Pregledni rad

UDK 811.163.42'322.6(091)
81'322.6(091)
Pregledni rad
Prihvaćen za tisak: 14. listopada 2024.
https://doi.org/10.22210/suvlin.2024.098.07

Dario Poljak, Kristina Kocijan
Filozofski fakultet Sveučilišta u Zagrebu, Zagreb
dpoljak@ffzg.unizg.hr, krkocijan@ffzg.unizg.hr

Odjeci inovacija – putovanje kroz krajolik sustava za sintezu govora na hrvatskom jeziku

Ovaj rad nudi sveobuhvatan pregled razvoja tehnologija sinteze govora, s naglaskom na hrvatski jezik i njegovu usporedbu s drugim slavenskim i nesrodnim jezicima. Analizirajući povijesni napredak od mehaničkih do elektroničkih sustava, rad istražuje kako su se tehnike i metode razvijale te njihov utjecaj na učinkovitost i kvalitetu sintetiziranog govora.

Posebna pozornost posvećena je istraživanju suvremenog pejzaža tj. suvremenim digitalnim sustavima, stavljajući u fokus napredne tehnike i algoritme koji su revolucionirali sintezu govora u jezicima poput engleskog i mandarinskog. Usporedbom s međunarodnim trendovima, rad identificira ključne izazove s kojima se suočava hrvatski jezik, posebice u kontekstu ograničenih jezičnih resursa.

Kao odgovor na ove izazove, rad nas uvodi u doktorsku disertaciju usmjerenu na stvaranje označenog korpusa i razvoj specijaliziranih modela dubokog učenja za hrvatski jezik. Cilj je postaviti temelje za unapređenje sinteze govora na hrvatskom, pridonoseći time globalnom napretku u ovom dinamičnom području.

1. Uvod

Sinteza govora star je problem, otprije poznat istraživačima, i koristi se kao krovni termin za postupke kojima se pisani korpus prevodi i ozvučava u *kvazi* govorni oblik (Lazić 2006). Kao multidisciplinarno područje koje spaja lingvistiku, informatiku i akustiku, sinteza govora predstavlja jedan od ključnih izazova u računalnoj obradi prirodnog jezika. Od prvih mehaničkih pokušaja oživljavanja pisane riječi do suvremenih digitalnih sustava, ovaj rad istražuje evoluciju tehnologija sin-

teze govora s posebnim osvrtom na hrvatski jezik i njegovu interakciju sa srodnim slavenskim jezicima te nesrodnim jezičnim obiteljima.

Razvoj povijesti sinteze govora možemo pratiti od druge polovice 18. stoljeće (Story 2019) s pojavom prvih mehaničkih uređaja koji prethode pojavi prvih računala (1938. – 1942.[1]), kao i pojavi prvog telefona (1877.[2]), pa sve do revolucionarnog VODER–a (engl. *Voice Operating Demonstrator*)[3] Homera Dudleya koji je označio početak **elektroničke** ere sinteze govora (Story 2019; Schroeder 1993). Povijesni pregled koji se pruža u ovom radu nije samo retrospektiva; on je neophodan temelj koji nam omogućuje da kontekstualiziramo sadašnje stanje tehnologije i predvidimo buduće smjerove razvoja. Razumijevanje korijena i evolucije sinteze govora ključno je za identificiranje važnih prekretnica i inovacija koje će oblikovati naš pristup ovoj disciplini u godinama koje dolaze.

Cilj ovog rada ipak nije detaljno analizirati sve sustave za sintezu govora niti njihovu povijest, već usredotočiti se na aktualnu **digitalnu** eru, odnosno razvoj **računalnih sustava za sintezu govora**. Story (2019) vjeruje da digitalna era počinje nedugo nakon razvoja VODER–a. Naime, to je bio prvi put da se cjelokupni proces sinteze pojednostavio pomoću računala. U početku je bilo nužno specijalizirano sklopovlje za sintezu govora, ali vrlo brzo cijeli proces prelazi na softversku razinu i u potpunosti se odvija na računalu. Takav razvoj, kao i postepeno povećanje dostupnosti računala za opću namjenu, smanjio je tehničke zahtjeve (prvenstveno pristup računalnim centrima, te specijalizirano sklopovlje) za sintezu govora. Više o kategorijama računalnih sustava za sintezu govora bit će riječi u narednom poglavlju kojim ćemo dati temelje za prikaz podataka u preostalim poglavljima.

S obzirom na ograničene jezične resurse i specifičnosti hrvatskog jezika, rad se nadalje usredotočuje na analizu i usporedbu s metodama i dostignućima u srodnim slavenskim jezicima (bosanski, srpski, slovenski, makedonski), kao i na razmatranje pristupa primijenjenih na jezicima s bogatim resursima. Kroz ovu analizu, rad identificira ključne faktore koji utječu na razvoj tehnologija sinteze govora i naglašava potrebu za stvaranjem robustnih jezičnih skupova podataka i modela dubokog učenja prilagođenih hrvatskom jeziku.

Ograničavajući opseg rada na najnovija dostignuća do prvog kvartala 2023. godine, ovaj rad pruža temeljit i aktualan pregled stanja tehnologija sinteze govora, postavljajući temelje za buduća istraživanja i razvoj u ovom dinamičnom području.

1 Računalo. *Hrvatska enciklopedija, mrežno izdanje.* Leksikografski zavod Miroslav Krleža, 2021. Pristupljeno 20. 5. 2023.

2 Bell, Alexander Graham. *Hrvatska enciklopedija, mrežno izdanje.* Leksikografski zavod Miroslav Krleža, 2021. Pristupljeno 20. 5. 2023.

3 Napravljen po uzoru na VOCODER (*VoiceCoder* – šifrant glasova) iz Bell laboratorija sredinom 1930–ih.

2. Računalni sustavi za sintezu govora

Nekoliko je kategorija i potkategorija računalnih sustava za sintezu govora. Poslužit ćemo se ovdje podjelom koju predlažu Ning i suradnici (2019) a odnosi se upravo na sustave za sintezu govora u digitalnoj eri. Naime, ovisno o načinu na koji se pristupa sintezi razlikujemo konkatenativne (korpusne) i parametarske sustave, a oni su dodatno podijeljeni u potkategorije koje su definirane specifičnostima pojedine sinteze.

2.1. Konkatenativna sinteza

Pojam *konkatenacije* (ili nadovezivanja) nalazimo u teoriji formalnih jezika za potrebe operacije spajanja nizova znakova. Iz toga se jasno može zaključiti da kad govorimo o konkatenativnoj sintezi govora, govorimo u stvari o spajanju unaprijed snimljenih govornih elemenata (nizova) u svrhu dobivanja sintetiziranog govora (Bakran & Lazić 1998). Unaprijed snimljeni elementi (u ovom slučaju zvukovi) mogu biti na nekoliko razina granularnosti: od difona, trifona, fonema ili slogova pa sve do riječi i potpunih rečenica. Odabir zvučnih jedinica jednako je važan kao i njihova posljedična obrada te konačna sinteza govora.

Suzić (2019) posebno ističe problem koartikulacije koji je prisutan kod mnogih sustava konkatenativne sinteze govora. Koartikulacija, utjecaj glasova jednih na druge[4], u prirodnom govoru daje glatkost i tečnost, odnosno postiže se prirodan ritam i intonacija. Dok je kod konkatenativne sinteze izazov oponašanje tog procesa koristeći ograničeni skup prethodno snimljenih zvučnih segmenata.

Nerijetko se u literaturi može naići i na termin linearno prediktivno kodiranje (engl. *Linear Prediction Coding* – **LPC**). Ova je metoda jedna od najraširenijih u kodiranju govora s primjenama i u prepoznavanju govora, prepoznavanju govornika i sintezi govora (Gupta & Gupta 2016). Petrinović (2009) opisuje kako se originalno LPC–om koristilo Američko ministarstvo obrane za kodiranje govora pri brzinama 2.4 kbit/s, odnosno maksimalnom brzinom tadašnjih modema, a kako bi govor i dalje bio razumljiv. Usredotočen na učinkovitost, a ne prirodnost, ovako dobiven rezultat često je s metalnim prizvukom, odnosno zvukom koji asocira na zujanje. Kako saznajemo od Charpentiera i Stelle (1986), unatoč svojim manjkavostima, LPC metoda je svojevremeno bila popularan pristup u sintezi govora.

S ciljem unaprjeđenja konkatenativne sinteze govora te ispravljanja manjkavosti LPC pristupa sintezi, nastali su tzv. *namjenski algoritmi*. Jedan od njih poznat je pod akronimom **PSOLA** (engl. *Pitch Synchronous Overlap and Add*), a navodimo ga ovdje jer je doveo do poboljšanja koja su se naročito iskazala kod dužih sekvenci govora kao i kod mogućnosti upravljanja prozodijom (Charpentier & Stella 1986). Nešto kasnije, nadograđujući se na PSOLA, predstavljen je i drugi algoritam akronima **MBROLA** (engl. *Multi–Band Resynthesis OverLap Add*) kojem je cilj bio omo-

4 http://struna.ihjj.hr/naziv/koartikulacija/55046/

gućiti besplatnu dostupnost što sveobuhvatnijeg rješenja za sintezu govora (Dutoit *et al*. 1996).

