Зборник радова Факултета техничких наука, Нови Сад

**УДК: 4.934**
**ДОИ: https://doi.org/10.24867/35BE23Vitkovic**

# Анализа перформанси Whisper модела у превођењу српског говора на енглески језик

# *Performance analysis of the Whisper model in translating Serbian speech into English*

Милана Витковић, *Факултет техничких наука, Нови Сад*

**Студијски програм – ЕНЕРГЕТИКА, ЕЛЕКТРОНИКА И ТЕЛЕКОМУНИКАЦИЈЕ**

**Кратак садржај –** *Овај рад анализира методе аутоматског превођења српског говора на енглески језик, фокусирајући се на Whisper моделе и велике језичке моделе. Поређени су директни end-to-end приступ и каскадни систем који раздваја препознавање говора и превођење. Евалуација је спроведена на више тест скупова користећи BLEU, METEOR и WER метрике. Резултати показују да каскадни приступ са Whisper транскрипцијом и преводом помоћу великих језичких модела даје боље перформансе од директног превођења Whisper модела. Дообука Whisper модела за транскрипцију на српском значајно побољшава квалитет транскрипције и превода. Велики језички модели показују робусност чак и уз грешке у транскрипцији, што указује на њихов потенцијал у овој области. Као будући кораци препоручује се даља дообука и интеграција језичких модела ради унапређења квалитета превођења.*

**Кључне речи (три до пет):** *аутоматско превођење говора, whisper модел,*

**Abstract** – *This paper analyzes methods for automatic speech translation from Serbian to English, focusing on Whisper models and large language models (LLMs). It compares a direct end-to-end approach with a cascade system that separates speech recognition and translation. Evaluation was conducted on multiple test sets using BLEU, METEOR, and WER metrics. Results show that the cascade approach, combining Whisper transcription and LLM translation, outperforms direct Whisper translation. Fine-tuning Whisper for automatic speech recognition in Serbian significantly improves transcription and translation quality. LLMs demonstrate robustness even with transcription errors, highlighting their potential in this domain. Future work includes further fine-tuning and integration of language models to enhance translation accuracy.*

**Keywords: (three to five):** *automatic speech translation, whisper model*

---

**НАПОМЕНА: Овај рад проистекао је из мастер рада чији ментор је био др Синиша Сузић, доцент.**

## 1. УВОД

Аутоматско превођење говора (енг. *automatic speech translation*, AST) представља једно од најдинамичнијих истраживачких подручја у оквиру технологија говора и природног језика, подстакнуто напретком дубоког учења и развојем великих мултијезичких модела. Традиционални системи за превођење говора најчешће се заснивају на каскадној архитектури која обухвата два одвојена корака: препознавање говора (АПГ; енг. *automatic speech recogntion*, ASR) и машинско превођење (МП; енг. *machine translation*, MT). У овом приступу, говорни сигнал се најприје транскрибује у текст, након чега се добијена транскрипција преводи на циљни језик. Иако ефикасни, каскадни системи су подложни проблемима пропагације грешака, јер свака грешка настала у фази транскрипције директно утиче на квалитет финалног превода [1, 2].

Развој *end-to-end* неуронских архитектура омогућио је да се превођење говора у текст врши у оквиру једног модела. Ови модели, који се у великој мјери ослањају на трансформер архитектуру [3], истовремено уче акустичке и лингвистичке представе, што резултује већом робусношћу и бољом кохерентношћу превода у односу на традиционални приступ. Ипак, овакви модели захтијевају велике количине паралелних говорно-текстуалних података, који су често недоступни за језике са ограниченим ресурсима, као што је српски [1-3].

С обзиром на ове изазове, важно је анализирати како се постојећи АПГ и МП системи, како каскадни, тако и *end-to-end*, понашају на задацима превођења српског говора. Ово је посебно значајно у сценаријима гдје недостају велики паралелни корпуси и гдје се модели ослањају на мултилингвалну предобуку или *zero-shot* способности. *Zero-shot* приступ означава способност модела да извршава задатке превођења за језике или језичке парове који нису били директно обухваћени током обуке. Ова способност се заснива на мултилингвалној предобуци и заједничким семантичким репрезентацијама које омогућавају пренос знања између различитих језика. У том контексту, овај рад испитује перформансе различитих система, *Whisper* модела [4], *Google*

*Translate*-а и савремених модела великих језика (енг. *Large Language Model*, LLM). Посебна пажња посвећена је упоређивању директног превођења говора и каскадног приступа који комбинује транскрипцију помоћу *Whisper*-а и превођење коришћењем великих језичких модела. Циљ рада је да се идентификују предности и ограничења сваког од наведених приступа.

