Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/12BE33Suvajdzic

ЕВАЛУАЦИЈА СОФТВЕРСКИХ АЛАТА ЗА ОБРАДУ И ПРЕПОЗНАВАЊЕ ГОВОРА

EVALUATION OF SOFTWARE TOOLS FOR SPEECH PROCESSING AND RECOGNITION

Милан Сувајџић, *Факултет техничких наука, Нови Сад*

Област – Рачунарство и аутоматика

Кратак садржај – *У овом раду представљене су неке од грана вештачке интелигенције, као што су машинско учење, неуронске мреже и обрада природног језика, које су повезане с областшу препознавања говора. Дат је детаљан опис структуре и начина функционисања система за препознавање говора. Представљена је SpeechRecognition библиотека преко које је извршена примена три тренутно најпопуларнија сервиса за препознавање говора. Урађена је њихова детаљна анализа и евалуација по задатим критеријумима, на основу које су изведени закључци о томе колико је сваки од њих погодан за коришћење.*

Кључне речи: *Софтверски алатал, обрада говора, препознавање говора*

Abstract – *This paper presents some of the branches of artificial intelligence, such as machine learning, neural networks and natural language processing, which are related to the field of speech recognition. A detailed description of the structure and functioning of the speech recognition system is given. The SpeechRecognition library was presented, through which the application of 3 currently most popular speech recognition services was performed. Their detailed analysis and evaluation according to the given criteria was done, on the basis of which conclusions were made about how suitable each of them is for use.*

Keywords: *Software tool, speech processing, speech recognition*

1 УВОД

Човек, као друштвено биће, живи и ради у заједници, у оквиру које комуницира са другим људима. Језик је човеково најважније средство комуникације, а говор примарни медиј. Говорна интеракција својствена је људским бићима, те се по томе разликује од осталих бића на земљи.

Како није увек у ситуацији да директно оствари вербални контакт са другим људима, човек се у данашње време служи средствима модерне технологије (рачунаром, мобилним телефоном и многим другим апаратима). Интеракција коју човек остварује посредно, путем тастатуре, миша, управљачких конзола и других приступа, наилази на разне потешкоће, као што су: губитак информација, грешке и време које се троши на пренос и интерпретацију. Стога се неизбежно намеће потреба имплементације говорне интеракције и у случају интеракције човека - рачунар.

Рад обухвата говорне технологије које се данас најчешће користе и претварање говора у аналогни и дигитални таласни облик разумљив рачунарима. Препознавање говора или конверзија говора у текст укључује хватање и дигитализацију звучних таласа, претварање у основне језичне јединице или фонеме, изградњу речи од фонема, и контекстуалне анализе речи.

Препознавање говора је способност рачунара да препозна опште, природне изразе од стране различитих корисника.

Нагласак је на моделирању говорних јединица и граматике на основи скривеног Марковљевог модела, као и на примени вештачких неуронских мрежа у процесу конверзије говора.

Препознавање говора омогућава унос улазних података користећи се гласом. Примене су небројене, од медицине, авио индустрије до роботике итд.

Иако постоји још много простора за унапређења, данашњи системи дају задовољавајуће резултате, те потврђују да се ова технологија развија у правом смеру.

2 ОСНОВНИ ТЕОРИЈСКИ КОНЦЕПТИ

Ово поглавље даје преглед основних теоријских појмова неопходних за разумевање методологије коришћене у експерименту, као и метода за евалуацију добијених резултата.