Dutoit je 1994. godine proveo i usporednu analizu ova tri navedena sustava po (i danas važnim) metrikama: **razumljivost**, **prirodnost** i **tečnost** sintetiziranog govora (Tablica 1). LPC model je zbog svoje učinkovitosti u vidu memorije bio i najlošiji s rezultatom od 54,6 % za razumljivost, 44,5 % prirodnost i 50,4 % tečnost. Unaprijeđeni TD–PSOLA algoritam (Moulines & Charpentier 1989) postigao je najbolji rezultat za razumljivost (78 %), te 68,3 % za prirodnost i 65 % za tečnost, dok je MBROLA (tada MBR–PSOLA) postigla 72,8 % za razumljivost, 68,3 % za prirodnost i 75,6 % za tečnost.

	razumljivost	prirodnost	tečnost
LPC	54,6 %	44,5 %	50,4 %
TD–PSOLA	**78,0 %**	**68,3 %**	65,0 %
MBROLA	72,8 %	**68,3 %**	**75,6 %**

Tablica 1. Usporedna analiza algoritama konkatenativne sinteze za engleski jezik

Iz rezultata prikazanih u Tablici 1, vidljiva je dominantnost TD–PSOLA algoritma u području razumljivosti i MBROLA algoritma u području tečnosti, dok iz perspektive prirodnosti postižu isti rezultat. No, postoje i druge razlike između PSOLA i MBROLA algoritama, a najznačajnija jest format njihove licence: MBROLA program se ne može prodavati niti ugraditi u neku drugu cjelinu koja se prodaje, bez odobrenja njegovih autora (Bakran & Lazić 1998). Iako na prvu ruku plemenit cilj, takva restriktivna licenca negativno je utjecala na veću rasprostranjenost rješenja temeljenih na MBROLA–i. Iz perspektive izrade komercijalnih rješenja 10 % tečniji rezultat ne znači ništa ako ga je teško ili nemoguće naplatiti. Unatoč tomu, ne možemo zanemariti doprinos MBROLA programa u potrazi za prirodnijim sintetizatorom govora koja se nastavlja u smjeru parametarske sinteze.

2.2. Parametarska sinteza

Kada se govori o *parametarskoj* sintezi, naglasak je na parametrima (varijablama). U ovom se slučaju cijeli proces vokalizacije, uključujući i sam vokalni trakt, promatra kao nešto što se može oblikovati pomoću parametara. Prema Ningu i suradnicima (2019) parametarskoj sintezi tipično se može pristupiti na nekoliko načina: artikulacijskom sintezom (engl. *articulatory synthesis*), formantnom sintezom (engl. *formant synthesis*), prikrivenim Markovljevim modelima (engl. *Hidden Markov Models – HMM*) i dubokim neuronskim mrežama (engl. *deep neural networks – DNN*). Ključna razlika između pristupa jest u tehničkim zahtjevima provedbe sinteze (od manje zahtjevnih prema zahtjevnijima) te kompleksnosti izrade modela u odnosu na prirodnost sintetiziranog govora.

Kod **artikulacijske sinteze** koristi se matematički model koji opisuje oblik govornog trakta, mehanička kretanja artikulatora i kretanje zračne struje kroz trakt. Na temelju tog modela vrši se sinteza govora. Takvi modeli nerijetko se razvijaju uz pomoć rendgenskih snimaka ljudskog vokalnog trakta za vrijeme izgovora (Pobar, Martinčić–Ipšić & Ipšić 2008). No zbog inherentne dvodimenzionalnosti rendgenskih snimaka, proces modeliranja nije znatno olakšan jer je stvarni vokalni trakt prirodno trodimenzionalan (Lazić 2006). Kada se dodatna kompleksnost pri izradi modela uzme u obzir, rezultati obično budu lošiji od formantne i konkatenativne sinteze, a teže ih je ostvariti (Tan, Qin, Soon & Liu 2021).

Za razliku od artikulacijske sinteze, **formantna sinteza** tvori glas propuštanjem odgovarajućih pobudnih signala kroz linearni filtar podešen prema rezonantnim frekvencijama ljudskog govornog trakta (Pobar, Martinčić–Ipšić & Ipšić 2008). Jedan od najpoznatijih istraživača u području formantne sinteze govora je Dennis Klatt i njegov sintetizator *Klattalk* (Klatt 1982) s bogatom kolekcijom sintetiziranih glasova. Među poznatijim glasovima je i "*Perfect Paul*" kojim se dugo vremena koristio fizičar Stephen Hawking (Story 2019). Uz *Klattalk,* neizostavan je i komercijalni paket *eSpeak*, kojeg je od 2010. godine Google počeo koristiti na svom mrežnom servisu za prijevode, *Google Translate*. Zahvaljujući, između ostalog, i niskim tehničkim zahtjevima formantne sinteze, Google je ovim alatom omogućio sintezu govora na čak 27 različitih jezika (Sciforce 2020). S kvalitativne strane, za proizvodnju razumljivog govora obično su potrebna najmanje tri formanta, a do pet formanata omogućuje proizvodnju govora izrazito visoke kvalitete (Lazić 2006). Unatoč već spomenutim tehničkim prednostima formantne sinteze, velika mana ostaje prirodnost, odnosno nedostatak prirodnosti u sintetiziranom govoru. Petrinović (2009) to dodatno pojašnjava opisujući kako prilikom govora vokalni trakt u izvođenju samih glasova prolazi i kroz niz međustanja (efekt koartikulacije) te izostavljanje bilo kojeg od njih negativno utječe na prirodnost.

Kako bi se to ublažilo, umjesto ručnog namještanja parametara, korištenjem **HMM**–a taj posao prepušta se statističkoj aproksimaciji stohastičkih procesa kod govora. Pobar i suradnici (2008) tvrde da HMM sustavi na temelju analize unaprijed snimljena govora i njegove fonetske transkripcije, konstruiraju tzv. vektore karakteristika. Odnosno, na temelju prirodnog izgovora, bilježe karakteristike o kontekstu što posljedično pozitivno utječe na prirodnost sintetiziranog govora. U tom pogledu HMM sustavi također se navode u literaturi kao sustavi za statističku parametarsku sintezu govora (Ning *et al.* 2019). Međutim, prema Kuligowskoj i suradnicima (2018), rezultati sustava za sintezu govora temeljenih na HMM–a su po prirodnosti manjkavi u odnosu na sustave temeljene na konkatenativnoj sintezi koji su prema njihovom mišljenju također manjkavi i to zbog domene ograničene veličinom korpusa. Stoga oni zagovaraju potrebu za novim, po mogućnosti hibridnim, pristupom.

Relativno nedavno takav pristup se i iskristalizirao u vidu sustava za sintezu govora temeljenom na **dubokim neuronskim mrežama** gdje se pomoću neuron-

skih mreža oba sustava mogu upotpuniti (Story 2019). Naime, duboke neuronske mreže izrazito su učinkovite u izdvajanju karakteristika iz različitih oblika podataka (Ning *et al.* 2019). Kroz primjenu dubokih neuronskih mreža za akustičnu analizu, sintetizatori postaju sve kvalitetniji, bilo kao jedinstvena rješenje poput sustava *TacoTron* (Wang *et al.* 2017), ili kao dio već postojećih sustava za sintezu govora kao što su *WaveNet* (van den Oord *et al.* 2016) ili *WaveGlow* (Prenger 2018). Sustavi za sintezu govora temeljeni na neuronskim mrežama pokazuju zavidne rezultate u mnogim područjima, no dolaze s visokim tehničkim i podatkovnim zahtjevima (Thompson *et al.* 2020). Tako su za potrebe izrade modela *TacoTron*, Wang i suradnici (2017) koristili interni podatkovni set od 24,6 sati govora jedne profesionalne spikerice. Budući je do takvih podatkovnih modela dosta teško doći, složili bi se s Lazićem (2006) koji navodi da iako duboke neuronske mreže u domeni sinteze govora nisu nova ideja, tj. spominju se još od sredine 90–ih, zahtjevi za podacima i računalnom snagom usporili su njihovu širu primjenu, ali to je posljednjih godina sve manji problem.

Pogotovo kada se u obzir uzmu rezultati prirodnosti po MOS (Mean Opinion Score) ljestvici gdje model *WaveGlow* postiže 4,21 od maksimalnih 5 bodova (van den Oord *et al.* 2016). Bodovanje se vrši kroz Likertovu skalu od 1 do 5, gdje 1 predstavlja najgori a 5 najbolji rezultat, za prirodnost govora. Kada taj rezultat pretvorimo u postotak kako bi ga približili modelu ocjenjivanja koji je koristio Dutoit, prikazan u tablici 1, vidjeli bi kako je tih 4,21/5 zapravo 84,2% što je značajan pomak u pozitivnom smislu.

3. Hrvatska

S obzirom na visok stupanj interdisciplinarnosti teme i raspršenost istraživača, u Hrvatskoj se, u odnosu na ostatak Europe, tema sinteze govora počela relativno kasno istraživati. Ta raspršenost osjeti se još i danas u terminologiji koja je često direktan prijevod ili interpretacija bez obaziranja na prijašnja istraživanja u području. Tako se u radovima konkatenativna sinteza još naziva i korpusnom (Lazić 2006), spojnom (Džijan 2020), ulančavanjem (Pobar, Martinčić–Ipšić & Ipšić 2008), i sinteza odabirom jedinica (Pobar 2014). Vrednovanje sustava isto tako nije standardizirano, već svaka skupina autora koristi drukčije, međusobno neusklađene, mjere.

