UDC 004.943.5 Info M str. 63-66

СИСТЕМИ ЗА СИНТЕЗУ И ПРЕПОЗНАВАЊЕ ГОВОРА (TTS, STT): ПРИМЕНА У ПРЕТРАЖИВАЧУ ЗВУКА УНИВЕРЗИТЕТСКЕ БИБИЛОТЕКЕ "СВЕТОЗАР МАРКОВИЋ"
SPEECH SYNTHESIS AND RECOGNITION SYSTEMS (TTS, STT): APPLICATION IN THE SOUND SEARCH ENGINE OF THE UNIVERSITY LIBRARY "SVETOZAR MARKOVIĆ"

Ивана Гавриловић, Никола Смоленски, Адам Софронијевић

РЕЗИМЕ: Овај рад се бави темама: технологија синтезе говора (енг. *text-to-speech — TTS*) и технологија препознавања говора (енг. *Speech-to-Text — STT*), са осврtом на њихову примену у претраживачу звука Универзитетске библиотеке "Светозар Марковић" (УБСМ). У уводном делу се тема разрађује теоретски и из историјске перспективе, почев од механичких уређаја из 18. века, попут Краценштајновог модела вокалног тракта из 1779. године, до савремених система заснованих на вештачкој интелигенцији и дубоком учењу. Поменут је и развој ових технологија за српски језик, од првих система из средине 1980-их до данашњих решења. Рад описује имплементацију претраживача звучних записа са Јутјуб канала УБСМ, који користи *Whisper JAX* технологију за препознавање говора, достижући тачност препознавања од преко 90%. Детаљно су описани процеси прикупљања метаподатака, транскрипције говора, обраде добијених података и њиховог складиштења у бази података заснованој на *Lucene/Solr* систему. Систем омогућава претрагу транскрибованог садржаја уз подршку за проналажење сличних речи коришћењем Левенштајнове удаљености, што повећава ефикасност претраге упркос могућим грешкама у препознавању говора. Разматрани су и изазови попут недостатка временских предметних одредница у метаподацима и зависности квалитета препознавања од квалитета изворног снимка. Закључак је да је развијени систем користан алат који олакшава приступ аудио-визуелном садржају библиотеке, уз планове за даљи развој и проширење његове функционалности.

КЉУЧНЕ РЕЧИ: обрада природног језика, синтеза говора, препознавање говора, претрага звучних записа, претрага текста

ABSTRACT: This paper addresses the topics of speech synthesis technology (text-to-speech — TTS) and speech recognition technology (speech-to-text — STT), with a focus on their application in the sound search engine of the University Library "Svetozar Marković". The introductory section elaborates on the subject both theoretically and from a historical perspective, starting with mechanical devices from the 18th century, such as Kratzenstein's vocal tract model from 1779, up to modern systems based on artificial intelligence and deep learning. The development of these technologies for the Serbian language is also mentioned, from the first systems in the mid-1980s to today's solutions. The paper describes the implementation of a search engine for audio recordings from the University Library's YouTube channel, which uses *Whisper JAX* technology for speech recognition, achieving recognition accuracy of over 90%. The processes of metadata collection, speech transcription, data processing, and storage in a *Lucene/Solr*-based database are described in detail. The system enables searching transcribed content with support for finding similar words using Levenshtein distance, which increases search efficiency despite possible errors in speech recognition. Challenges such as the lack of temporal subject descriptors in metadata and the dependence of recognition quality on the quality of the source recording are also discussed. The conclusion is that the developed system is a useful tool that facilitates access to the library's audiovisual content, with plans for further development and expansion of its functionality.

KEY WORDS: natural language processing, speech synthesis, speech recognition, audio search, text search

1.1. УВОД

Технологије препознавања говора (енг. *speech recognition*) и синтезе говора (енг. *speech synthesis*) су савремени алати дигиталне комуникације који суштински мењају начин на који људи комуницирају са рачунарима. Ове напредне технологије су постале саставни део савремених дигиталних система. Захваљујући њима, контакт између човека и машине постаје све приступачнији, интуитивнији и природнији (Jurafsky & Martin, 2025).

Како је дигитални свет све више део нашег свакодневног живота, технолошки напредак у области природне обраде језика игра важну улогу у олакшавању приступа информацијама и комуникацији са различитим дигиталним уређајима. Технике које омогућавају рачунарима да разумеју и генеришу људски језик заједнички се зову обрада природног језика.

