ПРОГРЕС У РАЗВОЈУ И ПРИМЕНИ ГОВОРНИХ ТЕХНОЛОГИЈА У ЕРИ ВЕШТАЧКЕ ИНТЕЛИГЕНЦИЈЕ

PROGRESS IN THE DEVELOPMENT AND APPLICATIONS OF SPEECH TECHNOLOGIES IN THE AI ERA

ВЛАДО ДЕЛИЋ[1], МИЛАН СЕЧУЈСКИ[2]
СИНИША СУЗИЋ[3], ТИЈАНА НОСЕК[4]
ВУК СТАНОЈЕВ[5]

Прегледни рад
DOI: 10.5937/VI25035D

Резиме: Прогрес у развоју говорних технологија у последњој деценији базиран је на дубоком учењу као парадигми вештачке интелигенције, а и њихова примена представља типичан пример онога што подразумевамо под вештачком интелигенцијом. Развој говорних технологија за неки језик мотивисан је потребом да се том језику да глас у ери вештачке интелигенције, односно, да се на том језику омогући говорна комуникација с машинама. У раду су представљена настојања да се говорне технологије развију за српски и сродне језике. Приказана су достигнућа у том развоју и прве шире примене, као и трендови у светлу промена које вештачка интелигенција доноси у индустрији и образовању.

Кључне речи: препознавање говора, синтеза говора, очување српског језика

Abstract: Progress in speech technology development over the past decade has been based on deep learning, and their application is a typical example of what we understand as artificial intelligence. Developing speech technologies for a language is motivated by the need to provide that language with a voice in the era of artificial intelligence, that is, enable spoken

[1] Владо Делић, Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад, vlado.delic@uns.ac.rs, ORCID број: 0000-0002-4558-9918
[2] Милан Сечујски, Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад, secujski@uns.ac.rs, ORCID: 0000-0002-3426-3277
[3] Синиша Сузић, Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад, sinisa.suzic@uns.ac.rs, ORCID: 0000-0002-0511-6729
[4] Тијана Носек, Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад, tijanadelic@uns.ac.rs, ORCID: 0000-0002-3707-0286
[5] Вук Станојев, Универзитет у Новом Саду, Факултет техничких наука, Трг Доситеја Обрадовића 6, Нови Сад, vukst@uns.ac.rs, ORCID: 0000-0003-2517-3728

communication with machines in that language. This paper presents efforts to build such technologies for Serbian and related languages, outlining major achievements and initial broad applications. It also discusses emerging trends shaped by the growing influence of artificial intelligence in industry and education.
Key Words: Speech recognition, speech synthesis, preservation of the Serbian language

1. Увод

Све је израженија потреба човека да разговара не само са људима, него и са уређајима у свом окружењу, као што су роботи, апарати у домаћинству и говорни агенти у позивним центрима. Лични говорни асистенти постају системи за диктирање, хватање белешки, преслушавање текстова, па и уређаји за превођење говора, као и помагала за особе са различитим врстама инвалидитета. Посебно је важно да се то свакоме омогући на матерњем језику, што захтева развој говорних технологија за аутоматско препознавање говора и синтезу говора на основу текста за сваки језик засебно.

Овај рад представља достигнућа у развоју говорних технологија за српски и сродне јужнословенске језике у светлу промена које долазе с вештачком интелигенцијом. Рад је организован на следећи начин.

У другом поглављу са нешто више детаља представљене су поменуте говорне технологије, изазови и достигнућа у њиховом развоју за српски језик.

Треће поглавље приказује прве примене у Србији и државама у региону.

У четвртом поглављу сагледани су трендови развоја и примене говорних технологија у светлу промена које вештачка интелигенција доноси на тржишту рада и у систему образовања.

2. Ретроспектива и перспективе развоја говорних технологија за српски

2.1. Актери у контексту вештачке интелигенције

Иницијатори развоја говорних технологија за српски језик на Факултету техничких наука (ФТН) Универзитета у Новом Саду почели су са тим истраживањима пре 30 година (генерација X) када су вештачка интелигенција и говорна комуникација човек-машина били у домену научне фантастике. Носиоци тог развоја данас (генерације Y и Z) користе машинско учење и вештачку интелигенцију као алате, али нису одрасли уз њих као уз интернет и мобилну телефонију са којима су се потпуно сродили. Насупрот њима, данашња деца из α- и β-генерације одрастају окружена вештачком интелигенцијом и за њих је могућност говорне комуникације човек-машина сасвим нормална. Можемо само да маштамо о томе шта њих очекује у контексту убрзане појаве нових технологија.