## 2. WHISPER МОДЕЛ

*Whisper* представља савремени модел за аутоматско препознавање и превођење говора, развијен од стране компаније *OpenAI*. Обучен је на око 680.000 часова аудио података на 99 различитих језика, прикупљених у условима слабог надгледаног учења, што му омогућава изузетну робусност и стабилне перформансе на различитим језицима, акцентима и условима снимања. Модел постоји у више конфигурација (*tiny*, *base*, *small*, *medium*, *large*, *large-v2*), које се разликују по укупном броју параметара. Мањи модели су доступни у енглеској и вишејезичној варијанти, док су највећи модели тренирани искључиво у вишејезичном режиму. Вишејезични модели подржавају и препознавање говора и директно превођење говора на енглески језик. Као *end-to-end* систем, *Whisper* интегрише препознавање језика, транскрипцију, превођење и сегментацију говора унутар једне архитектуре, што га чини једноставнијим за употребу у односу на класичне АПГ системе који се ослањају на више одвојених компоненти [4].

### 2.1. Архитектура модела

*Whisper* је заснован на кодер–декодер трансформер архитектури, која се показала изузетно успјешном у задатку секвенцијалног моделовања. Улазни аудио сигнал се најприје ресемплује на 16 kHz и претвара у 80-канални логаритамски мел спектрограм, који представља стандардна обележја у обради говорног сигнала [5]. Овај спектрограм се нормализује и служи као улаз у кодер, чији први слојеви користе плитку конволуциону структуру [6] са GELU (енг. *Gaussian Error Linear Unit*) активацијом [7] како би се смањила временска резолуција и припремио сигнал за серију трансформер блокова. У тим блоковима, уз примјену синусоидалних позиционих ембединга [8] и резидуалних веза [9], модел гради високоапстрактне и контекстуализоване репрезентације аудио садржаја.

Декодер функционише као ауторегресивни језички модел који, на основу информација из кодера и претходно генерисаних токена, производи текст корак по корак. Он се ослања на механизам пажње који омогућава ефикасно повезивање звучног контекста са језичким обрасцима. *Whisper* подржава више различитих задатака унутар једне архитектуре, па се одређени задатак моделу задаје путем специјалних токена. Свака обрада почиње токеном <|*startoftranscript*|>, након чега се додаје токен језика (нпр. <|*sr*|> или <|*en*|>), затим токен који одређује да ли модел треба да транскрибује (<|*transcribe*|>) или преводи (<|*translate*|>), као и токен <|*notimestamps*|> уколико се не предвиђају временске одреднице. На тај начин модел може флексибилно да прелази између различитих функција као што су препознавање говора, превођење, идентификација језика или детекција тишине [4].

Текст се на излазу генерише коришћењем *byte-level* BPE (енг. *Byte Pair Encoding*) токенизације [10, 11], која је прилагођена да равноправно обрађује и енглеске и вишејезичне моделе. Ова врста токенизације омогућава ефикасно кодирање различитих писама и језика, што омогућава да *Whisper* стабилно функционише и на српском језику [4].

## 3. ЕВАЛУАЦИОНЕ МЕТРИКЕ

Процјена квалитета система за аутоматско превођење говора захтијева употребу поузданих и уједно осјетљивих мјера које могу да обухвате лексичку, структурну и семантичку тачност превода. У овом раду примијењене су три најзаступљеније метрике у области машинског превођења и препознавања говора: BLEU, METEOR и WER. Наведене метрике међусобно се допуњују и омогућавају свеобухватну евалуацију различитих модела и приступа.

**BLEU** (енг. *Bilingual Evaluation Understudy*) [12] мјери сличност између генерисаног превода и референтног превода на основу преклапања n-грама. Основна претпоставка је да ће квалитетан превод садржати значајан степен преклапања са референтним преводом у виду појединачних ријечи и фраза различите дужине. Општи облик формуле гласи [12]:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \qquad (1)$$

гдје $p_n$ представља процењат одговарајућег броја n-грама садржаног у кандидату у поређењу са референтним преводом, $w_n$ тежинске коефицијенте (у пракси се обично рачуна као $1/N$ ), а $BP$ фактор казне за краткоћу, који умањује резултат уколико је генерисани превод значајно краћи од референце.

За разлику од BLEU метрике, која даје предност лексичком преклапању, **METEOR** (енг. *Metric for Evaluation of Translation with Explicit ORdering*) тежи да обухвати семантичку сличност између превода и референце. У процесу поређења узима у обзир тачно поклапање ријечи, лематизоване облике, синониме, фразну непрекидност и редослијед ријечи [13].