Обрада природног језика (ОПЈ) је грана вештачке интелигенције која се бави интеракцијом између рачунара и људског језика а за циљ има омогућавање рачунарима да „разумеју", интерпретирају, анализирају и произведу људски језик на смислен и користан начин. Обрада природног језика је настала средином 20. века на раскрсници између лингвистике и вештачке интелигенције (Nadkarni, 2011). Она користи мултидисциплинаран приступ разумевању и манипулацији људским језиком.

У почетним декадама развоја обраде природног језика, истраживачи су се суочавали са бројним изазовима превођења и разумевања, од којих су најупечатљивији били хомографи и метафоре, који су онемогућавали једноставне преводе реч по реч (Nadkarni, 2011). Важан теоријски допринос дао је Ноам Чомски 1956. године анализом граматика језика, што је касније утицало на развој формалних нотација попут Бакус-Наурове форме за спецификацију контекстно-слободних граматика (Nadkarni, 2011). Током осамдесетих година долази до великог преокрета у приступу, те се уместо дубинских анализа уводе једноставније ме-

тоде машинског учења, уз коришћење анотираних корпуса текста за тренирање алгоритама (Nadkarni, 2011). Савремена ера у развоју природне обраде језика се карактерише развојем сложенијих система који омогућавају напредније разумевање људског језика у различитим доменима.

ОПЈ обухвата велики број техника и алгоритама који омогућавају рачунарима да обављају задатке као што су препознавање говора, стварање говора, анализа реченица, препознавање ентитета, машинско превођење, класификација текста, синтактичка и семантичка анализа итд. *ОПЈ* омогућава рачунарима да обраде људски језик на напредни начин што је кључно за развој раличитих система као што су виртуелни асистенти, четботови, машинско превођење, аутоматско генерисање сажетака текстова и многих других.

1.2. ТЕРМИНОЛОГИЈА

У области обраде природног језика постоји више сродних термина који су семиотички блиски, али семантички различити. Из тог разлога је неопходно прецизније дефинисати поједине термине који се користе у овом раду.

Обрада природног језика (енг. *natural language processing — NLP*) је грана вештачке интелигенције која се бави анализом, разумевањем и генерисањем људског језика. Важно је напоменути да се иста скраћеница *NLP* користи и за неуролингвистичко програмирање, (енг. *neurolinuistic programming*) што је врста психотерапије и у том значењу се не користи у контексту машинске обраде природног језика. Такође се може срести појам природна обрада језика, што је лош превод енглеског *natural language processing.*

Препознавање језика (енг. *language recognition*) представља аутоматску идентификацију језика писаног или говорног текста. Ова технологија је често први корак у процесу синтезе или препознавања говора, јер омогућава избор одговарајућег језичког модела за даљу обраду.

Препознавање говора (енг. *speech recognition*) је технологија која преводи говорни сигнал у текст. У оквиру ове области разликујемо:

Препознавање изолованих речи: способност система да препозна појединачне, унапред дефинисане речи.

Препознавање континуалног говора (енг. *speech-to-text — STT*) је напреднија технологија која може да препозна природан, неограничен говор.

Синтеза говора (енг. *text-to-speech — TTS*) је технологија која омогућава претварање писаног текста у вештачки генерисан говор.

Препознавање гласа (енг. *voice recognition*) је специфична технологија која се бави идентификацијом појединачних говорника на основу њихових гласовних карактеристика. Ова технологија се понекад користи у безбедносне сврхе и за разликовање говорника у звучним записима који имају више говорника.

У даљем тексту се детаљније приказују процеси синтезе и препознавања говора и начина на који се препознавање говора користи за претрагу корпуса говорног материјала.

2. СИНТЕЗА ГОВОРА

Синтеза говора или *TTS* је технологија која омогућава рачунарима претварање писаног садржаја у говор. Анализом текста и генерисањем звука који одговара садржају текста, *TTS* има широку примену у ситуацијама када корисници желе да чују информације које су написане, али немају могућност или би било компликовано да их прочитају. На пример, *TTS* се користи у навигационим системима, апликацијама за помоћ особама са оштећењем вида и у различитим асистентима као што су *Siri* и *Google Assistant*.