Početna istraživanja u Hrvatskoj bila su na Odsjeku za fonetiku Filozofskog fakulteta u Zagrebu tek 80–ih godina 20. stoljeća, a najveći korak pritom bio je napravljen na samom kraju stoljeća kada se izdala zvučna baza za izgovor, temeljena na algoritmu MBROLA (Bakran & Lazić 1998). Povezano s tim radom i iskustvom u MBROLA–i, Lazić (2006) brani doktorsku temu "Modeliranje strojnih postupaka za izgovaranje teksta pisanoga hrvatskim jezikom".

Početkom 21. stoljeća, Martinčić–Ipšić i Ipšić (2003) predstavljaju govorni korpus VEPRAD. Nedugo nakon, Martinčić–Ipšić sa suradnicima (2004) predstavljaju

Korpus Hrvatskog Govora te opisuju postupak izrade korpusa uz koji navode ukupno trajanje tonskog zapisa od 19 sati i 24 minute. Predstavljeni korpus sastoji se od dva modula: VEPRAD (VrEmenske Prognoze–RADio) te TV HRBCN (Televizijski HRvatski BroadCast News). VEPRAD je podijeljen na snimke s radija u trajanju od 6 sati i 17 minuta, te na snimke telefonskih izvještaja meteorologa u trajanju od 5 sati i 6 minuta dok se TV HRBCN sastoji od snimki s televizije u trajanju od 3 sata i 28 minuta. Ukupno su Martinčić–Ipšić i suradnici prikupili i po SAMPA pravilima fonetski označili gotovo 15 sati hrvatskoga govora (vidi Bakran & Horga 1996 za više primjera fonetski označenog govora). S obzirom na broj objavljenih radova od trenutka objave korpusa do kraja 2010–ih (Martinčić–Ipšić & Ipšić 2006; Pobar, Martinčić–Ipšić & Ipšić 2008; Pobar & Ipšić 2011), možemo reći i da su među najaktivnijim domaćim autorima u domeni sinteze, ali i prepoznavanja govora.

Početkom 2010–ih, u Hrvatskoj se pojavljuju i sustavi temeljeni na gotovim programskim rješenjima poput sustava za sintezu govora *Festival* (Šoić 2010; Pobar & Ipšić 2011) te *SPICE* (Šoić & Dembitz 2011). Rezultat jednog od tih istraživanja bio je sustav *Hascheck Voice* sastavljen 2011. godine (*Hascheck Voice*, 2011)[5]. Pojavljuju se i kompletna programska rješenja za sintezu govora hrvatskih autora poput sutava *CroSS – Croatian Speech Synthesizer* (Dunđer 2013). Svi navedeni radovi koriste varijantu konkatenativne sinteze pomoću difona ili statističke parametarske sinteze temeljene na HMM–ima.

Sinteza temeljena konkatenacijom prirodnija je dok se god radi o sintezi govora iz domene iz koje je i baza zvučnih jedinica. Čim se domena mijenja u "nepoznatu" pada i kvaliteta rezultirajućeg sintetizatora govora. Sustavi temeljeni na HMM–u generalno su razumljiviji bez obzira na domenu, ali im je prirodnost relativno niska. To potkrepljuje i Paulin sa suradnicima (2020) te pojašnjava kako problem domene i prirodnosti rješavaju sustavi temeljeni na dubokim neuronskim mrežama.

Međutim, kako je početkom 2010–ih došlo do strelovita razvoja dubokih neuronskih mreža, u Hrvatskoj istraživanja u domeni sinteze govora polako stagniraju. Istraživanja su ograničena na studentske radove (Džijan 2020; Strmečki Stakor 2020; Marić 2020; Lochert 2020) koji zbog same svoje prirode predstavljaju individualne napore koji mogu biti tek početak, a ne zamjena za projekt skupine istraživača. U tablici 2 dostupan je pregled postignuća u sintezi govora za hrvatski jezik poredan kronološki. Primijetiti ćemo kako dominiraju istraživanja temeljena na konkatenativnoj i HMM sintezi, a tek s velikim vremenskim odmakom dolazimo do dubokih neuronskih mreža.

5 Web sjedište Hascheck Voicea više nije aktivno i dostupna je samo preslika zabilježena u arhivi *Internet Archive*

Autori	Tip sustava	Model	Godina
Bakran, Lazić	Konkatenativna sinteza	MBROLA	1998., 2006.
Martinčić–Ipšić, Ipšić	HMM	–	2006.
Pobar, Martinčić–Ipšić, Ipšić	Konkatenativna sinteza	TD–PSOLA	2008.
Šoić	Konkatenativna sinteza	Festival	2010.
Pobar, Ipšić	Konkatenativna sinteza i HMM	Festival	2011.
Šoić, Dembitz	Konkatenativna sinteza	SPICE	2011.
Dunđer	Formantna sinteza i HMM	CroSS	2013.
Džijan, Lochert, Marić, Strmečki–Stakor	Neuronske mreže	Tacotron2	2020.

Tablica 2. Vremenski poredana postignuća u sintezi govora za hrvatski jezik

Unatoč vrlo individualiziranoj prirodi diplomskih radova, prvi sustav za sintezu govora temeljenog na dubokim neuronskim mrežama za hrvatski jezik upravo je bio rezultat diplomskih radova. Ono što se u tim radovima (Džijan 2020; Strmečki Stakor 2020; Marić 2020; Lochert 2020) navodi je možda izvorni razlog zbog čega za hrvatski jezik još nema radova temeljenih na dubokim neuronskim mrežama, a radi se o nedostupnosti materijala za provedbu istraživanja. Dok je Džijan (2020) snimio vlastiti korpus govora trajanja 5 sati i 39 minuta, drugi radovi spominju korpus prikupljen sa snimaka HRT–a za *Hascheck Voice,* a kod Marića (2020) i Strmečki–Stakora (2020) nalazimo podatak o trajanju korištenog korpusa od 3 sata 7minuta i 58 sekundi.

Jedan od javno dostupnih korpus dovoljnih gabarita za izradu sustava za sintezu govora temeljenog na dubokim neuronskim mrežama napravile su Kuvač Kraljević i Hržica (2016) kroz kontinuirani rad od nekoliko godina. *Hrvatski korpus govornog jezika – HrAL* opisan u istoimenom radu, korpus je razgovornog hrvatskog jezika i uključuje ukupno 617 govornika. Problem ovog korpusa je upravo tako veliki broj različitih govornika. Dok trenutno najveći dostupni korpus, ParlaSpeech–HR 2.0 (Ljubešić *et al.* 2024), unatoč svojoj zavidnoj veličini, pati od sličnih problema, no daje priliku da se kvalitetnije snimke jednog govornika izdvoje u zaseban podkorpus. Naime, kako bi se smanjila kompleksnost zadatka, korpus bi trebalo svesti isključivo na jednog govornika koji se snimao u izoliranim uvjetima (Džijan 2020). Ovo je važno u prvim fazama treniranja sustava temeljenog na dubokim neuronskim mrežama. Ovakvi, veliki korpusi primjereniji su sustavima za prepozna-

vanje govora, baš zbog izazovnog zapisa koji se sastoji od šuma, jeke i spontanog govora. Pregled navedenih korpusa zajedno s podacima o njihovoj dostupnosti kao i duljini tonskog zapisa donosimo u nastavku (Tablica 3).

Gledajući sveobuhvatno područje računalne obrade prirodnog jezika, prema svemu izloženom moramo se složiti s Tadićem (2023) da je područje sinteze govora trenutno najslabije zastupljeno područje jezičnih tehnologija u Hrvatskoj i da za sada ne postoji javno dostupno rješenje na industrijskoj razini. No Hrvatska tu nije jedina. U naredna dva poglavlja pokušat ćemo usporediti hrvatski doprinos u području računalne sinteze govora s okolnim zemljama koje su nam jezikom bliske, a potom ćemo vidjeti gdje se nalazimo i u okvirima svjetske scene koja trenutno dominira u ovom području.

Naziv	Trajanje zapisa	Tip pristupa
Korpus Hrvatskoga Govora	19h 24m	zatvoren
HrAL	27h 36m 21s[6]	otvoren
Džijan	5h 39m 00s	zatvoren
HascheckVoice	3h 07m 58s	zatvoren
ParlaSpeech–HR 2.0	3 061h 00m 00s	otvoren

Tablica 3. Korpusi hrvatskog govora sa trajanjem zapisa i tipom pristupa

4. Srodni jezici

Dembitz (2017) tvrdi kako je u nekoliko navrata pokušao ostvariti suradnju s raznim javnim službama u svrhu izrade hrvatskih govorno–tehnoloških rješenja. No, njegova nastojanja nisu pala na plodno tlo. Stoga je već spomenuti *Hascheck Voice* načinjen od svega 17 minuta govora. Za aktualna govorno–tehnološka rješenja koja se koriste u hrvatskom javnom prostoru tvrdi da su potekla iz Slovenije i Srbije. Samim time, vrijedi pokriti i razvoj područja u nama susjednim zemljama jer je i najveća vjerojatnost da ćemo pri izradi govorno–tehnoloških rješenja imati suradnju s najbližim susjedima, što geografski, a što po jezičnoj bliskosti.

4.1. Slovenski jezik

Slično kao i u Hrvatskoj, područje sinteze govora u Sloveniji nije se znatno razvijalo do pojave računalne sinteze govora, no, od tada je razvoj bio munjevit. U manje od 10 godina u Sloveniji je od prvih istraživanja sinteze govora na razini riječi (Weilguny 1993) došlo do potpunih sustava za sintezu govora (Šef *et al.* 1998). U isto vrijeme kada je u Hrvatskoj objavljena difonska baza temeljena na algoritmu MBROLA (Bakran & Lazić 1998), u Sloveniji, pod okriljem Instituta Jožef Štefan

6 Izračun temeljen na zbrajanju trajanja zvučnih zapisa preuzetih s mrežnog sjedišta TalkBank–a za hrvatski jezik (https://ca.talkbank.org/access/Croatian.html)

napravljen je potpuni sustav za sintezu govora: EMA (engl. *EMployment Agent*). Jedna je to od prvih aplikacija koja sadržava sustav za sintezu govora, a bilo ju je moguće koristiti u mrežnom okružju (Šef *et al.* 1998).