Вештачка интелигенција, за разлику од претходних технолошких револуција, не само да је много брже наступила него и јасно показује могућности да мења људе – и то не само на тешким, физичким и понављајућим пословима, него и на бројним умним пословима. Уз све то, она долази у време инертних образовних система са споро променљивим курикулумима. Вештачка интелигенција уводи нас у уске калупе са садржајима које бирамо кликовима, нуди нам да чујемо и видимо најпре оно са чиме се већ слажемо, а тако несвесно губимо контакт с реалношћу.

Истраживачи на ФТН су још пре 20 година почели с увођењем предмета из области говорних технологија и машинског учења у поједине студијске програме. Група за акустику и говорне технологије као део јединог акредитованог центра изузетних вредности на ФТН дуже од две деценије има низ националних и европских пројеката. Из те групе је 2003. године поникло и предузеће АлфаНум које се годинама једино у континуитету бавило говорним технологијама за српски и сродне јужнословенске језике на просторима бивше Југославије. Први кораци у развоју говорних технологија за српски ишли су у корак с развојем говорних технологија за велике светске језике, али је разлика у величини тржишта фаворизовала предузећа која су говорне технологије развијала за енглески и друге велике светске језике.

Захваљујући истрајности поменутих истраживача можемо се похвалити вредним резултатима у развоју говорних технологија за српски језик. Нажалост, вишегодишње настојање да се у развој језичких технологија за српски организовано укључи шири круг истраживача и организација уз подршку државе тек у последње време добија изгледе за успех. Да је ова иницијатива препозната од наше државе по узору на бројне државе сличне и мање величине (нпр. Словенија), данас смо могли имати језичке технологије и велике језичке моделе на знатно вишем нивоу, што би директно допринело конкурентности наших компанија и ефикасности институција, као и добробити становника на српском говорном подручју.

Велике светске компаније лансирају све моћније језичке моделе, и то по правилу један општи модел уз наменске моделе за поједине области као што су медицина, правосуђе и др. Користећи огромне количине података и енормне рачунарске ресурсе, они данас развијају и вишејезичне моделе, што омогућава да се модел за мањи језик добије прилагођењем вишејезичног модела коришћењем знатно мање ресурса. Међутим, велики језички модели обучени су на одређеним подацима, а ти подаци су различити за различите државе и језике, те народе са различитом културом, историјом и политиком. Зато је за сваку државу важно да има своје језичке моделе који су наменски развијени за дати језик, макар и прилагођењем вишејезичних модела властитим подацима.

Програм развоја језичких технологија у Србији добио је своје место тек у другој верзији Стратегије развоја у области вештачке интелигенције у Србији 2024-2030, али још није креиран пратећи акциони план. Велике светске компаније из САД и Кине показују интересовање да развијају велике језичке моделе за српски језик, што јесте у складу с наведеним Програмом, али су им за то потребни ресурси попут дигиталног репозиторијума Народне библиотеке Србије, као и говорни и језички ресурси у архиви Јавног медијског сервиса Радио телевизије Србије. Поред самих ресурса потребно је и ангажовање научне и технолошке заједнице за њихову припрему за обуку језичких и говорних модела. Програм зато предвиђа истраживачко-развојне активности на решавању проблема везаних за специфичности српског језика као што су коришћење два писма (ћирилица и латиница), екавски и ијекавски изговор, као и посебне изазове као што је избегавање употребе дијакритичких знакова латиничног писма у одређеним неформалним комуникационим ситуацијама.

Преводиоци текста и преводиоци говора појављују се и као апликације за мобилне телефоне, али и као наменски уређаји који имају микрофонске низове и софтвер за детекцију уста говорника и поништавања буке из других праваца, а имају и гласне спикерфоне, тако да омогућавају коришћење и у прилично бучној средини. Уз то, имају камере и софтвер за оптичко препознавање карактера, могућност приступа интернету кроз мрежу мобилне телефоније итд. Појављују се и компактни таблет уређаји с микрофонским низовима који у реалном времену исписују оно што се говори, раздвајају записе по говорницима, а добијени транскрипт у моменту преводе на други језик и/или формирају сажетак сниманог разговора. Међутим, све ове функционалности су на високом нивоу квалитета доступне само на одређеном броју већих језика и тек предстоје наши напори да све то имамо и на српском и сродним јужнословенским језицима. У наставку ће бити више речи о дометима у развоју говорних технологија за српски језик.