Метрика се заснива на хармонијској средини прецизности и одзива, на коју се примјењује казнени фактор који умањује резултат уколико су поклапајуће ријечи распоређене у већем броју неповезаних сегмената. На тај начин METEOR кажњава фрагментоване преводе, односно преводе у којима је редослијед ријечи или структура значајно нарушена. Формално, метрика се изражава као [13]:

$$METEOR = F_{mean} \cdot (1 - Penality), \qquad (2)$$

гдје $F_{mean}$ представља хармонијску средину прецизности и одзива, а $Penalty$ казну која расте са степеном фрагментације превода.

Ова метрика се сматра посебно погодном за језике богате морфологијом, као што је српски, јер боље препознаје сродност између различитих облика исте ријечи и мање је осјетљива на варијације у синтаксичком редослиједу.

**WER** (енг. *Word Error Rate*) је стандардна метрика за оцјену квалитета препознавања говора, али се може примијењивати и на анализу структурне тачности превода. Она мјери разлику између генерисаног текста и референтног текста, на основу броја извршених операција уређивања [14]:

$$WER = \frac{S + D + I}{N}, \quad (3)$$

гдје $S$ представља замјене ријечи (substitutions), $D$ брисања ријечи (deletions), $I$ уметања ријечи (insertions) и $N$ укупан број ријечи у референтном тексту [14].

## 4. ПРЕВОЂЕЊЕ СРПСКОГ ГОВОРА У ЕНГЛЕСКИ ТЕКСТ ПОМОЋУ WHISPER МОДЕЛА

Аутоматско превођење говора са српског језика представља изазован задатак због језичких специфичности и релативно ограничене заступљености српског у великим мултилингвалним корпусима. Иако *Whisper* спада у најнапредније моделе за препознавање и превођење говора, количина података на српском била је ограничена, што може утицати на перформансе у сценаријима без додатне дообуке.

Међутим, захваљујући свом опсежном мултилингвалном предтренингу и способности да учи опште акустичке и језичке обрасце, *Whisper* је у стању да обавља транскрипцију и превођење српског говора у *zero-shot* режиму. Управо то омогућава његово тестирање и поређење са другим системима, иако је српски био ограничено заступљен (28 сати за транскрипцију и 136 сати за превођење) у оригиналним тренинг подацима [4].

Постојећа литература показује да модели обучени на универзалним, доминантно енглеским и високоресурсним корпусима често имају потешкоћа са језицима који имају богату флексију, слободнији редослијед ријечи и специфичне морфосинтаксичке структуре, што су управо карактеристике српског језика. Због тога се у бројним студијама истиче значај адаптације модела (дообука на језички специфичним подацима или доменска адаптација), јер она значајно побољшава тачност, флуентност и робусност система за аутоматско превођење говора за нискоресурсне језике као што је српски [15-17].

### 4.1. Поређење постојећих метода превођења

У првој фази истраживања извршено је поређење више приступа превођењу говора са српског на енглески језик. Анализирани су *Whisper-small* (оригинални модел), *Whisper-medium* (оригинални модел), *Google Translate* (текст → текст), *GPT-5 mini* (текст → текст).

У овој поставци Whisper функционише у режиму директног превођења говора (енг. *speech-to-text translation*): модел као улаз добија аудио запис на српском језику, а затим директно генерише енглески текст, без посредног АПГ корака на српском. Овај *zero-shot* приступ омогућава процјену способности модела да преводи са језика који је био ограничено заступљен у тренинг подацима.

Евалуација је спроведена над три различита тест скупа:

- Јужне вести – 50 реченица из домена новинарског, спонтаног говора [1]
- Common Voice – 30 кратких диктираних реченица са различитим говорницима [2]
- FLEURS – 30 реченица из стандардизованог мултилингвалног корпуса [3].

Сваки скуп представља различите стилове говора (спонтани новинарски, диктирани, контролисани), што омогућава процјену робусности модела у више реалистичних сценарија.