Koju godinu kasnije sustav koji se koristio u EMA–i, nakon modifikacija i znatnih poboljšanja u procesu sinteze, dobiva novo ime, GOVOREC, i objavljuje se besplatno za nekomercijalne svrhe (Šef & Gams 2003). Govorno–tehnološko područje u Sloveniji od samih početaka razvija se oko instituta Jožef Štefan, a kasnije se u izgradnju sustava uključuju i privatne tvrtke (Šef 2016). Projekti sinteze govora u Sloveniji primarno se koriste konkatenativnim TD–PSOLA i formantnim HMM metodama. Velik napor ulože se u što bolju prirodnost sustava, ali i njegovu široku primjenjivost. Tako Šef (2016) komentira kako je prilikom razvoja sustava za mobilne uređaje, posebna pažnja dana korištenju izvornih *Android Speech* instrukcija, kako bi integracija bila što jednostavnija.

Najaktualniji sustav u Sloveniji je eBralec 4, koji za sintezu koristi HMM metodu (Žganec Gros *et al.* 2020). Razvijen je u suradnji dvije privatne tvrtke uz Institut Jožef Štefan te u sebi sadrži glasove i iz sustava GOVOREC (eBralec 2020). Od istraživanja u području sinteze govora pomoću neuronskih mreža, jedino dostupno istraživanje bilo je u sklopu srednjoškolskog maturalnog rada ali s nepovoljnim rezultatima (Grebovšek & Vrečer 2020).

Slovensko govorno područje iako relativno malo, primjer je dobre suradnje akademske zajednice i privrede. Ovakvim modelom suradnje razvijanje govorno–tehnoloških rješenja ne mora biti ograničeno na studentske radove ili samostalne pothvate pojedinih istraživača već ima podršku cijele akademske institucije poput, u ovom slučaju, instituta Jožef Štefan.

4.2. Srpski jezik

Istraživanje sinteze govora u Srbiji započelo je nešto kasnije u usporedbi s Hrvatskom i Slovenijom, ali je brzo dobilo zamah kao zaseban projekt na Fakultetu tehničkih nauka Sveučilišta u Novom Sadu. Projekt, poznat kao **AlfaNum**, prerastao je akademske okvire i 2003. godine evoluirao u istoimenu tvrtku koja i danas predstavlja ključnog igrača u razvoju tehnologija za sintezu i prepoznavanje govora, ne samo za srpski jezik već i za hrvatski, crnogorski, bosanski i makedonski (AlfaNum 2003).

U Srbiji, kao i u Sloveniji, suradnja između akademskih institucija i privatnog sektora pokazala se iznimno uspješnom. Unatoč kasnijem početku, Srbija je imala značajnu prednost u odnosu na ostale zemlje u regiji zahvaljujući tvrtki AlfaNum koja je već 2013. godine raspolagala s impresivnim korpusom od 200 sati govora od kojih je otprilike 190 sati iz zvučnih knjiga koje čita 50 spikera, a ostalih 10 sati su studijske snimke još 120 spikera (Delić *et al.* 2013). Danas, taj korpus obuhvaća gotovo 2000 sati precizno anotiranog materijala, što je temelj za razvoj sustava automatskog prepoznavanja govora (ASR).

Za razliku od ASR korpusa, tvrtka AlfaNum također raspolaže s posebno fonemski označenim korpusom za sintezu govora, koji je prvotno sadržavao 4 sata srpskog govora i dodatnih 1 sat i 30 minuta hrvatskog govora. Trenutno stanje govornih baza uključuje resurse[7] za srpski [Danica 3h20m; Snežana 3h20m; Jovan 55m; Stefan 70m] ali i za hrvatski [Ivana 4h50m; Marica 3h30m; Mario 55m] i crnogorski [Mia 30m; Danilo 30m] (osobna komunikacija 2024).

S obzirom na ovo bogatstvo resursa, ne iznenađuje što su se srbijanski istraživači već 2016. godine okrenuli korištenju dubokih neuronskih mreža za postizanje što prirodnije sinteze govora (Delić & Sečujski 2016). U tom kontekstu, zanimljivo je usporediti pristupe u Hrvatskoj (Džijan 2020) i Sloveniji (Grebovšek & Vrečer 2020), gdje su istraživači koristili model *TacoTron* (Wang *et al.* 2017). U tim radovima, autori su zbog velike potrebe za podacima i nužnog predznanja potrebnog za korištenje modela *TacoTron*, navodili kako su kvalitetnije rezultate dobili korištenjem HMM–a. No, Delić i Sečujski (2016) su za sintezu srpskoga jezika temeljenu na neuronskim mrežama odlučili primijeniti *Merlin* (Wu *et al.* 2016) sustav otvorenog koda i ostvarili izvanredne rezultate koji svojom prirodnošću pariraju onima u *TacoTron* radu koji u tom trenutku još nije bio ni objavljen.

Ocjena prirodnosti prvog *TacoTron* modela za engleski jezik bila je 3,82 od maksimalnih 5, dok su Delić i Sečujski svojim modelom za srpski jezik postigli rezultat od 4,54 na istoj skali, koristeći samo 4 sata i 9 minuta govora, u usporedbi s 24 sata materijala koje su autori *TacoTron* modela imali na raspolaganju. Međutim usporedba na toj razini možda i nije u potpunosti korektna jer je *TacoTron* model sustava za sintezu govora *s kraja na kraj*. U ovakvim se sustavima dubinsko poznavanje fonetike i sinteze govora supstituira velikom količinom podataka i time smanjuje mogućnost pogreške zbog eventualnog neznanja istraživača (Wang *et al.* 2017).

S druge strane, Delić i Sečujski (2016) su za izradu njihova sustava koristili dvije neuronske mreže nad fonetski označenim korpusom, te vokalne modele napravljene u posebnom alatu kojeg je napravila i kojeg održava ista skupina istraživača u AlfaNumu. Kroz niz radova, AlfaNum tim je eksperimentirao s različitim vokoderima poput WORLD, WaveRNN i HiFi–GAN, prilagođavajući ih potrebama i resursima, čime su dodatno unaprijedili kvalitetu sinteze govora (Suzić *et al.* 2022).

Značajno je istaknuti kako AlfaNum kontinuirano proširuje svoje tehnološke horizonte ne ograničavajući se samo na sintezu govora na srpskom jeziku. Tako je od verzije 3.0, njihova platforma za sintezu govora obogaćena uslugama za hrvatski jezik, a od verzija 6.0 i za crnogorski jezik (*Demonstracija TTS*, pristupljeno 3.3.2024.). Ovaj multilingvalni pristup odražava njihovu predanost pružanju kvalitetnih usluga sinteze govora širem spektru korisnika. Uz to, istraživački tim AlfaNuma pokazuje visoku razinu aktivnosti, ne samo u razvoju čiste sinteze govora (Suzić *et al.* 2022), već i u istraživanjima usmjerenim na efikasnu sintezu s ograničenim resursima i na višejezičnu sintezu (Nosek *et al.* 2023), što je vidljivo u njihovim nedavnim publikacijama.

7 Vremenici za navedene jezične resurse uključuju i tišinu i odnose se na glasove koji su komercijalno iskorišteni.

4.3. Makedonski jezik

Istraživanja u području sinteze govora u Makedoniji koincidiraju s onima u Sloveniji i Hrvatskoj. Josifovski i suradnici (1997) navode kako je razvoj u području krenuo ranih 90–ih s ciljem pomoći osobama oštećena vida kroz optičko prepoznavanje znakova, njihovo sintetiziranje u govor i ispis brajice. Njihov sustav za sintezu govora bio je zamišljen kao sustav koji se temelji na konkatenaciji slogova u vremenskoj domeni, a sastoji se od slogovne baze s 1 275 zvučnih zapisa. Međutim, prema Gerazovu i suradnicima (2008) sustav nije nikada dovršen već je ostao u fazi koncepta.

Unatoč ranom početku, područje sinteze govora u Makedoniji stagnira do trenutka kada AlfaNum objavljuje rad u kojem se navode govorno–tehnološka rješenja za hrvatski, srpski i makedonski (Delić *et al.* 2006). Iz tog se razloga Gerazov i suradnici (2008) referiraju na prijašnja istraživanja kao nepotpuna jer su se koristila hrvatskom ili srpskom (Delić *et al,* 2006) difonskom bazom za sintezu makedonskog govora. Stoga oni predlažu novu vrstu zvučne jedinice "kvazi–difon" i koriste konkatenativnu sintezu na temelju algoritma TD–PSOLA. U sljedećih nekoliko godina Gerazov se zadržava na difonskoj analizi (Gerazov *et al.* 2008), modeliranju intonacije (Gerazov *et al.* 2010), te generiranju prozodije (Gerazov & Ivanovski 2011), a sve sa svrhom sinteze govora u sustavu "*Speak Macedonian*".