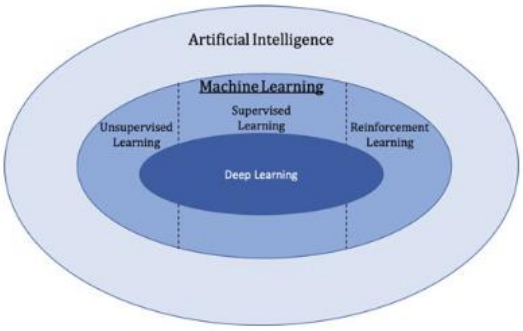
2.1 Машинско учење

Машинско учење (eng. ***Machine Learning***) је поткатегорија вештачке интелигенције (eng. ***Artificial Intelligence*** – AI), која представља способност дигиталног рачунара или рачунара контролисаног робота да извршава задатке уобичајено повезане са интелигентним бићима. Машинско учење представља проучавање рачунарских алгоритама који се аутоматски побољшавају кроз искуство. Категорије машинског учења су: **учење под надзором** (*Supervised Learning*) - ослања се на учење из скупа података са ознакама, **учење без надзора** (*Unsupervised Learning*) - одређује кластере, на основу података на без ознака и **учење ојачавањем** (*Reinforcement Learning*) - фокусира се на максимизирање награде за дату радњу или низ предузетих радњи [1].

НАПОМЕНА:
Овај рад је проистекао из мастер рада чији ментор је био др Александар Купусинац, ванр. проф.

2.2 Дубоко учење

Дубоко учење (eng. ***Deep Learning***) је поље машинског учења које се бави алгоритмима инспирисаним структуром и функцијом мозга званом вештачке неуронске мреже. Подручје дубоког учења покрива све три категорије машинског учења, а заједно са машинском учењем упада у област вештачке интелигенције. (Слика 1)



Слика 1. *Однос дубоког учења, машинског учења и вештачке интелигенције [1]*

2.2.1 Вештачке неуронске мреже

Вештачка неуронска мрежа (eng. ***Artificial Neuron Network*** - ANN) је систем за обраду информација састављен од огромниг броја међусобно повезаних процесних чворова званих неурони. Неурони се групишу у слојеве, а веза између два неурона се назива ивица. Неуронске мреже које се користе у дубоком учењу имају више од три слоја и називају се дубоке неуронске мреже (*Deep Neural Network* – DNN). Постоје два типа неуронских мрежа: неуронске мреже без повратног преноса (eng. *Feedforward Neural Network*) и неуронске мреже са повратним преносом (eng. *Feedback Neural Network* или *Recurrent Neural Network* – RNN).

2.3 Обрада природног језика

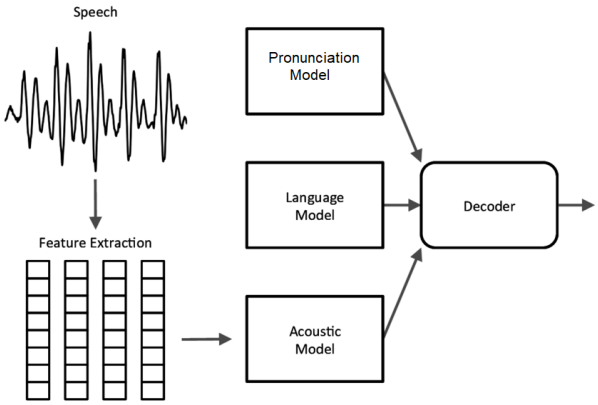
Обрада природног језика (eng. ***Natural Language Processing*** - NLP) је грана вештачке интелигенције која помаже рачунарима да разумеју, тумаче и манипулишу људским језиком. NLP црпи из многих дисциплина, укључујући рачунарску науку и рачунарску лингвистику, у својој потрази за попуњавањем јаза између људске комуникације и разумевања рачунара [1].

2.4 Препознавање говора

Препознавање говора (eng. ***Speech Recognition***) је интердисциплинарно потпоље рачунарске науке и рачунарске лингвистике које развија методологије и технологије које омогућавају препознавање и превођење говорног језика у текст помоћу рачунара. Такође је познато као аутоматско препознавање говора (eng. ***Automatic Speech Recognition*** - ASR), рачунарско препознавање говора или говор у текст (eng. *Speech-to-Text*).

2.5 Структура ASR система

Дизајн и развој система за препознавање говора зависе од различитих компоненти, као што су представљање и пред-обрада говора, класе говора, различите врсте техника издвајања карактеристика, коришћени класификатори, база података и перформансе система. Слика 2 репрезентује типичну структуру аутоматског система за препознавање говора.