2.2. Домети у развоју синтезе говора на основу текста за српски језик

Упркос убрзаном напретку говорних технологија на светском нивоу у последњој деценији, најквалитетнији постојећи синтетизатор говора за српски језик, као и неке друге сродне језике, и даље је онај развијен у Србији, у сарадњи ФТН и предузећа АлфаНум. Овај синтетизатор свој квалитет дугује систематичном развоју језичких ресурса за ове језике (акценатско-морфолошких речника, те алата за морфолошку анотацију текста и анализу реченичне структуре). Ови ресурси обезбеђују готово непогрешив изговор и интонацију огромног броја речи у свим њиховим појавним облацима (нпр. падежни облици именица). Штавише, иако најсавременији тзв. end-to-end синтетизатори који се обучавају на огромним количинама снимака говора без улажења у

граматичку структуру текста постижу изузетну природност интонације, с времена на време још увек греше у акцентовању појединих речи – што квари утисак при оцењивању и поређењу квалитета синтетизованог говора [1].

Први значајни резултати у развоју синтезе за српски језик овај истраживачки тим постигао је 2002. године увођењем елемената природне интонације [2]. У то време су доминирали тзв. конкатенативни синтетизатори, који су спајали унапред снимљене говорне сегменте уз настојање да ублаже нагле прелазе на спојевима. Уз брижљиво припремљене базе говорних сегмената и усавршаване алгоритме њиховог избора спрам датог текста постизани су значајни резултати у погледу разумљивости, па чак и природности. Међутим, промена гласа или стила говора била је немогућа без снимања нове говорне базе, те су временом преовладали параметарски синтетизатори говора способни да „уче“ из података уместо да их репродукују, и то прво на бази скривених Марковљевих модела (HMM) [3], а потом и помоћу дубоких неуронских мрежа (DNN) [4]. Први синтетизатор за српски језик на бази HMM развијен је од стране истраживача са ФТН [5]. Иако флексибилнији за промене гласа и стила говора, имао је одређена ограничења квалитета синтезе због пробабилистичког приступа, што је био проблем са којим се суочила читава научноистраживачка заједница и који је превазиђен тек појавом синтезе на бази DNN.

Развој синтезе на бази DNN одвијао се у две фазе. У првој фази се још увек користе системи са класичном језичком обрадом текста, док у последње време доминирају end-to-end синтетизатори говора. Наше публикације [6], [7], [8], [9], [1] прате прогрес у развоју синтетизатора говора на основу текста на српском језику у последњој декади. Даљи развој заснивaће се на комбиновању предности које доносе развијени језички ресурси са предностима које доносе решења заснована на дубоком учењу.

Све шира примена синтезе је у двосмерној говорној комуникацији човекмашина у аутоматизованим позивним центрима (енг. voice bots). За утисак двосмерног разговора у реалном времену потребно је да кашњење од добијања текста за синтезу до почетка репродукције првих синтетизованих речи буде испод једне секунде, што постојећи синтетизатор за српски језик већ постиже. За још ширу примену синтезе потребна је варијабилност стила и динамике, као и могућност промене гласа говорника. Пут ка томе отварају велики језички модели, који би брзом анализом текста могли да дају инструкције синтетизатору каквим би гласом и на који начин требало изговорити поједине реченице.

2.3. Домети у развоју препознавања говора за српски језик

Аутоматско препознавање говора има ширу и разноврснију примену од синтезе говора на основу текста, те су решења до којих су дошли светски

гиганти на високом нивоу, укључујући и српски језик. Међутим, ова решења најчешће подразумевају препознавање на серверима којима се приступа преко интернета, што је повезано с низом безбедносних и етичких питања – од неовлашћеног пресретања аудио записа говора на интернету до начина чувања и коришћења снимака јер је глас говорника биометријско обележје. Зато многе институције, као што су медицинске и правосудне, инсистирају на решењу „у кући“ (енг. on premise), а АлфаНум је једина компанија која развија и пласира оваква решења за српски језик, поштујући специфичне потребе ових институција, нпр. специфични вокабулар медицинских налаза.