Slično kao i u hrvatskom i slovenskom jeziku, sustav za sintezu makedonskog govora temeljen na neuronskim mrežama pojavljuje se tek 2020. godine. Za razliku od ranije spomenutih jezika, radi se o skupini istraživača koji se okupljaju s ciljem izrade modela za sintezu govora otvorena pristupa. Rezultat istraživanja je MAKEDONKA sustav za sintezu pomoću neuronskih mreža temeljen na otprilike 20 sati govora. Sustav postiže rezultat po MOS skali od 3,93, što je iznimno dobar rezultat uzmemo li u obzir da po istoj skali stvarni govor ima rezultat 4,62 (Mishev *et al.* 2020). S obzirom na predanost skupine autora da produciraju izrazito kvalitetan sustav za sintezu govora, njihov rad dotiče se modela *TacoTron* ali i njegovih mnogih varijacija i nasljednika što ga čini vrlo vrijednim resursom za bilo koje buduće istraživanje u ovom području.

4.4. Bosanski jezik

Analizirajući prostorno i jezično bliski bosanski jezik, postaje jasno da je on u kontekstu istraživanja sinteze govora znatno manje prisutan u regionalnim studijama. Iako su radovi iz tvrtke AlfaNum pružili temelj, osim njih, istaknut je samo jedan znanstveni rad (Mujčić 2005) koji se bavi problematikom sinteze bosanskog govora, a koji sebe smatra začetnikom u ovom polju.

Taj je rad rezultirao razvojem sustava za sintezu govora koji se oslanja na konkatenativnu metodu unutar softverskog okruženja *MatLab*. Autor (Mujčić 2005) ističe niz izazova koji zahtijevaju daljnje analize i rješenja, međutim, prema dostupnim informacijama, čini se da nije došlo do značajnijeg napretka ili novih istraživanja koja bi nadogradila ove početne nalaze. Ovo ukazuje na široko polje mo-

gućnosti za buduće istraživače koji su voljni proširiti granice znanja i tehnologije u sferi sinteze bosanskog jezika, te tako doprinijeti razvoju ovog važnog segmenta lingvističke tehnologije.

Na kraju ovog poglavlja donosimo i usporednu analizu resursa među srodnim nam jezicima (Tablica 4). Ono što se može primijetiti iz tablice je jasna evolucija po godinama objava modela od konkatenativnih pristupa prema neuronskim mrežama. Takva situacija korelira sa činjenicom da su smanjeni tehnički zahtjevi i prepreke za treniranje dubokih neuronskih mreža, a čiji rezultati su nerijetko razumljiviji i prirodniji od pristupa konkatenacijom ili HMM–ima.

Jezik	Naziv /Autor	Tip sustava	Model	Veličina korpusa	Godina
slovenski	EMA	Konkatenativna sinteza	TD–PSOLA	–	1998.
	GOVOREC	Konkatenativna sinteza	TD–PSOLA	–	2003.
	eBralec	HMM	–	–	2020.
srpski hrvatski crnogorski	AlfaNum	Konkatenativna sinteza	TD–PSOLA	2000 h – asr	2006.
		Neuronske mreže	Merlin, interni program	32h 50m srpski, 9h 15m hrvatski, 1h crnogorski	do sada
makedonski	*Josifovski et al.*	Konkatenativna sinteza	–	1 275 slogova	1997
	Gerazov et al.	Konkatenativna sinteza	TD–PSOLA	–	2008–2011.
	MAKEDONKA	Neuronske mreže	–	20 sati	2020.
bosanski	*Mujčić*	Konkatenativna sinteza	–	–	2005.

Tablica 4. Usporedna analiza srodnih jezika

5. Nesrodni jezici

Da bi se očuvala fokusiranost i jasnoća ovog rada, ograničit ćemo pregled na metode sinteze govora koje koriste neuronske mreže, s posebnim naglaskom na suvremena istraživanja u jezicima koji nisu srodni našem, poput engleskog i manda-

rinskog. Cilj nam je istražiti najnovija postignuća u području sinteze govora temeljene na dubokim neuronskim mrežama, koja će poslužiti kao putokaz za buduće inovacije.

5.1. Engleski jezik

Uzimajući u obzir do sada spomenuta istraživanja za srodne jezike provedena na modelu *TacoTron* (Wang *et al.*, 2017), posebno ćemo se osvrnuti na engleski jezik, za koji je model prvotno razvijen. Većina pionirskih istraživanja u vidu sinteze govora temeljene na neuronskim mrežama odvijala se naime za engleski jezik. Tome su prvenstveno pridonijeli količina dostupnih resursa kao i jednostavnost engleskog jezika (Ning *et al.* 2019).

U području sinteze govora temeljene na dubokim neuronskim mrežama, prema taksonomiji koju predlažu Tan i suradnici (2021) razlikujemo četiri ključne kategorije: **tekstualnu analizu**, **akustične modele**, **vokodere** i **sustave s kraja na kraj**.

Tekstualna analiza odnosi se na obradu ulaznog teksta kako bi se generirali odgovarajući akustični znakovi za sintezu govora. To uključuje različite tehnike kao što su grafemsko fonemsko mapiranje, prozodijsko obilježavanje i segmentacija teksta. Istraživači s primarnim fokusom na prirodnost govora koriste duboke neuronske mreže tek kao jedan korak u izradi sustava za sintezu govora, kao što su to radili Delić i Sečujski (2016) za srpski jezik.

Akustični modeli koriste se za predviđanje akustičnih značajki na temelju ulaznog teksta koristeći tehnike poput HMM–a ili DNN–a (duboke neuronske mreže) kako bi naučili statističke veze između teksta i zvuka. Među poznatijim engleskim radovima navodimo "WaveNet: A Generative Model for Raw Audio" na kojem je radio van den Oord sa svojim suradnicima (2016). Prirodnost takvog sustava znatno je viša od čiste konkatenativne ili parametarske sinteze. Van den Oord i suradnici (2016) prijavljuju rezultat od 4,21 po MOS ljestvici u odnosu na 4,55 koliko su ispitanici ocijenili prirodni govor[8] te 3,86 za konkatenativnu sintezu i 3,67 za parametarsku sintezu.

Vokoderi su zaduženi za generiranje konačnog zvuka na temelju prediktivnih akustičkih značajki. Često se temelje na tehnologijama poput konkatenativne sinteze ili parametarske sinteze. Kako bi se smanjila potreba za širokim poznavanjem materije i smanjio nužan broj koraka da se dobije konačni sintetizator govora, Wang i suradnici (2017) predlažu metodu sinteze govora **s kraja na kraj**. Odnosno, zaobilazi se treniranje posebnih vokodera, grafemsko fonemskih modela i prozodijskog obilježavanja teksta, te se sav taj posao delegira neuronskoj mreži. Sve što se od znanstvenika traži jest priprema dovoljne količine ulaznog objekta u obliku parova tekst/govor (24,6 sati u originalnom radu). Nakon treninga, sustav je spre-

8 Vrijedi spomenuti kako se prilikom provedbe anketnog ispitivanja za dobivanje MOS rezultata testira i prirodni govor koji je dio govornog korpusa, a koji je izostavljen iz seta za treninranje sintetizatora. Na taj način se kvaliteta sintetizatora može provjeriti na govoru koji "nije viđen".

man za upotrebu sa zadovoljavajućim rezultatima. Već spomenuta, prva iteracija modela TacoTron, po tom je principu postigla rezultate prirodnosti od 3,82 prema MOS ljestvici.

Unatoč nižoj ocjeni prirodnosti u usporedbi s rezultatom od 4,21 po MOS–u koji je naveden za WaveNet model, TacoTron je olakšao pristup području sinteze govora temeljene na dubokim neuronskim mrežama, omogućujući čak i srednjoškolcima (Grebovšek & Vrečer, 2020) i studentima (Džijan, 2020) bez formalnog obrazovanja iz fonetike da se upuste u istraživanja u ovom području. I dok neka gotova aplikacijska rješenja poput paketa Merlin (Wu *et al.*, 2016) iziskuju prolaženje kroz nekolicinu naprednih koraka kako bi se osigurali svi preduvjeti za početak sinteze govora, možemo reći da su dostupnost i pristupačnost TacoTron modela značajno pridonijele demokratizaciji sinteze govora temeljene na dubokim neuronskim mrežama, otvarajući vrata novim generacijama istraživača.

Ne čudi stoga da su noviji radovi u sferi sinteze govora temeljene na dubokim neuronskim mrežama za engleski jezik preuzeli hibridni pristup. Shen i suradnici (2018) predstavljaju model *TacoTron 2* koji uzima najbolje karakteristike iz prve inačice *TacoTron* modela i pridružuje ih metodama korištenim prilikom izrade *WaveNet* vokodera. Zahvaljujući tom pristupu ostvaruju rezultat od 4,526 prema MOS skali u odnosu na 4,582 tada ocijenjenog stvarnog govora.

Međutim, autori (Shen *et al.* 2018) naglašavaju problem vremena nužnog za sintezu govora pomoću neuronskih mreža i tehničke zahtjeve koji još uvijek koče njihovu širu primjenu. U tom pogledu razvijaju se sustavi koji se bave prilagodbom postojećih modela za druge jezike (Zhao *et al.* 2021) ili oni koji nastoje smanjiti vrijeme nužno za izradu sama sustava (Ren *et al.* 2019). S druge strane razvijaju se i sustavi poput *Deep Voice*–a koji je prošao kroz 3 iteracije u zadnjih nekoliko godina i postavio pritom rekord u veličini podatkovnog seta za trening treće inačice s preko 800 sati govora i 2000 govornika (Ping *et al.* 2019).

