitivna za korišćenje, sa što manje kompleksnih funkcija.

8.2. Predlog našeg rešenja za ovaj problem

Kao što je gore pomenuto, mobilna Android aplikacija je rešenje koje bismo napravili i koja bi uz pomoć algoritama sinteze govora kao što su konkatenativna i formantna sinteza, pretvarala tekst u glas. Prednost bi imao srpsko-hrvatski jezik za tržište celog Balkana, jer trenutno ta opcija na tržištu ne postoji.

Aplikacija bi bila napravljena u Android Studio programu i imala bi jednostavan interfejs: Početni ekran bi bili svi razgovori sa osobama koji bi se čuvali u bazi podataka na serveru, i bilo bi omogućeno aplikaciji da preko prepoznavanja govora otvori specifičan čet sa osobom koja je naglašena. Prilikom otvaranja četa, poruka bi bila označena bojom koja je upečatljiva i koja se dosta razlikuje od pozadine aplikacije, a sve u cilju da slabovida osoba ima mogućnost da prepozna da je to poruka koju želi da odsluša. Jednostavnim klikom na tu poruku, ili glasovnom komandom, za potpuno slepe osobe, moguće je poslušati poruku koja je sintetizovana. Model koji može poslužiti kao glas za čitanje poruka može biti neki vokalni glumac, koji ima dubok tonalitet i lepu boju glasa, npr. Petar Benčina.

Koristili bi kao open-source biblioteku koja se može integrisati u mobilne aplikacije kao što je https://f-droid.org/packages/com.github.olga_yakovleva.rhvoice.android/ koja ima čak i podršku za ruski jezik, jezik vrlo sličan srpsko-hrvatskom, koji ujedno koristi i ćirilično pismo. Još jedan open-source sistem koji može da se iskoristi za sisteme sa minimalnom verzijom Android 4.0 je i https://github.com/espeak-ng/espeak-ng. Za prepoznavanje govora moguće je iskoristiti https://developer.android.com/reference/android/speech/SpeechRecognizer funkciju koja je integrisana u Android okruženje.

Različiti jezici bi se čuvali u posebnim bazama podataka, a bilo bi moguće izabrati i druge jezike.

Još jedna ideja je čitanje sa usana i prepoznavanje reči koje se izgovaraju ili znakovnog jezika i prevođenje u tekst i slanje osobi koja bi mogla da čita ili čuje poruku, ukoliko se npr. dopisuju gluva i slepa osoba.

9. Zaključak

Ova aplikacija bi imala veliku primenu kod svih ljudi koji imaju slabiji vid ili govornu manu sa prostora Balkana i pričaju srpsko-hrvatskim jezikom. Potencijalni broj korisnika je veći od 12.000, s obzirom da samo u Srbiji ima ovoliko slepih i slabovidih osoba.

Aplikacija bi bila besplatna za korišćenje i koristila bi sve open-source besplatne biblioteke i funkcije kako bi održali aplikaciju besplatnom.

Umnogome može poboljšati dopisivanje i izražavanje svih osoba koje imaju problema sa vidom i govorom.

Literatura

- https://en.wikipedia.org/wiki/Speech_synthesis
- https://www.techopedia.com/definition/3647/speech-synthesis
- https://vivoka.com/how-to-speech-synthesis-tts/
- http://research.spa.aalto.fi/publications/theses/lemmetty_mst/contents.html
- https://en.wikipedia.org/wiki/Audio_deepfake
- https://www.mdpi.com/1999-4893/15/5/155
- https://www.ibm.com/topics/machine-learning