Дуги низ година основна парадигма за развој система за препознавање говора били су пробабилистички модели – скривени Марковљеви модели (HMM) [10], са развојем DNN поједини делови HMM система били су постепено препуштани неуронским мрежама [11], да би у последњој деценији end-to-end DNN системи са повећањем ресурса за обуку (подаци и процесорски капацитети) надмашили све претходне системе по тачности препознавања говора, па чак и човека [12]. Предности оваквих система долазе до изражаја у отежаним условима препознавања (веома бучна средина, више симултаних говорника), где комбиновањем више модалитета (аудио и видео) може да се формира прецизан транскрипт. Ово је управо један од праваца истраживања на пројекту Multimodal multilingual human-machine speech communication – AI-SPEAK код Фонда за науку Републике Србије, чији је носилац управо истраживачки тим са ФТН, и очекује се да резултати пројекта додатно унапреде обе говорне технологије за српски језик. Поред препознавања говора, у говору је могуће аутоматски препознати и говорника, па и његово тренутно расположење, емоције, али и ставове и намере у дијалошким системима. То су типични примери онога што подразумевамо под вештачком интелигенцијом, тако да и развој и примена говорних технологија представљају угледне примере вештачке интелигенције. Најзначајнији резултати у препознавању говора на српском језику постигнути у сарадњи истраживачког тима са ФТН и предузећа АлфаНум публиковани су у радовима [13], [14] и [15].

3. Примене говорних технологија на српском језику

Неке од примена говорних технологија базирају се само на једној говорној технологији, док њихова комбинација омогућује двосмерну говорну комуникацију човек-машина. Као потенцијално најшире примене могу се издвојити персонални (говорни) асистенти и говорни агенти (енг. *voice bots*), који могу да пруже услуге позивних центара великом броју корисника истовремено јер се значајан део услуга-одговора може у потпуности реализовати аутоматски применом говорних и језичких технологија.

3.1. Примене синтезе говора на основу текста

Прве примене синтезе говора на основу текста на српском језику биле су намењене слепим корисницима рачунара. Они су први прихватили ову говорну технологију још у време конкатенативних сиснтетизатора скромног квалитета, јер су по први пут добили елементе природне интонације у синтетизованом говору на српском језику.

До сада највећи број примена је на веб сајтовима разних институција и медијских кућа – који омогућавају преслушавање текстова са тих веб сајтова помоћу опције „Читај ми“. Примера ради, свака вест приликом постављања на сајт се аутоматски конвертује и у говор, те се поред наслова те вести појављује дугме за покретање, паузирање и прекидање преслушавања.

Примери ове примене синтезе говора су на веб сајтовима влада (Србије и Црне Горе), већег броја јавних институција (скупштине, градови), јавних медијских сервиса и приватних новинских агенција у Србији, Црној Гори, као и Босни и Херцеговини, па и у Хрватској. Поред тога, почеле су и примене у разгласу у јавном превозу, нпр. на станицама брзе пруге у Србији, неких аеродрома у региону, као и такси удружења. Поред тога, произведен је већи број аудиоиздања књига и часописа, прво за потребе савеза слепих, касније и за шири круг корисника, укључујући студенте – пример је Централна аудиобиблиотека Универзитета у Новом Саду [16].

У неким ситуацијама корисно је да синтетизовани говор буде праћен приказом анимираног лица (аватара) који покретима усана и изразима лица прати оно што изговара. То омогућава боље разумевање, па чак и у бучној средини. Анимирање лица аватара није толико језички зависно као синтеза говора, већ је облик и израз лица више је повезан са акустичким својствима онога што се тренутно изговара, независно од језика. Лик аватара може одговарати говорнику (нпр. особи која држи предавање), али и некој историјској личности коју је на овај начин могуће „оживети“.

Јасно је, међутим, да клонирање нечијег гласа, а поготово синхронизованих покрета усана и израза лица, може бити и предмет злоупотребе (енг. deepfake). Клониране гласове, па и видео-снимке говора, све је теже разликовати од природних, а при томе је вештачка интелигенција успешнија у откривању вештачки генерисаног садржаја од човека.