Ono što vrijedi primijetiti je kako se u engleskom sve češće i češće u istraživačkim radovima uz akademske institucije navode i privatne tvrtke kao što su Google, Mozilla, nVidia, Baidu i slične (Valle *et al.* 2021; Shen *et al.* 2018; Ping *et al.* 2019; Zhao *et al.* 2021). Tako visoko zasićenje područja sinteze govora temeljenom na dubokim neuronskim mrežama i jest razlog njegova strelovitog razvoja na engleskom govornom području.

5.2. Mandarinski kineski

U kineskom kao tonskom jeziku, bitan faktor kod interpretacije značenja pojedinih riječi je intonacija, što je ponekad slučaj i u hrvatskom jeziku (Lazić 2006). Interpretiranje intonacije time povećava kompleksnost problema, kako kod sustava za prepoznavanje govora (Petrinović 2009) tako i kod sustava za sintezu govora. Ne čudi stoga da je, uz engleski jezik, mandarinski kineski često naveden, u nekim prominentnijim radovima vezanim za sintezu govora. Navodi se kao jezik visoke

razine kompleksnosti, ali i kao dobar test kvalitete sintetizatora govora (van den Oord *et al.* 2016).

Prvi *WaveNet* osim engleskog MOS rezultata predstavlja i rezultate za mandarinski kineski. Unatoč visokoj kompleksnosti jezika, prirodnost sintetizatora mandarinskog temeljenog na dubokim neuronskim mrežama je 4,02 naspram 4,25 za stvarni govor, a 3,79 za parametarsku te 3,47 za konkatenativnu sintezu (van den Oord *et al.* 2016).

Međutim, učinkovitost *TacoTron* modela i njegova široka primjenjivost dovela ga je i do mandarinskog. Yang i suradnici (2019) predstavili su model za sintezu mandarinskog jezika *s kraja na kraj* uz poboljšanja namijenjena prirodnosti koja parira sustavima u kojima je obvezno označavanje prozodije u tekstu. Usporedna ocjena sustava u radu prezentirana je korištenjem MOS ljestvice. Prva inačica modela *TacoTron* dobila je ocjenu 3,76, inačica s prozodijskim obilježjima dobila je ocjenu 3,89, dok su predloženi modeli *SAE–TacoTron* i *SAG–TacoTron* dobili ocjene 3,83 i 3,87, redom, dok je testni prirodni govor ocijenjen sa 4,21. Iako su dobiveni rezultati nešto niži nego kod originalnog *WaveNet*–a, treba napomenuti da se radi o sintezi temeljenoj isključivo na jednostavnom tekstualnom prikazu kineskih znakova, bez dodatnih prozodijskih obilježja.

Jian Zhu (2019) s druge strane fokusira svoje napore na *TacoTron 2* modelu i unaprijed treniranoj dubokoj neuronskoj mreži BERT za poboljšanje automatskog prepoznavanja prozodije i koartikulacije. U radu testira obični *TacoTron 2* model i model s prilagodbom za korištenje BERT–a. Rezultati po MOS ljestvici običnom modelu daju 3,65, a prilagođenom modelu 4,04 dok je stvarni govor ocijenjen sa 4,39. Vidljivo je osjetno unaprjeđenje u odnosu na ranije navedene rezultate, pogotovo kada uzmemo u obzir rezultate parametarske i konkatenativne sinteze.

5.3. Drugi jezici i eksperimenti

Ne uživaju svi svjetski jezici veliko bogatstvo jezičnih resursa u vidu označenih govornih korpusa u trajanju nekoliko sati poput engleskog LjSpeech (Ito & Johnson 2017) ili kineskog Aishell–3 (Shi et al, 2021). Neke istraživačke skupine bave se isključivo primjenom metoda sinteze govora, i drugih jezičnih zadataka, pomoću neuronskih mreža u domeni jezika s oskudnim jezičnim resursima. To je i od posebna interesa za jezike Europske unije jer se prema zadnjem istraživanju META–NETa pokazalo da od 30 proučenih jezika, 21 jezik ima slabe ili nikakve šanse preživjeti digitalnu eru zbog manjkavih ili nepostojećih digitalnih jezičnih resursa (Meta–Net 2018).

Zhang i suradnici (2019) u svom radu navode sustav temeljen na modelu *TacoTron 2* kojeg treniraju na engleskom, španjolskom i kineskom jeziku, ali i prilagođavaju trenirane modele s jednim jezikom na drugi kako bi smanjili potrebu za jezičnim resursima. Rezultati prirodnosti su i više nego zadovoljavajući. Početni engleski model, a odredišni španjolski postiže prirodnost od 4,20, odnosno 3,94

za mandarinski. Početni španjolski model postiže prirodnost od 4,28 za engleski i 3,85 za kineski, a kineski model ocijenjen je sa 4,49 za engleski i 4,56 za španjolski.

Slične pretpostavke testiraju Tu i suradnici (2019) navodeći sustav *s kraja na kraj* za jezike oskudnih jezičnih resursa (engl. *low–resource langauges*). U radu se engleski model treniran na 100 sati govora smatra izvornikom, dok su odredišni jezici (mandarinski kineski, njemački i francuski) trenirani na 30, 25 i 15 minuta materijala. Test prirodnosti proveden je isključivo za kineski jezik u varijanti treniranoj s 25 minuta i 15 minuta materijala za nekoliko modela. Najbolji model uspijeva postići rezultat prirodnosti od 4,01 s 25 minuta i 3,48 s 15 minuta materijala. Možemo reći da su to izrazito obećavajući rezultati za jezike s oskudnom količinom dostupnih resursa.

Xu i suradnici (2020) predlažu sustav za sintezu govora s kraja na kraj, za jezike s izrazito oskudnom količinom resursa (engl. *extremely low–resource languages*), sličan ranije navedenim pristupima s engleskim početnim modelom. Unatoč činjenici da Litavski korpus *Liepa* sadrži preko 100 sati govora (Laurinčiukaitė *et al.* 2018), eksperimenti u radu provedeni su s 3,7 minuta i 1,29 sati govora. Rezultati prirodnosti litavskog govora za model temeljen na 1,29 sati su 3,65 u odnosu na 4,01 prirodna govora. Rekli bi smo da su ovo obećavajući rezultati za sve istraživače koji se namjeravaju baviti sintezom govora u domeni jezika s oskudnom količinom jezičnih resursa, kao što je i hrvatski.

6. Zaključak

Ovaj rad predstavlja temeljit pregled evolucije sinteze govora, s posebnim osvrtom na hrvatski jezik i njegovu povezanost s drugim slavenskim jezicima poput slovenskog, bosanskog, srpskog, makedonskog, te pruža uvid u globalne trendove i inovacije u ovom području. Analizirajući povijesni razvoj i suvremene pristupe, rad otkriva kako se sinteza govora transformirala od ranih eksperimenata do naprednih računalnih modela.

U prvom dijelu, rad istražuje pionirske pothvate u području sinteze govora u Hrvatskoj, ističući kako su rane tehnike i metode oblikovale temelje za današnje sustave. Ovaj retrospektivni uvid pruža ne samo poštovanje prema inicijalnoj inventivnosti znanstvenika već i dublje razumijevanje kako su te rane ideje evoluirale u napredne alate koji oblikuju moderne sustave. U širem kontekstu, rad ističe značajne doprinose dubokih neuronskih mreža u generiranju visokokvalitetnog govora, posebice za jezike s obiljem jezičnih resursa (u ovom slučaju, primarno tonskim zapisima), poput engleskog i kineskog, ali i za jezike s oskudnijim jezičnim resursima poput njemačkog, francuskog ili pak litavskog.

Međutim, usporedba hrvatske situacije s globalnim dostignućima ukazuje na izazove s kojima se suočava domaća istraživačka zajednica, uključujući nedosljednost terminologije i nedostatak jezičnih resursa. Dok druge zemlje poput Litve

ulažu u razvoj jezičnih resursa, Hrvatska se suočava s rizikom ovisnosti o stranim tehnološkim rješenjima.

Kao odgovor na navedene izazove, doktorska disertacija, u koju nas ovaj pregledni rad uvodi, ima za cilj razviti označeni korpus i modele dubokog učenja prilagođene hrvatskom jeziku. Ovaj pristup ne samo da će potaknuti napredak u sintezi govora na hrvatskom jeziku, već će i osigurati njegovu održivost i neovisnost u digitalnoj budućnosti.

Literatura

AlfaNum (18.6.2003.). ALFANUM, dostupno na http://alfanum.ftn.uns.ac.rs, pristupljeno 1.3.2021.

Bakran Juraj i Nikolaj Lazić (1998). Fonetski problemi difonske sinteze hrvatskoga govora. *Govor*, *Vol. 15 No. 2,* 103–114.

Bakran, Juraj i Damir Horga (1996). SAMPA za hrvatski. *Govor XIII, 1–2,* 99–104.

Charpentier, Francis i Stella Michael (1986) Diphone synthesis using an overlap–add technique for speech waveforms concatenation, ICASSP '86. *IEEE International Conference on Acoustics, Speech, and Signal Processing, Tokyo*, Japan, 1986, pp. 2015–2018, doi: 10.1109/ICASSP.1986.1168657. .

Delić, Tijana i Milan Sečujski (2016). Sinteza govora na srpskom jeziku zasnovana na veštačkim neuralnim mrežama, *24th Telecommunications forum TELLFOR 2016*, Beograd, Srbija, 22–23.11.2016.

Delić, Vlado, Milan Sečujski, Darko Pekar, Nikša Jakovljević i Dragiša Mišković (2006). A Review of AlfaNum Speech Technologies for Serbian, Croatian and Macedonian, *IS–LTC 06*, Ljubljana, Slovenia, 9–10.10.2006.