Први покушаји да се опонаша људски говор укључивали су механичке уређаје попут модела вокалног тракта које је 1779. израдио Кристијан Краценштајн и акустично-механичке машине за говор Волфганга фон Кемпелена из 1791. године. Тридесетих година 20. века, Бел лабораторије развиле су вокодер, а 1939. Хомер Дадли је представио Водер, синтетизатор гласа којим се управљало путем тастатуре. Рачунарска синтеза говора започела је касних 1950-их, са значајним достигнућима попут синтезе песме «Daisy Bell» на ИБМ 704 1961. године и развоја линеарног предиктивног кодирања (енг. *linear predictive coding — LPC*) шездесетих година. Током осамдесетих и деведесетих, појавили су се системи попут ДЕЦталк-а и вишејезични синтетизатори говора из Бел лабораторија, заједно са едукативним и забавним уређајима са функцијом синтезе говора. До 2010-их, напредак је значајно унапредио квалитет синтетизованог говора, иако је он и даље остајао препознатљив у односу на људски глас (Сасирехка и Чандра, 2024). Још једна занимљива примена синтезе говора су вокалоиди. Вокалоид је синтетизатор говора који има могућност промене тона, брзине, динамике, и других аспеката говора, што му даје напредне могућности као што је могућност да пева (Bell 2015).

За израду синтетизатора говора могућа су два приступа: први је потпуна синтеза вештачког гласа, а други је снимање природног говорника, сечење његовог говора на погодне делове (појединачне фоне или слогове), и њихово поновно спајање у жељени текст.

Најстарији пронађени синтетизатори говора за српски језик потичу из средине 1980-их година (Jovanović, 1985) (Ogrizek, 1986). Данас најраширенији синтетизатор говора за српски језик развила је компанија АлфаНум на основу истраживања рађених на Факултету техничких наука у Новом Саду (Сечујски 2003).

3. ПРЕПОЗНАВАЊЕ ГОВОРА

Препознавања говора је технологија која преводи говор у текстуални облик, а која се развија више од педесет година. Овде можемо направити разлику између разазнавања присуства појединачних речи у одређеном говору, и много сложенијих технологија препознавања континуалног говора без ограничења у речнику које називамо превођењем говора у текст (*STT)*.

Препознавање говора је много сложеније од синтетизације говора. Ова комплексна технологија заснива се на напредним математичким моделима попут Бајесових и

Марковљевих, који омогућавају анализу акустичких образаца и њихову трансформацију у дигитални текст. Технички изазови су значајни: обухватају разлике у изговору између говорника, акустичке двосмислености, варијације у брзини и начину изговора, те утицај контекста на разумевање речи. Препознавање говора захтева прецизне алгоритме који истовремено обрађују лингвистичка правила, моделе људског понашања и специфичности акустичног улаза. Примена ове технологије је широка, почев од помоћи особама са моторичким инвалидитетом, контроле уређаја, до писања текста диктирањем и чак потенцијалне рехабилитације говорних поремећаја. Упркос сталном напретку, технологија и даље захтева значајна унапређења у тачности и поузданости препознавања говора.

Развој препознавања говора усмерен је ка повећању броја препознатих речи, разумевању без прилагођавања говорнику, и повећању брзине препознавања. Први системи за препознавање говора направљени су 1950-их година, и могли су да препознају само појединачне речи, само једног одређеног говорника. Крајем 1960-их, системи су добили могућност препознавања континуалног говора, на ограниченом речнику. У 1980-им развој се усмерио на статистичку анализу говора, што је довело до првих практично употребљивих *STT* система. У 2000-им је започета употреба вештачких неуронских мрежа и технологија дубоког учења, те су *STT* системи поузданости упоредиве са људском достигнути тек крајем 2010-их.

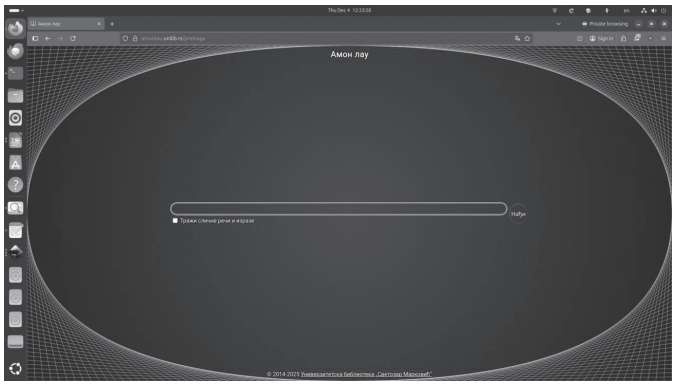
Најстарији пронађен препознавач говора за српски језик је такође из средине 1980-их (Antunović и Ćurić, 1986).