Једино решење за спречавање злоупотреба ових технологија је у доношењу одговарајућих прописа који би предвиђали обавезу да сваки дигитални садржај (аудио и видео) садржи „потпис“ аутора у виду тзв. воденог жига (енг. watermark) који не нарушава аудио/видео садржај, али омогућује да се лако утврди да ли иза садржаја заиста стоји особа наведена као аутор.

3.2. Примене препознавања говора

С развојем аутоматског препознавања говора ширио се и опсег примене. До пре десетак година доминирале су примене ограниченог обима речника и то углавном на великим светским језицима. Већ то је омогућавало издавање говорних команди у концепту паметне куће или навигацију кроз једноставнији дијалошки систем. Померање парадигме развоја говорних технологија с HMM на DNN концепт знатно је повећало тачност и робусност препознавања говора, а велики језички модели повећали су применљивост дијалошких система.

АлфаНум је у сарадњи са ФТН развијао сопствена решења за аутоматско препознавање говора. Овај развој је последњих десетак година фокусиран на области као што су медицина и правосуђе. Да би се омогућило диктирање медицинских налаза било је неопходно омогућити правилну интерпретацију разних медицинских термина, скраћеница и мерних јединица, а језички модел развијен за српски језик морао је бити проширен латинским изразима. С друге стране, вокабулар и говорни стил потпуно су другачији у правосуђу, где, примера ради, судски списи садрже мноштво властитих имена, назива предузећа, улица, те географских појмова. Уз интензивну сарадњу с партнерима из области медицине и правосуђа, како у погледу обуке језичких модела, тако и функционалности које су потребне корисницима у овим доменима, настали су производи „Медикта“ и „Јурисдикта“. Ови системи записују диктат у реалном времену са изузетно високом тачношћу, при чему се препознавање обавља на серверу који се физички налази у оквиру саме институције, а уз то могу да се интегришу у софтвер који се у институцији већ користи.

Прилагођавањем отворених модела као што је препознавач говора Whisper компаније *OpenAI* [17] или Parakeet модел компаније NVIDIA [18], исти истраживачки тим развио је и препознавач говора опште намене под називом „Транскрипта“, који може да се користи за транскипцију говора снимљеног са нпр. састанка или предавања. Добијени транскрипт синхронизован је с аудио снимком, што омогућава брзу корекцију уочених штампарских грешака, а постоји и могућност дијаризације, тј. раздвајања транскрипта по говорницима.

Илустрације ради, препознавач говора Whisper био је обучен са 680.000 сати снимљеног говора и пратећег транскрипта (у верзији 3 и преко 1.000.000 сати) – већином на енглеском, доста на великим светским језицима, а веома мало на српском (око 20-так сати). Колико су то огромни ресурси за обуку најбоље говори чињеница да је АлфаНум успео да припреми за обуку мање од 2.000 сати говора с прецизним транскриптом – што је много више у односу на остале постојеће базе за српски језик. Пре прилагођавања за српски језик Whisper (medium) имао је стопу грешке од 30% на нивоу речи, али се прилагођавањем на основу расположивих ресурса АлфаНума грешка може спустити

на 5%. Међутим, Whisper није предвиђен за препознавање у реалном времену него за транскрипцију већ снимљеног говора, што онемогућава примене у позивним центрима и voice bot апликацијама, које захтевају мало кашњење.

Напредни персонални говорни асистенти би, поред синтезе говора на основу текста и аутоматског препознавања говора (шта је речено), требало да препознају и говорника (ко је рекао), па и његово расположење, намере и ставове. Дакле, такви говорни асистенти треба да обухвате све поменуте говорне технологије и вештачку интелигенцију у пуној мери.

3.3. Примене говорних као асистивних технологија

Говорне технологије могу да помогну великом броју особа са различитим врстама инвалидитета. Слепе особе могу без помагала да читају само текст на Брајевом (рељефном) писму помоћу чула додира, те не могу да користе рачунар или сличан уређај без синтетизетора говора. Слепи у Србији користе АлфаНум синтезу (верзија 4), под називом „anReader“, углавном за слушање обимнијег текста, док за навигацију кроз менију на рачунару чешће користе скромнији синтетизатор због бржег одзива. Очекују се пројекти који ће омогућити да се квалитетнија синтеза (верзија 5 или 6) пружи слепим корисницима рачунара. Важна улога синтетизатора говора је и у продукцији аудиокњига, што је много практичније од ангажмана професионалних спикера, али захтева синтезу изузетно високог квалитета, ону која би у идеалном случају свој говор прилагодила контексту у мери у којој је то у стању и професионални спикер.