Слика 2. *Структура ASR система [2]*

Улаз у систем за аутоматско препознавање говора је говорни **аудио сигнал**. Аудио сигнали су електронски прикази звучних таласа, настали као резултира преноса енергије са једног молекула на други. Постоје два типа аудио сигнала: аналогни (континуирани приказ сигнала током одређеног временског периода) и дигитални (дискретни приказ сигнала током одређеног временског периода).

Издвајање карактеристика звука је прва фаза ASR система. Главни циљ екстракције карактеристика у ASR систему је пронаћи неки скуп вектора или карактеристика које могу дати тачан приказ улазног аудио сигнала.

Акустични модел је следећа фаза, где се помоћу издвојених вектора карактеристика проналази статистички модел. Најмањи разликовни или препознатљиви део језика назива се фонема. Однос између аудио сигнала и фонема представљен је **акустичним моделом**. Креира се база података говора и на њу се примењују алгоритми обуке како би се створио статистички приказ сваке фонеме, који се називају скривени Марков модел (eng. ***Hidden Markov Model*** – HMM).

Модел изговора је једини модел у ASR систему који не пролази кроз фазу учења. Он помаже у организовању фонема да би се створила изговорена смислена реч.

Језички модел, с друге стране, помаже у организовању речи да би се створила изговорена смислена реченица. То је у основи статистички модел који показује колико је вероватно да се низ речи може појавити заједно.

Резултати свих претходно описаних модела, улазе у **декодер**. Декодер тада идентификује најтачнију транскрипцију изговорене реченице.

Из перспективе декодера, то је главни проблем претраживања. Комбинујући акустични модел, модел изговора и језички модел у декодеру се формира графикон претраживања [2].

3 ОПИС РЕШАВАНОГ ПРОБЛЕМА

Системи за препознавање говора имају своје апликативне програмске интерфејсе, путем којих се користе у оквиру других већих система. Већина апликативних програмских интерфејса за препознавање говора, доступна је на интернету и бесплатна један пробни период, а касније се наплаћује у зависности од количине потреба корисника.

3.1 Мотивација

Да би се разумео процес препознавања говора, потребно је извршити његову детаљну анализу у склопу система за препознавање говора. Тек након створене комплетне слике процеса препознавања говора, долази се до потребе евалуације система који врше тај процес. Мотивација овог рада јесте да на основу евалуације различитих система за препознавање говора, примењених преко својих апликативних програмских интерфејса, прикаже предности и мане сваког од њих по задатим критеријумима.

3.2 Преглед стања у области

Од 1920-их, систем препознавања говора се постепено развијао. Прва комерцијална играчка која се може сматрати првом машином за препознавање говора појавила се 1922. године по називом "*Radio Rex*" [1].
IBM је развио и демнострирао *Shoebox* шездесетих година прошлог века, што је био пионирски рад у данашњем систему препознавања говора. Овај јединствени уређај могао је да препозна 16 изолованих речи [1].
Седамдесетих година прошлог измишљен *HARPY* напредни систем препознавања говора заснован на вештачкој интелигенцији. Могао је препознати речник који садржи 1000 речи [1].
Статистички модели постали су популарни од периода 1980-их, а најчешће коришћени модел у вези с тим био је скривени Марков модел. Величина речника је постајала све већа и користило се око 10 000 речника [1].
Током 2000-их наставља се фокус на унапређење машинског учења. Мреже дубоког веровања примењене су на препознавање телефона, постижући врхунске перформансе на ТИМИТ корпусу. Касније је представљен хибридни модел дубоких неуронских мрежа са проширењем скривеног Марковог модела.
Активно истраживање препознавања говора довело је до доласка различитих јавних и лиценцираних софтвера алата. Најпопуларнији јавни софтверски алати су Sphinx, HTK, RWTH и Kaldi. Најпопуларнији лиценцирани софтверски алати су Siri, IBM Watson, Google Speech API, Bing Speech API, Amazon Alexa, Cortana [3].

