Зборник радова Факултета техничких наука, Нови Сад

УДК: ()
ДОИ: ()

АНАЛИЗА И ОПТИМИЗАЦИЈА Wav2vec 2.0 МОДЕЛА ЗА ПРЕПОЗНАВАЊЕ ГОВОРА НА СРПСКОМ ЈЕЗИКУ

ANALYSIS AND FINE-TUNING OF THE Wav2vec 2.0 MODEL FOR SERBIAN SPEECH RECOGNITION

Тамара Бановац, Владо Делић, *Факултет техничких наука, Нови Сад*

Област – ЕНЕРГЕТИКА, ЕЛЕКТРОНИКА И ТЕЛЕКОМУНИКАЦИЈЕ

Кратак садржај – *Овај рад се бави применом Wav2vec 2.0 модела за препознавање говора на српском језику. Рад обухвата анализу оригиналног модела, обученог техником самонадгледаног учења на нелабелираним подацима, и fine-tuning фазу на српском језику. Оригинална имплементација, са 53000 сати нелабелираних и 10 минута лабелираних података, постигла је WER од 4.8/8.2, што потврђује ефикасност коришћених метода. Циљ је испитати применљивост ових метода на аудио подацима на српском језику. Постигнути резултати на српском језику показују WER од 10.3% и CER од 3.4%, уз могућност даљих унапређења.*

Кључне речи: *Препознавање говора, самонадгледано учење, српски језик, Wav2vec модел, трансформер*

Abstract – *This paper focuses on the application of the Wav2vec 2.0 model for speech recognition in Serbian. It includes an analysis of the original model, trained using a self-supervised learning technique on unlabeled data, and the fine-tuning phase for Serbian. The original implementation, with 53,000 hours of unlabeled and 10 minutes of labeled data, achieved a WER of 4.8/8.2, demonstrating the effectiveness of the applied methods. The goal is to explore the applicability of these methods to Serbian audio data. The results for Serbian show a WER of 10.3% and a CER of 3.4%, with room for further improvements.*

Keywords: *Speech recognition, self-supervised learning, Serbian language, Wav2vec model, transformer*

1. Увод

Надгледано учење је техника која користи унапред лабелиране податке, где свака улазна вредност има одговарајућу ознаку. Иако даје одличне резултате у областима попут препознавања говора, често је тешко обезбедити довољну количину лабелираних података за тренинг напредних модела. Самонадгледано учење (енг. *self-supervised learning*, *SSL*) решава овај проблем тако што омогућава моделима да се тренирају на великим количинама нелабелираних података, чиме се генеришу репрезентације података.

НАПОМЕНА:
Овај рад настао је из мастер рада чији ментор је био проф. др Владо Делић.

Један од најпознатијих *SSL* модела је *Wav2Vec 2.0* [1], који је трениран на нелабелираним сировим аудио подацима, а касније дообучен за препознавање говора користећи релативно мало лабелираних података. Оригинална имплементација модела, са 53000 сати нелабелираних и 10 минута лабелираних података, постигла је *WER* од 4.8/8.2.

У овом раду истражује се ефективност ових метода на српском језику. Коришћена су два *Wav2Vec 2.0* модела, са основном и проширеном архитектуром, а они су доступни у оквиру *torchaudio* библиотеке. Модели су дообучени на три базе података: *Common Voice Corpus 18.0* [2] (7 сати валидираних аудио података), Јужне Вести [3] (50.55 сати са одређеним непоклапањима између транскрипта и снимака) и *Serbian ParlaSpeech-RS* [4] (896 сати аудио снимака, такође са одређеним непоклапањима).

У наставку је описана архитектура *Wav2Vec 2.0* [1] модела и представљен је и процес тренинга, који обухвата фазу претренирања на великој количини нелабелираних података, као и фазу дообуке. Посебно је анализиран процес дообуке модела за препознавање говора на српском језику, као и коришћене базе података. Након тога, су приказани резултати које модел постиже, уз анализу метрика као што су стопа грешке у препознавању карактера и речи.