Одређено унапређење квалитета експресивне синтезе говора може се постићи уз помоћ инструкција како би нешто требало изговорити које може пружити одговарајући велики језички модел, па је могуће и читање гласовима више говорника у случају потребе. Неопходни су и TTS модели тренирани на већим количинама експресивних података. Треба напоменути да су аудиоиздања корисна не само слепима већ и свима онима који не могу лако да листају књиге, као што су особе са церебралном парализом, дистрофијом и квадриплегијом, а слично се односи и на особе са дислексијом.

Синтетизатори говора веома су корисни особама с потешкоћама у говору, без обзира да ли се ради о немим особама, особама које су подвргнуте ларингектомији или су привремено изгубиле моћ говора. Сви они могу на телефону да напишу шта желе да кажу и телефон ће помоћу синтетизатора претворити дати текст у гласан и јасан говор. У случају особа подвргнутих ларингектомији чак је могуће реализовати и синтезу говора гласом исте особе снимљеним пре ларингектомије.

С друге стране, аутоматско препознавање говорних команди може да послужи особама с физичким инвалидитетом да говорним командама управ-

љају уређајима у окружењу (нпр. паметној кући), као и особама са било којим другим типом инвалидитета, под условом да могу разговетно да издају те команде. Препознавање говора од користи је и особама које не чују говор јер им ставља на располагање транскрибовани текст. Најзад, како би се у потпуности схватиле могућности говорних технологија за помоћ особама с инвалидитетом, треба приметити да оне омогућују комуникацију помоћу говора и/или текста између потпуно глуве, слепе и неме особе.

4. Трендови развоја и примене говорних технологија

Тренд нестанка малих језика, који је свакако присутан, могао би се интензивирати у ери вештачке интелигенције за језике за које се не буду пратили трендови развоја и примене говорних и језичких технологија, јер ће се млади нараштаји ослањати на језике на којима су им расположиве нове технологије, те корисне и забавне апликације. С друге стране, вишејезични модели и релативно лак развој говорних и језичких технологија за сваки наредни језик омогућује да се ове технологије развију и за језике са малим бројем говорника. Пример језика за који су развијене говорне технологије, које доприносе и његовом очувању, јесте језик Лужичких Срба, који данас говори свега око 60.000 говорника у Немачкој [19]. Чињеница да данас деца (α- и β-генерација) одрастају у окружењу у ком је нормално да машине причају као људи и разумеју не само шта човек говори, него и ко и како говори, омогућиће брже и лакше прихватање свих нових апликација и сервиса базираних на говорној комуникацији човек-машина. Свако ће у џепу имати приватног туристичког водича, преводиоца, дактилографа, спикера, па и некога с ким може увек попричати, а ко ће свог власника добро познавати, препознавати нијансе у његовом гласу, расположење и намере, служити као извор свих жељених информација, подсетник, забављач, заштитник од превара и саветник у пословним одлукама. Свака нова технологија може бити и злоупотребљена, па се томе мора посветити посебна пажња. Један правац је глобална сарадња на изради законских решења и сузбијању злоупотреба, али ћемо на крају морати сами да критички преиспитамо све што може да буде лажно, и то вероватно уз помоћ персоналних асистената. На крају ће најуспешнији бити они који највештије буду користили добробити које нам вештачка интелигенција доноси, и зато је важно дати глас сваком језику у ери у коју смо већ ушли, и за коју нам досадашње искуство говори да ће бити пуна динамичних промена.

5. Закључци

Говорне и језичке технологије које се ослањају на велике језичке моделе, развијене за одређени језик, повећавају продуктивност и конкурентност при-

вредних субјеката, унапређују ефикасност јавне управе и олакшавају коришћење дигиталних технологија на датом говорном подручју.