Delić, Vlado, Milan Sečujski, Nikša Jakovljević, Darko Pekar, Dragiša Mišković, Branislav Popović, Stevan Ostrogonac, Milana Bojanić i Dragan Knežević (2013). Speech and Language Resources within Speech Recognition and Synthesis Systems for Serbian and Kindred South Slavic Languages. In: Železný M., Habernal I., Ronzhin A. (eds.) *Speech and Computer. SPECOM 2013.* Lecture Notes in Computer Science, vol 8113. Springer, Cham. https://doi.org/10.1007/978–3–319–01931–4_42.

Dembitz, Šandor (2017) Strojna obrada hrvatskog jezika – mađarski doprinosi, *Kolo: časopis Matice hrvatske, 4,* 108–122.

Demonstracija TTS. Pristupljeno 03. Ožujak 2024., od https://www.alfanum.co.rs/index.php/sr/demonstracija/demonstracija–tts

Dunđer, Ivan (2013). CroSS 2.0: Croatian Speech Synthesizer. CroSS 2.0: Croatian Speech Synthesizer. Računalni programski paket.

Dutoit, Thierry (1994). High quality text–to–speech synthesis: a comparison of four candidate algorithms, *Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing*, Adelaide, SA, Australia, pp. I/565–I/568 vol.1, doi: 10.1109/ICASSP.1994.389231.

Dutoit, Thierry, Vincent Pagel, Nicolas Pierret, Francois Bataille i Olivier van der Vrecken (1996). The MBROLA project: towards a set of high quality speech synthesizers free of

use for non commercial purposes, *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Philadelphia, PA, USA, 1996, pp. 1393–1396 vol.3, doi: 10.1109/ICSLP.1996.607874.

Džijan, Matej (2020). Sinteza govora iz teksta upotrebom dubokog učenja, diplomski rad, Fakultet elektrotehnike, računarstva i informacijskih tehnologija, Osijek.

eBralec 4 (2020). dostupno na https://ebralec.si, (pristupljeno 27.2.2021.).

Gerazov, Branislav i Zoran Ivanovski (2009). Diphone Analysis of the Macedonian Language for the Purpose of Text–To–Speech Synthesis, ICEST 2009, Veliko Tarnovo, Bugarska.

Gerazov, Branislav Zoran Ivanovski (2011). Prosody Generation Module for Macedonian Text–to–Speech Synthesis, *AES 130*, London, Velika Britanija.

Gerazov, Branislav, Goce Shutinoski i Goce Arsov (2008) A Novel Quasi–Diphone Inventory Approach to Text–To–Speech Synthesis. *MELECON 2008 – The 14th IEEE Mediterranean Electrotechnical Conference*, Ajaccio, France, 2008, pp. 799–804, doi: 10.1109/MELCON.2008.4618533.

Gerazov, Branislav, Zoran Ivanovski i Ružica Bilibajkić (2010). Modeling Macedonian intonation for text–to–speech synthesis, *DOGS 2010*, Iriski Venac, Srbija.

Grebovšek, Nik i Marko Vrečer (2020). Sinteza govora, završni rad, Srednja škola za kemiju, elektrotehniku i računarstvo, Celje.

Gupta, K., Gupta, D. (2016). An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system. 2016 6th International Conference – Cloud System and Big Data Engineering (Confluence), 493–497. https://doi.org/10.1109/CONFLUENCE.2016.7508170

Hascheck Voice (2011). dostupno na http://hascheck.tel.fer.hr/voice/, (pristupljeno 24.2.2021.).

Ito Keith i Linda Johnson (2017). The LJ speech dataset, dostupno na: https://keithito.com/LJ–Speech–Dataset, (pristupljeno 4.3.2021.).

Josifovski, Ljubomir, Dragan Mihajlov i Dejan Gjorgjevikj (1997). Speech Synthesizer Based on Time Domain Syllbale Concatenation, *SPECOM*, 27–30.10.1997.

Klatt, Dennis (1982). The Klattalk text–to–speech conversion system, Acoustics,Speech, and Signal Processing, *IEEE International Conference on ICASSP'82,7,* IEEE, pp. 1589–1592.

Kuligowska, Karolina, Paweł Kisielewicz i Aleksandra Włodarz (2018). Speech synthesis systems: Disadvantages and limitations. *International Journal of Engineering and Technology (UAE).* 7. 234–239. 10.14419/ijet.v7i2.28.12933.

Kuvač Kraljević, Jelena i Gordana Hržica (2016). Hrvatski korpus govornog jezika (HrAL), *FLUMINENSIA: časopis za filološka istraživanja*, Vol. 28 No. 2. 87–102.

Laurinčiukaitė, S., Telksnys, L., Kasparaitis, P., Kliukienė, R., & Paukštytė, V. (2018). Lithuanian Speech Corpus Liepa for Development of Human–Computer Interfaces Working in Voice Recognition and Synthesis Mode. Informatica, 29(3), 487–498. https://doi.org/10.15388/Informatica.2018.177

Lazić, Nikolaj (2006). Modeliranje strojnih postupaka za izgovaranje teksta pisanoga hrvatskim jezikom. Zagreb: Sveučilište u Zagrebu, Filozofski fakultet.

Lochert, Karlo (2020). Primjena modela WaveGlow za strojnu tvorbu hrvatskoga govora (Završni rad). Zagreb: Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva. Preuzeto s https://urn.nsk.hr/urn:nbn:hr:168:781677

Ljubešić, N., Koržinek, D., & Rupnik, P. (2024). Parliamentary spoken corpus of Croatian ParlaSpeech–HR 2.0.
https://www.clarin.eu/content/parlamint–towards–comparable–parliamentary–corpora.
https://www.clarin.si/repository/xmlui/handle/11356/1914

Marić, Eduardo (2020). Primjena umjetnih neuronskih mreža za strojnu tvorbu hrvatskoga govora (Diplomski rad). Zagreb: Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva. Preuzeto s https://urn.nsk.hr/urn:nbn:hr:168:173160

Martinčić–Ipšić Sanda i Ivo Ipšić (2003). VEPRAD: A Croatian speech database of weather forecasts. U: Budin, L., Lužar–Stiffler, V., Bekić, Z. & Hljuz–Dobrić, V. (ur.) *25th International Conference on Information Technology Interfaces, ITI 2003. proceedings,* 321–326.

Martinčić–Ipšić Sanda i Ivo Ipšić (2006). Context–Dependent Acoustic Modelling of Croatian Speech. *Proceedings of the 9th International Multiconference Information Society IS 2006,* 9th–10th October 2006, Ljubljana, Slovenia, 251–256.

Martinčić–Ipšić Sanda, Mihaela Matešić i Ivo Ipšić (2004). Korpus hrvatskoga govora, *Govor, Vol. 21 No. 2*, 135–150.

Martinčić–Ipšić, Sanda i Ivo Ipšić (2006b). Croatian HMM Based Speech Synthesis. *28th International Conference on Information Technology Interfaces, ITI 2006,* Cavtat, Croatia, 251–256.

Meta–Net (2018). White Paper Series: Press Release, dostupno na: http://www.meta–net.eu/whitepapers/press–release, (pristupljeno, 10.3.2021.).

Mishev, Kostadin, Aleksandra Karovska Ristovska, Dimitar Trajanov, Tome Eftimov i Monika Simjanoska (2020). MAKEDONKA Applied Deep Learning Model for Text–to–Speech Synthesis in Macedonian Language, *Applied Sciences 10(19), 6882,* https://doi.org/10.3390/app10196882.

Moulines, Eric i Francis Charpentier (1989). Pitch Synchronous waveform Processing techniques for Text–To–SpeechSynthesis using diphones, *Speech Communication, Vol. 9, n°5–6*, 453–467, https://doi.org/10.1016/0167–6393(90)90021–Z.

Ning, Yishuang, Sheng He, Zhiyong Wu, Chunxiao Xing i Liang–Jie Zhang (2019). A Review of Deep Learning Based Speech Synthesis *Appl. Sci. 9, no. 19*: 4050. https://doi.org/10.3390/app9194050

Nosek, T., Suzić, S., Delić, V., & Sečujski, M. (2023). Cross–lingual Text–to–Speech with Prosody Embedding. *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, 1–5. https://doi.org/10.1109/IWSSIP58668.2023. 10180259

Paulin, Goran, Marina Ivašić–Kos i Ivo Ipšić (2020). Mogućnost primjene govora u računalnim igrama temeljenim na lokaciji, *Govor, Vol 37., No. 1*, 2020, 31–59, https://doi.org/10.22210/govor.2020.37.02.

Petrinović, Davor (2009). *Uvod u digitalnu obradbu govora korištenjem Matlaba*, Udžbenici Sveučilišta u Zagrebu, Fakultet Elektrotehnike i Računarstva.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). Deep Voice 3: Scaling Text–to–Speech with Convolutional Sequence Learning (arXiv:1710.07654). arXiv. http://arxiv.org/abs/1710.07654

Pobar Miran i Ivo Ipšić (2011). Development of Croatian unit selection and statistical parametric speech synthesis, *34th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2011): Proceedings. Vol. 3 : Computers in technical systems. Intelligent systems*, Bogunović, Nikola i Slobodan Ribarić (eds.), Rijeka: Croatian Society for Information and Communication Technology, Electronics and Microeletronics – MIPRO, 2011, 306–311.