4 ОПИС РЕШЕЊА ПРОБЛЕМА

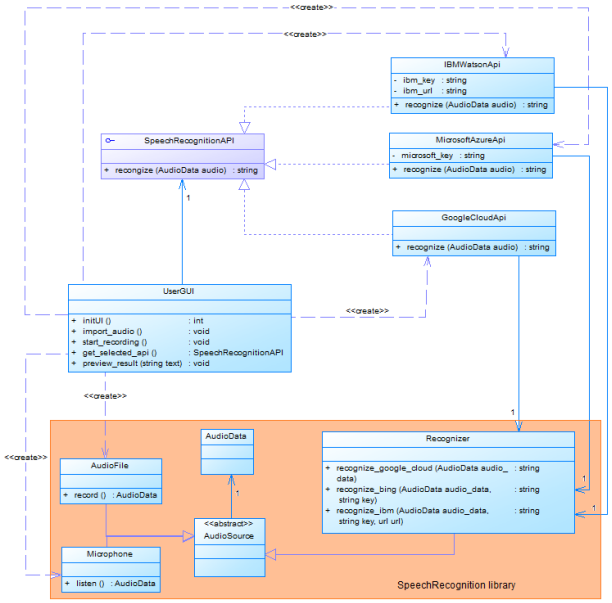
Детаљан опис решења проблема дат је кроз експериментелан пројекат у ком се примењују различити системи за препознавање говора, преко одговарајућих API-ја, захваљујући *SpeechRecognition* библиотеци. Спецификација пројекта дефинисана је кроз две фазе: анализа захтева и спецификација дизајна.

4.1 Анализа захтева

Главна сврха система за препознавање говора се дефинише приликом формулисања захтева у процесу анализе захтева.
Функционални захтеви апликације која представља експериментални пројекат у овом раду су: одабир API-ја за препознавање говора (*Google Cloud Speech to Text*, *Microsoft Azure Speech to Text* или *IBM Watson Speech to Text*), снимање говора путем микрофона, отпремање аудио датотеке, позивање API-ја за препознавање говора и приказивање резултата. Док функционални захтеви дефинишу шта систем ради, нефункционални захтеви одређују како систем то треба да ради. У нефункционалне захтеве спадају: латенција, тачност, прецизност, језичка подршка и цена.

4.2 Спецификација дизајна

Објектни дизајн дефинише решење које би требало да премости јаз између модела анализе и хардверске / софтверске платформе дефинисане током дизајнирања система. То укључује прецизан опис интерфејса објекта и подсистема. Слика 3 детаљно приказује дијаграм класа (eng. ***Class Diagram***) апликације за препознавање говора која се анализира у овом раду.



Слика 3. *Дијаграм класа експерименталне апликације*

4.3 SpeechRecognition библиотека

Једна од најпопуларнијих *Python* библиотека за препознавање говора јесте ***SpeechRecogniton*** библиотека, која се налази на PyPI репозиторијуму. Она олакшава преузимање аудио улаза из ког се врши препознавање говора. Препознавање говора захтева унос аудио звука, а *SpeechRecogniton* олакшава његово преузимање. Уместо прављења скрипте за приступ микрофонима и обраду аудио датотека од нуле, *SpeechRecogniton* обавља те процесе за само неколико минута [4].

4.3.1 Recognizer класа

Све функционалности у *SpeechRecognition* библиотеци врше се преко ***Recognizer*** класе.

Инстанца *Recognizer* класе долази са разним поставкама и функцијама за препознавање говора из аудио извора (аудио датотека или говора преко микрофона) [4].
Свака *Recognizer* инстанца има седам метода за препознавање говора из аудио извора помоћу различитих API-ја [4]:

- recognize_google(): Google Web Speech API
- recognize_google_cloud(): Google Cloud Speech
- recognize_bing(): Microsoft Bing Speech
- recognize_ibm(): IBM Speech to Text
- recognize_houndify(): Houndify by SoundHound
- recognize_sphinx(): CMU Sphinx – захтева инсталацију PocketSphinx
- recognize_wit(): Wit.ai

4.4 Имплементација

Како се сви наведени API-ји, који се користе у експерименталној апликацији, налазе у "облаку", било је потребно имплементирати и аутентификацију за сваки од њих. Приликом аутентификације добијају се додатни параметри за позивање API-ја.
Препознавање говора преко *Google CloudApi*-ја обавља готова ***recognize_google_cloud*** метода класе *Recognizer*. Препознавање говора преко *MicrosofAzureApi*-ја обавља готова ***recognize_bing*** метода класе *Recognizer*. Препознавање говора у оквиру *IBMWatsonApi*-ја обавља готова ***recognize_ibm*** метода класе *Recognizer*.