У раду су представљени домети у развоју говорних технологија за српски и сродне језике, перспективе даљег развоја и примене у ери вештачке интелигенције. Представљене су прве примене ових технологија у Србији и региону, како синтезе говора на основу текста тако и препознавања говора, са посебним осврtом на примену говорних технологија као асистивних.

Иако развој вештачке интелигенције може да продуби јаз између великих и малих држава и нација, те да се у ери дигитализације мали језици изгубе, прилагођење великих језичких модела у циљу развоја говорних технологија за мале језике може и да допринесе њиховом очувању. На тај начин сваки језик може да добије глас у ери вештачке интелигенције у коју улазимо кроз 4. и 5. индустријску револуцију, а одговарајући напори морају се уложити и у развој законских решења неопходних за сузбијање злоупотреба ових технологија.

6. Захвалница

Истраживање спроведено уз подршку Фонда за науку Републике Србије, #7449, *Multimodal multilingual human-machine speech communication* – AI-SPEAK

7. Литература

[1] T. Nosek et al, End-to-End Speech Synthesis for the Serbian Language Based on Tacotron. *Proc. SPECOM 2024*, Belgrade, Serbia, pp. 219-229, 2024.

[2] M. Sečujski et al, AlfaNum System for Speech Synthesis in Serbian Language. *Proc. of TSD'02*, Brno, Czech Republic, pp. 237-244, 2002.

[3] T. Yoshimura et al. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *Proc. of EUROSPEECH*, Vol. 5, pp. 2347–2350, 1999.

[4] H. Zen, A. Senior, M. Schuster, Statistical parametric speech synthesis using deep neural networks. *Proc. ICASSP 2013*, pp. 7962–7966, 2013.

[5] Е. Пакоци, Синтеза говора на бази скривених Марковљевих модела за српски језик, *Дипломски мастер рад*, Факултет техничких наука, 2012.

[6] T. Delić, M. Sečujski, S. Suzić, A review of Serbian parametric speech synthesis based on deep neural networks. *Telfor Journal*, Vol. 1, No. 9. pp. 32–37, 2017.

[7] M. Sečujski et al, Speaker/style-dependent neural network speech synthesis based on speaker/style embedding. *JUCS*, Vol. 4, No. 26, pp. 434–453, 2020.

[8] S. Suzić et al, HiFi-GAN based text-to-speech synthesis in Serbian. *Proc. 30th EUSIPCO*, Belgrade, Serbia, pp. 2231–2235, 2022.

[9] S. Suzić et al, Style transplantation in neural network based speech synthesis. *Acta Polytech. Hung.*, Vol. 6, No. 16, pp. 171–189, 2019.

[10] L. R. Bahl et al, Speech recognition with continuous-parameter hidden Markov models. *Computer Speech & Language*, Vol. 2, Issues 3–4, pp. 219-234, 1987.

[11] G. Hinton et al, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Processing Magazine*, Vol. 29, No. 6, pp. 82-97, 2012.

[12] H. Ahlawat, N. Aggarwal, D. Gupta, Automatic speech recognition: a survey of deep learning techniques and approaches. *IJCCE*, Vol. 6, pp. 201-237, 2025.

[13] B. Popović et al, Voice assistant application for the Serbian language. *Proc. 23rd TELFOR*, Belgrade, Serbia, pp. 858–861, 2015.

[14] B. Popović, E. Pakoci, D. Pekar, Transfer learning for domain and environment adaptation in Serbian ASR. *Telfor Journal*, Vol. 2, No. 12, pp. 110–115, 2020.

[15] E. Pakoci et al, Overcoming data sparsity in automatic transcription of dictated medical findings. *Proc. 30th EUSIPCO*, pp. 454–458, 2022.

[16] V. Delić et al, Central audio-library of the University of Novi Sad. *Proc. of Intelligent Distributed Computing XIII*, pp. 467–476, 2020.

[17] A. Radford et al, Robust speech recognition via large-scale weak supervision. *Proc. of the International Conference on Machine Learning*, pp. 28492-28518, 2023.

[18] O. Kuchaiev et al, Nemo: a toolkit for building ai applications using neural modules, *arXiv*:1909.09577, 2019.

[19] I. Kraljevski et al, Preserving Language Heritage Through Speech Technology: The Case of Upper Sorbian. *Proc. SPECOM 2024*, Belgrade, Serbia, pp. 3-22, 2024.