Табела 1. Поређење различитих метода превођења – Јужне вести (50 реченица)

| Model | BLEU | METEOR | WER |
|---|---|---|---|
| Whisper-small | 13.88 | 33 | 80 |
| Whisper-medium | 21.74 | 44 | 72 |
| Google Translate | 29.33 | 59.89 | 68.64 |
| GPT-5 mini | 47.05 | 72.18 | 45.86 |

Табела 2. Поређење различитих метода превођења – Common Voice (30 реченица)

| Model | BLEU | METEOR | WER |
|---|---|---|---|
| Whisper-small | 18.28 | 32 | 86 |
| Whisper-medium | 24.78 | 47 | 71 |
| Google Translate | 41.97 | 67.19 | 43.70 |
| GPT-5 mini | 52.40 | 77.33 | 36.25 |

Табела 3. Поређење различитих метода превођења – Fleurs (30 реченица)

| Model | BLEU | METEOR | WER |
|---|---|---|---|
| Whisper-small | 22.46 | 44 | 78 |
| Whisper-medium | 36.47 | 62 | 56 |
| Google Translate | 66.79 | 88.14 | 24.88 |
| GPT-5 mini | 82.60 | 95.14 | 11.10 |

Перформансе су мјерене коришћењем BLEU, METEOR и WER метрика (Табеле 1, 2 и 3) у односу на референтни превод који је ручно креиран, прилагођен слободном и природном говору. Овај приступ омогућава објективнију и реалнију процјену квалитета машинских превода у контексту стварне употребе, избјегавајући претјерано дословне или неприродне преводе као референту вриједност. Резултати показују да оригинални *Whisper* модели постижу умјерен квалитет директног превођења, при

[1] https://www.clarin.si/repository/xmlui/handle/11356/1679?locale-attribute=en

[2] https://datacollective.mozillafoundation.org/datasets/cmflnuzw7y2kjfhed2p152qe5

[3] https://huggingface.co/datasets/google/fleurs/tree/main/data/sr_rs

чему *Whisper-medium* у свим случајевима надмашује *Whisper-small*. Ово је очекивано с обзиром на већи моделски капацитет и боље мултилингвалне репрезентације.

С друге стране, *Google Translate* и *GPT-5 mini* остварују значајно боље резултате, што се може приписати огромним паралелним корпусима на којима су обучени и њиховој способности да прецизно моделују структуру енглеског и српског језика у текстуалном домену. Ова разлика потврђује да директно *zero-shot* превођење говора представља изазов за српски, нарочито у одсуству језички специфичне обуке.

Иако *Whisper* не достиже перформансе специјализованих текстуалних система, важно је нагласити да један модел истовремено извршава препознавање говора и превођење, чиме представља ефикасну основу за даљи развој *end-to-end* система за аутоматско превођење говора на нискоресурсним језицима попут српског.

### 4.2. Превођење у два корака

У другој фази истраживања примјењен је каскадни приступ превођењу у два корака. Прво, *Whisper* модел је као улаз добијао аудио запис на српском језику и извршавао транскрипцију на српски. Након тога, добијени српски транскрипт је коришћен као улаз за превођење на енглески језик помоћу великог језичког модела, конкретно, *ChatGPT-4o*. Тиме је омогућена независна процјена утицаја квалитета транскрипције на коначни превод.

Тестиране су сљедеће варијанте *Whisper* модела: *Whisper-small* (оригинални), *Whisper-medium* (оригинални), *Whisper-small* (дообучен за транскрипцију на српском), *Whisper-medium* (дообучен за транскрипцију на српском).

За дообуку *Whisper-small* модела за транскрипцију на српском коришћен је *Common Voice* корпус, који садржи аудио снимке и валидиране транскрипције на српском језику. Модел је трениран у трајању од 5 епоха, уз стандардне параметре оптимизације. За дообуку *Whisper-medium* модела за транскрипцију на српском језику коришћена је комбинација тренинг скупа Јужне вести и корпуса *ParlaSpeech*. Модел је дообучен у трајању од 3 епохе. У оба случаја, током тренинга пратиле су се перформансе модела, а као метрика за одабир најбоље епохе коришћен је WER, при чему је модел са најнижом вриједношћу WER у евалуационој фази сачуван као финална верзија.

Табела 4. Поређење различитих модела за превођење у 2 корака – Јужне вести тест скуп

| Model | BLEU | METEOR | WER |
|---|---|---|---|
| Whisper-small | 43.58 | 70.39 | 47.41 |
| Whisper-medium | 50.36 | 75.9 | 40.87 |
| Whisper-small finetuned | 43.83 | 71.53 | 46.3 |
| Whisper-medium finetuned | 51.11 | 76.84 | 37.99 |

Перформасе на тест скупу „Јужне вести“ су приказане у табели 4. Анализа показује да каскадни приступ значајно надмашује директно превођење *Whisper* модела из прве фазе. *Whisper-medium* модели доследно постижу боље резултате од *Whisper-small*, што је у складу са већим капацитетом и бољом мултилингвалном репрезентацијом.