4.5 Резултати тестирања

Како се подразумева да функционални захтеви система за препознавање говора увек буду испуњени, у оквиру евалуације примењених API-ја у овом експерименталном пројекту, акценат ће бити више дат на нефункционалним захтевимаНа основу њих су формирани критеријуми евалуације, по којима је вршено тестирање.
Узорак говора је добијен из аудио књиге „Emma" од Jane Austin и дат је као улаз у апликацију. Узето је 19. поглавље књиге чији снимак траје 9 минута 7 секунди. Узорак говора снимљен је у тихом окружењу са микрофоном.
Што се тиче перформанси сервиса, Google Cloud STT и Microsoft Azure STT сервиси прелазе невероватних 97% тачности препознавања речи, док IBM Watson STT сервис има нешто мању тачност од 96.9% што је и даље задивљујућ резултат.
Време које је потребно за конверзију говора у текст је најмање код IBM Watson STT сервиса, док је Google Cloud STT бржи од Microsoft Azure STT сервиса за око 1 секунду.
Највећу предност је Google Cloud сервис за препознавање говора остварио у области језичке подршке где подржава скоро дупло више језика од Microsoft Azure STT сервиса. IBM Watson STT сервис подржава неупоредиво мање језика од претходна два.
Различите компаније система за препознавање говора наплаћују различито своје услуге. Ако се цена услуге посматра у форми $/минуту, долази се до закључка да Microsoft компанија има најповољнију цену својих услуга (0.016 $/минуту), док је Google најскупљи (0.036 $/минуту).

5. ЗАКЉУЧАК

У овом раду представљено је широко теоријско сазнање у области препознавања говора, које спада у једно од најактуелнијих области данас. Циљ рада је био да се упореде три најпопуларнија *cloud* сервиса за препознавање говора, развијаних од стране светских гигант компанија (Google, Microsoft и IBM).
Практично решење је била апликација у *Python* програмском језику, унутар које су имплементирани позиви ових сервиса преко њихових API-ја.
Објашњена је примена *SpeechRecognition* библиотеке преко које су извршени API позиви ових сервиса.
На основу резултата евалуације примењених сервиса, долази се до закључка да се не може посебно издвојити ни један сервис за препознавање говора који је најбољи по свим критеријумима.
Све је већа потражња за интерактивним говорним системима који укључују дијалог између људи и рачунара. У будућности ће бити потребно већа природност како у језику које човек користи, тако и у одговорима које генерише систем. Таква природност је вероватно остварива само ако рачунар има добар модел интеракције. Потешкоће се обично сматрају општим проблемима вештачке интелигенције. Иако су достигнућа на овом подручју импресивна, системима за препознавање говора се постављају додатни изазови.

6. ЛИТЕРАТУРА

[1] U. Kamath, J. Liu, J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Switzerland: Springer International Publishing. 2019. [Преузето: 1.09.2020]

[2] S. Sen, A. Dutta,N. Dey, *Audio Processing and Speech Recognition*. Singapore: Springer. 2019. [Преузето: 1.09.2020]

[3] J. Holmes, W. Holmes, *Speech Syntesis and Recognition*. New York: Taylor & Francis. 2003. [Преузето: 1.09.2020]

[4] "The Ultimate Guide To Speech Recognition With Python" https://realpython.com/python-speech-recognition/#recap-and-additional-resources. Приступљено октобра 2020

Кратка биографија:

Милан Сувајџић рођен је у Врбасу 1995. године. Мастер рад на факултету техничких наука из области Електротехнике и рачунарства – Рачунарство и аутоматика одбранио је 2020. год.