4. ПРЕТРАЖИВАЧ ЗВУКА УНИВЕРЗИТЕТСКЕ БИБЛИОТЕКЕ "СВЕТОЗАР МАРКОВИЋ"

Универзитетска библиотека "Светозар Марковић" од 6. маја 2009. године одржава канал библиотеке на платформи Јутјуб, доступан на Интернет адреси **https://www.youtube.com/@UBSMBeograd**. Канал редовно преноси предавања и догађаје одржане у УБСМ, и друге релевантне садржаје. Током петнаест година постојања канала, архива је нарасла и када се све сабере има аудио-визуелни материјал у укупном трајању од готово десет дана те се појавила се потреба за његовим ефикасним претраживањем.

2022. године *OpenAI* лабораторија развила је *Whisper* модел за препознавање говора (Radford et al. 2022) (OpenAI). Нешто каније, развијен је и *Whisper JAX*, оптимизована имплементација *Whisper* модела (Gandhi 2024). На тестирању у оквиру Универзитетске библиотеке, *Whisper JAX* је изабран због велике брзине, једноставности употребе и квалитета препознатог говора. Користи се у стандардној конфигурацији која истовремено препознаје све језике, због тога што аудио-визуелни материјал библиотеке и садржи разне језике.

Пројекат израде претраживача звука је назван Амон лау, по топониму из Толкинове митопеје који значи "Брдо слуха" а односи се на једно брдо са ког се може чути надалеко (вероватније познатије је њему суседно брдо Амон хен, што значи "Брдо вида", а са ког се може видети надалеко).



Слика 1: *Изглед корисничког интерфејса пројекта Амон Лау*

Први корак реализације претраживача звука Универзитетске библиотеке је био увоз метаподатака о Јутјуб видеима. Метаподаци се преузимају и чувају у текстуалном дигиталном документу у *JSONL* (*JSON Lines*) формату (Ward, Ian). Пошто су подаци прикупљени, приступа се њиховој преради и уносу у базу података. Пошто су подаци прерађени, из дигиталних докумената са метаподацима видеа и транскриптом звучног записа уносе се у базу података. Имплементацији претраживача се приступа преко симулираног веб прегледача по чему се резултати транскрипције смештају у локални текстуални дигитални документ. Као систем за управљање базом података користи се *Lucene/Solr* (The Apache Foundation, 2024) са додацима за српски језик развијеним за потребе Претраживе, највеће претраживе дигиталне библиотеке региона, великог пројекта Универзитетске библиотеке за претрагу штампаних материјала, доступног на адреси **https://**претражива.срб. Претрагу базе корисници могу да врше коришћењем корисничког интерфејса на Интернет адреси **https://amonlau.unlib.rs**. Претрагом је могуће прегледати исечке из транскрибованих видео записа, при чему је такође могуће прегледати сам видео запис, од момента на ком је пронађен жељени појам, тако да је могуће чути га у контексту.

Мада је претраживач звука заокружена целина која је корисна корисницима, постоје планови за даљи развој. Пре свега, на претраживач ће се додати више релевантних материјала. За ово може бити омогућен унос података са других платформи осим Јутјуба. Такође ће се за транскрипцију користити *Whisper JAX* на сопственом хардверу.

5. ЗАКЉУЧАК

Развој претраживача звука Универзитетске библиотеке «Светозар Марковић» представља значајан корак ка модернизацији библиотечких услуга који показује једну од бројних могућности примене технологија вештачке интелигенције у библиотечком окружењу. Пројекат је покушај одговора на растућу потребу за ефикасним претраживањем аудио-визуелне грађе, која постаје све значајнији део библиотечких колекција, посебно у контексту савремених тенденција ка мултимедијалном представљању садржаја и документовању јавних догађаја.

Током имплементације система уочени су одређени изазови, попут недостатка временских предметних одредница у метаподацима и зависности квалитета препознавања говора од квалитета изворног снимка. Показало се да су ови проблеми системске природе и захтевају пажљиво промишљање будућих решења, посебно у погледу стандардизације метаподатака аудио-визуелне грађе. Упркос овим ограничењима, систем је показао задовољавајуће резултате у пракси, посебно захваљујући имплементацији напредних функција претраге које компензују потенцијалне грешке у препознавању говора. Висока тачност препознавања говора од преко 90% и могућност проналажења сличних речи чине систем поузданим алатом за истраживање говорног садржаја.