Дообука *Whisper* модела за српски АПГ резултира побољшаним WER показатељима, што директно утиче на бољи квалитет превода мјерен BLEU и METEOR метрикама. Нарочито је видљив раст METEOR резултата, који је важан за српски језик јер боље одражава морфолошка и семантичка подударања. Такође, велики језички модел показује висок степен робусности у превођењу чак и када транскрипција *Whisper* модела садржи грешке.

Овим се потврђује да је за српски језик тренутно најпогоднији каскадни АПГ → МП приступ, док директно *end-to-end* превођење још увијек није довољно зрело за високу тачност.

## 5. ЗАКЉУЧАК

У овом раду испитани су различити приступи аутоматском превођењу српског говора на енглески језик. Резултати показују да директно превођење *Whisper* моделима има ограничен квалитет, док каскадни приступ који комбинује *Whisper* за транскрипцију и велики језички модел (*ChatGPT-4o*) за превод даје значајно боље резултате. Дообука *Whisper* модела за српски АПГ смањује грешке у препознавању, што позитивно утиче на квалитет превода, посебно мјере METEOR која је погодна за српски језик.

Велики језички модели показују висок ниво робусности у превођењу, чак и када транскрипције садрже грешке, што истиче потенцијал за даљу интеграцију ових технологија. Тренутно, каскадни АПГ → МП приступ представља најпоузданије рјешење за српски језик у условима ограничених ресурса.

Као сљедећи кораци у истраживању препоручују се: даља дообука *Whisper* модела на већим и разноликијим српским говорним корпусима, експериментисање са различитим великим језичким моделима, развој *end-to-end* модела прилагођених специфичностима српског језика, као и креирање српско–енглеског паралелног корпуса који садржи транскрипте на српском и њихове одговарајуће енглеске преводе. Овај корпус би омогућио циљану дообуку и адаптацију модела за задатак превођења говора, што би значајно допринијело побољшању прецизности, природности и робусности система за нискоресурсне језике.

ЛИТЕРАТУРА

[1] A. B´erard, O. Pietquin, C. Servan, and L. Besacier, “Listen and translate: A proof of concept for end-to-end speech-to-text translation,” in NIPS Workshop on

End-to-End Learning for Speech and Audio Processing, 2016.

[2] R. J. Weiss, J. Chorowski, N. Jaitly, Y. Wu, and Z. Chen, “Sequence-to-sequence models can directly translate foreign speech,” in Interspeech. ISCA, 2017, pp. 2625–2629.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” arXiv preprint arXiv:2212.04356, 2022.

[5] Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and trends® in signal processing*, *7*(3–4), 197-387.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016, ch. 9: “Convolutional Networks”.

[7] Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415, 2016.

[8] Press, O. and Wolf, L. Using the output embedding to improve language models. In Proceedings of the 15th Conference of the European Chapter of the Associa tion for Computational Linguistics: Volume 2, Short Papers, pp. 157–163, Valencia, Spain, April 2017. As sociation for Computational Linguistics. URL https: //aclanthology.org/E17-2025.

[9] Child, R., Gray, S., Radford, A., and Sutskever, I. Gen erating long sequences with sparse transformers. arXiv preprint arXiv:1904.10509, 2019.

[10] Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909, 2015.

[11] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[13] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evalua-tion with improved correlation with human judgments,” in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.

[14] A. Ali and S. Renalds, “Word Error Rate Estimation for Speech Recognition: e-WER,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pp. 20–24 Melbourne, Australia, July 15 - 20, 2018.

[15] V. Timmel, C. Paonessa, M. Vogel, D. Perruchoud and R. Kakooe, “Fine-tuning Whisper on Low-Resource Languages for Real-World Applications,” *arXiv preprint* arXiv:2412.15726, Apr. 2025.

[16] R. Ma, M. Qian, Y. Fathullah, S. Tang, M. Gales and K. Knill, “Cross-Lingual Transfer Learning for Speech Translation,” *arXiv preprint*, arXiv:2407.01130, Feb. 2025.

[17] P. Peng, B. Yan, S. Watanabe and D. Harwath, “Prompting the hidden talent of web-scale speech models for zero-shot task generalization,” *arXiv preprint*, arXiv:2305.11095v3, Aug. 2023.

**Кратка биографија:**

**Милана Витковић** је рођена 20.10.2001. године у Требињу. Основне академске студије је завршила у октобру 2024. године на Факултету техничких наука у Новом Саду. Мастер студије на Факултету техничких наука из области обраде сигнала уписала је 2024. године.
**Контакт**:
milanavitkovic2001@gmail.com