Pobar, Miran (2014). Sinteza hrvatskoga govora utemeljena na odabiru jedinica i stohastičkim modelima, (Disertacija) Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva, https://urn.nsk.hr/urn:nbn:hr:168:714523 (pristupljeno: 03.06.2023.).

Pobar, Miran, Sanda Martinčić–Ipšić i Ivo Ipšić (2008). Računalni sustav za tvorbu hrvatskoga govora. *Engineering review : znanstveni časopis za nove tehnologije u strojarstvu, brodogradnji i elektrotehnici*, 28 (2008), 2; 31–44.

Ren, Yi, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao i Tie–Yan Liu (2019). Fastspeech: Fast, robust and controllable text to speech. *ArXiv:1905.09263* [Cs, Eess]. http://arxiv.org/abs/1905.09263.

Schroeder, Manfred (1993). A Brief History of Synthetic Speech. *Speech Communication vol.13*, 231–237.

SciForce (13.2.2020.) Text–to–Speech Synthesis: an Overview, dostupno na https://medium.com/sciforce/text–to–speech–synthesis–an–overview–641c18fcd35f, (pristupljeno 15.2.2021.).

Sečujski, Milan, Radovan Obradović, Darko Pekar, Ljubomir Jovanovi Vlado Delić (2002). AlfaNum System for Speech Synthesis in Serbian Language. U Sojka, P., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue. TSD 2002. Lecture Notes in Computer Science(), vol 2448. Springer, Berlin, Heidelberg, 237–244. https://doi.org/10.1007/3–540–46154–X_32.

Sečujski, Milan, Vlado Delić, Darko Pekar, Radovan Obradović i Dragan Knežević (2007). An Overview of the AlfaNum Text–to–Speech Synthesis System, *Proceedings of 12th SPECOM (Speech and Computer)*, ISBN 6–7452–0110–x, Moscow, Russia, A3–A6.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry–Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis i Yonghui Wu (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. ArXiv:1712.05884 [Cs]. http://arxiv.org/abs/1712.05884.

Shi, Y., Bu, H., Xu, X., Zhang, S., & Li, M. (2021). AISHELL–3: A Multi–speaker Mandarin TTS Corpus and the Baselines (arXiv:2010.11567). arXiv. https://doi.org/10.48550/arXiv.2010.11567

Story, Brad (2019). *History of speech synthesis*. The Routledge Handbook of Phonetics, W. Katz & P. Assmann, Eds., Routledge, 9–32.

Strmečki Stakor, Lobel (2020). Primjena metoda strojnog učenja za tvorbu hrvatskoga govora (Diplomski rad). Zagreb: Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva. Preuzeto s https://urn.nsk.hr/urn:nbn:hr:168:519336

Suzić, S. (2019). *Parametarska sinteza ekspresivnog govora* (Doktorska disertacija, Sveučilište u Novom Sadu, Srbija).

Suzić, S., Pekar, D., Sečujski, M., Nosek, T., & Delić, V. (2022). HiFi–GAN based Text–to–Speech Synthesis in Serbian. *2022 30th European Signal Processing Conference (EUSIPCO)*, 1178–1182. https://doi.org/10.23919/EUSIPCO55093.2022.9909922

Šef, Tomaž (2016). Sinteza slovenskega govora na mobilni platformi Android, u *Slovenska konferenca o umetni inteligenci: zbornik 19. mednarodne multikonference Informacijska družba – IS 2016*, 12. 10. 2016, [Ljubljana, Slovenija]: zvezek A, 48–51.

Šef, Tomaž i Matjaž Gams (2003). SPEAKER (GOVOREC) A complete Slovenian Text–to Speech System, *Interlational Journal Of Speech Technology, 6*, 277–287.

Šef, Tomaž, Ales Dobnikar i Matjaž Gams (1998). Improvements in slovene Text–To–Speech synthesis, *The 5th International Conference on Spoken Language Processing*, Sydney Australia, doi: 10.21437/ICSLP.1998–42.

Šoić, Renato (2010). Sinteza hrvatskog govora uporabom sustava Festival (diplomski rad). Zagreb: Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva.

Šoić, Renato i Šandor Dembitz (2011). Automatsko prepoznavanje i sinteza govora – Mogućnosti sustava *SPICE*, *Govor, Vol 27 No. 2*, 2010, 145–158.

Tadić, Marko (2023). Language Report Croatian. U G. Rehm & A. Way (Eds.), *European Language Equality: A Strategic Agenda for Digital Language Equality,* 111–114). doi:10.1007/978–3–031–28819–7_9.

Tan, X., Qin, T., Soong, F., & Liu, T.–Y. (2021). A Survey on Neural Speech Synthesis (arXiv:2106.15561). arXiv. https://doi.org/10.48550/arXiv.2106.15561

Thompson, Neil C., Kristjan H. Greenewald, Lee Keeheon i Gaberiel F. Manso (2020). The Computational Limits of Deep Learning. *ArXiv*, abs/2007.05558.

Tu, T., Chen, Y.–J., Yeh, C., & Lee, H. (2019). End–to–end Text–to–speech for Low–resource Languages by Cross–Lingual Transfer Learning. http://arxiv.org/abs/1904.06508

Valle, Rafael, Kevin Shih, Ryan Prenger i Bryan Catanzaro (2020). Flowtron: An autoregressive flow–based generative network for text–to–speech synthesis. *ArXiv*:2005.05957 [Cs, Eess]. http://arxiv.org/abs/2005.05957.

van den Oord, Aaron, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior i Koray Kavukcuoglu (2016). WaveNet: A generative model for raw audio, *arXiv*:1609.03499.

Wang, W., Xu, S., & Xu, B. (2016). First Step Towards End–to–End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention. 2243–2247. https://doi.org/10.21437/Interspeech.2016–134

Weilguny, Simon (1993). Grafemsko–fonemski modul za sintezo izoliranih ebsed za sintezo slovenskega jezika. (Diplomski rad) Sveučilište u Ljubljani, Slovenija.

Wu, Z., Watts, O., King, S. (2016) Merlin: An Open Source Neural Network Speech Synthesis System. Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), 202–207, doi: 10.21437/SSW.2016–33

Xu, J., Tan, X., Ren, Y., Qin, T., Li, J., Zhao, S., & Liu, T.–Y. (2020). LRSpeech: Extremely Low–Resource Speech Synthesis and Recognition (arXiv:2008.03687). arXiv. https://doi.org/10.48550/arXiv.2008.03687

Yang, F., Yang, S., Zhu, P., Yan, P., & Xie, L. (2019). Improving Mandarin End–to–End Speech Synthesis by Self–Attention and Learnable Gaussian Bias. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 208–213. https://doi.org/10.1109/ASRU46091.2019.9003949

Zhao, S., Wang, H., Nguyen, T. H., & Ma, B. (2021). Towards Natural and Controllable Cross–Lingual Voice Conversion Based on Neural TTS Model and Phonetic Posteriorgram (arXiv:2102.01991). arXiv. https://doi.org/10.48550/arXiv.2102.01991

Zhu, Jian (2019). *Probing the phonetic and phonological knowledge of tones in Mandarin TTS models* (arXiv:1912.10915). arXiv. http://arxiv.org/abs/1912.10915

Žganec Gros, J., Romih, M., & Šef, T. (2020). *eBralec 4: Hibridni sintetizator slovenskega govora* (V. Pejović, M. Kljun, V. Groznik, D. Šoberl, K. Čopič Pucihar, B. Blažica, J. Žabkar, M. Pesek, J. Guna, & S. Kolmanič, Ur.; str. 17–20). Institut „Jožef Stefan“. https://plus.cobiss.net/cobiss/si/sl/bib/34882563

Echoes of innovation: journey through the Croatian speech synthesis landscape

As digital communication becomes increasingly prevalent, the development of speech synthesis systems for Croatian and related languages is of paramount importance. This paper provides an in–depth exploration into the field of speech synthesis, emphasizing the Croatian language. It chronologically charts the evolution of speech synthesis from its mechanical inception to the modern electronic age, culminating in an analysis of contemporary landscape of digital speech synthesis systems.

The study commences with a synthesis of previous research on Croatian speech synthesis, scrutinizing the methodologies and strategies implemented, and evaluating their effectiveness, constraints, and results. A comparative study is also presented, assessing advancements in related Slavic languages, including Serbian, Slovene, Bosnian, and Macedonian.

The discourse then widens to include the global landscape of speech synthesis. It highlights the latest breakthroughs, particularly cutting–edge techniques, frameworks, and algorithms that have yielded significant outcomes in languages with abundant linguistic resources, such as English and Mandarin Chinese. This comparison elucidates the notable gaps in speech synthesis progress on a global scale.

The paper also addresses the challenges posed by the scarce and suboptimal quality digital linguistic resources available in Croatia, which hinder the development of speech synthesis. In response to these challenges, the paper introduces a doctoral thesis dedicated to creating an annotated corpus and formulating deep learning models specifically tailored for Croatian speech synthesis. The ambition of this scholarly work is to catalyze advancement, remedy existing shortcomings, and pave the way for a robust future for Croatian speech synthesis technology.

In conclusion, this survey examines both the historical trajectory and the present state of speech synthesis in Croatian. It underscores the criticality of ongoing research in this area and the urgent necessity for enhanced linguistic resources and innovative methodologies. The paper also briefly touches upon the significant progress in speech synthesis for globally dominant languages, such as English and Mandarin Chinese, providing a benchmark for future investigations.

Ključne riječi: sinteza govora, povijesni razvoj, hrvatski jezik
Key words: speech synthesis, historical trajectory, Croatian language