Успешна имплементација овог система отвара пут даљем развоју сличних решења у области дигиталне хуманистике и библиотекарства. Планирана унапређења система указују на могућност његове примене и у другим институцијама културе које се суочавају са сличним изазовима у дигитализацији и претраживању звучне грађе. Интеграција оваквих система у свакодневни рад библиотека и архива може значајно унапредити доступност и претраживост аудио-визуелних колекција, чиме се остварује један од кључних циљева савременог библиотекарства: омогућавање лаког приступа знању у свим његовим облицима.

Значајан допринос имплементације система огледа се у примени алгоритма Левенштајнове удаљености, који омогућава проналажење речи сличних задатом појму и тиме компензује грешке настале током препознавања говора. На корпусу ове величине примена Левенштајнове удаљености не повећава приметно време претраге.

ЛИТЕРАТУРА

[1] Antunović, P., & Ćurić, I. (1986). „Raspoznavanje govora Spektrumom“. Moj Mikro, 36–38. (Oktobar 1986). Dostupno na: https://retrospec.elite.org/users/tomcat/yu/magshow.php?auto=&page=36&all=MMH_86_10

[2] Bell, S. A. (2015). „The dB in the .db: Vocaloid Software as Posthuman Instrument.“ Popular Music and Society, 39(2), 222–240. https://doi.org/10.1080/03007766.2015.1049041

[3] Garcia Gonzalez, R., & youtube-dl developers. (2023). „youtube-dl“. youtube-dl. (1. avgust 2023). Datum pristupa: 16. decembar 2024. Dostupno na: http://ytdl-org.github.io/youtube-dl/

[4] Gandhi, S. (2024). *Whisper* JAX: The Fastest *Whisper* API. „Hugging Face.“ Datum pristupa: 16. decembar 2024. Dostupno na: https://huggingface.co/spaces/sanchit-gandhi/*whisper*-jax

[5] Jovanović, D. (1985). Tišina, “Spectrum” govori. Svet kompjutera, 33–35. (Mart 1986). Dostupno na: https://retrospec.elite.org/users/tomcat/yu/magshow.php?auto=&page=33&all=SK_85_03

[6] Jurafsky, D., & Martin, J. H. (2025). „Speech and Language Processing“ (3rd ed. Draft). (24. avgust 2025). Dostupno na: https://web.stanford.edu/~jurafsky/slp3/ed3book_aug25.pdf

[7] Ogrizek, S. (1986). „Sintetizator govora za Spektrum. Moj Mikro“, 34–35. (Oktobar 1986). Dostupno na: https://retrospec.elite.org/users/tomcat/yu/magshow.php?auto=&page=34&all=MMH_86_10

[8] OpenAI. (2024). „*Whisper*.“ Hugging Face. Datum pristupa: 16. decembar 2024. Dostupno na: https://huggingface.co/openai/*whisper*-large-v3

[9] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). „Robust Speech Recognition via Large-Scale Weak Supervision“. arXiv. https://doi.org/10.48550/arXiv.2212.04356

[10] Sečujski, M., Delić, V., Pekar, D., Jovanov, L., & Đurić, N. (2003). „Govorne tehnologije i njihovo mesto u savremenim komunikacijama“. Info M, 2(6-7), 52–55.

[11] The Apache Software Foundation. (2024). „Apache Solr™ 9.7.0.“ The Apache Software Foundation. (30. oktobar 2024). Datum pristupa: 23. decembar 2024. Dostupno na: https://solr.apache.org/

[12] Ward, I. (2024). JSON Lines. „JSON Lines“. Datum pristupa: 16. decembar 2024. Dostupno na: https://jsonlines.org/

Ивана Гавриловић, библиотекарка саветница, Универзитетска библиотека "Светозар Марковић", Одељење за дигитализацију
Контакт: gavrilovic@unilib.rs
Област интересовања: библиотекарство и дигитална хуманистика, дигитализација библиотечких фондова, примена вештачке интелигенције, дигитална трансформација у образовању

Никола Смоленски, Универзитетска библиотека "Светозар Марковић", Одељење за дигитализацију
Контакт: smolenski@unilib.rs
Област интересовања: развој дигиталних библиотека и културних дигиталних пројеката, информатички алати за српски језик и језичке технологије, отворено знање, популарна наука и културни активизам

Др Адам Софронијевић, Универзитетска библиотека "Светозар Марковић", заменик управника
Контакт: sofronijevic@unilib.rs
Област интересовања: дигитализација културне баштине и библиотечких фондова, европске истраживачке инфраструктуре, примена вештачке интелигенције, менаџмент, читање штампаних књига