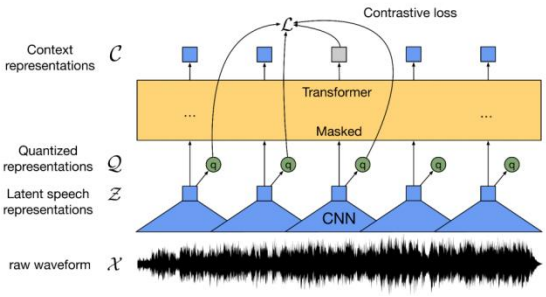
2. Архитектура *WAV2VEC 2.0* модела

Wav2Vec 2.0 модел се састоји из неколико основних целина које заједно омогућавају ефикасно учење репрезентација података погодних за различите задатке, укључујући и препознавање говора. Модел на свом улазу прима сирове аудио податке у временском облику. Прва целина је кодер обележја који се заснива на седмослојној конволуционој неуронској мрежи, а њена главна функција је да смањи димензионалност аудио података. Овај кодер пресликава скуп улазних података X у скуп вектора обележја Z. Скуп Z се састоји од T вектора $z_1, z_2, \ldots, z_T$, где је T број фрејмова који се налазе у улазном аудио фајлу, а сваки вектор z_t је повезан са по једним временским фрејмом x_t.

Након тога се примењује маскирање, где се одређени вектори обележја z_t постављају на нулу и затим шаљу на улаз трансформера [5] чија је функција да предвиди маскиране векторе. Трансформер генерише контекстуалне репрезента $c_1, c_2, \ldots, c_T$ које представљају предвиђања за маскиране векторе. Како

би се омогућило учење модела, вектори обележја z_t се квантизују путем квантизационог модула, који непрекидне вредности вектора $z_1, z_2, \ldots, z_T$ пресликава у дискретне јединице $q_1, q_2, \ldots, q_T$.

Квантизоване вредности $q_1, q_2, \ldots, q_T$ се користе за израчунавање контрастивног губитка (енг. *contrastive loss*) [6], који је кључан за обуку модела. Овај губитак пореди предвиђања трансформера са квантизованим векторима и мери степен сличности између њих, чиме модел учи репрезентације података. На слици 1 дат је приказ архитектуре модела.



Слика 1. Архитектура *Wav2vec 2.0* модела.

3. ТРЕНИНГ МОДЕЛА

Тренинг *Wav2Vec 2.0* модела за аутоматско препознавање говора се састоји из 2 фазе. Прва фаза обуке, позната и као претренирање, има за циљ учење општих репрезентација података које су довољно информативне да се могу користити за решавање различитих задатака машинског учења. Да би се ово постигло модел се тренира тако да решава контрастивни задатак (L_m) [6] уз коришћење неозначених података. Поред контрастивног губитка користи се и додатни губитак L_d (*diversity loss*), који подстиче модел да равномерно користи све елементе из кодних књига. Финална функција губитка која се минимизује у овом делу обуке модела има следећи облик:

$$L = L_m + \alpha L_d, \qquad (1)$$

где је α хиперпараметар.

Након претренирања, модел пролази кроз фазу дообуке (енг. *fine-tuning*), где се на врх трансформера додаје на случајно иницијализовани линеарни слој. Трансформер на свом излазу генерише матрицу димензија $N \times 768$ (односно $N \times 1024$ за проширену архитектуру), где је N број фрејмова у улазном аудио сигналу, а 768 представља димензију контекстуалних репрезентација за сваки фрејм. Линеарни слој који се додаје има димензије $768 \times C$, где је C број класа у речнику. Речник обухвата сва слова одговарајућег језика, у овом раду српског језика, размак и специјалне токене. Постоји само један такав линеарни слој, и исти скупови тежина примењују се паралелно на све векторе из матрице трансформера. Након примене линеарног слоја, сваки вектор димензије 768 се трансформише у вектор димензије C, што представља логит вредности за сваку класу у речнику. Затим се на ове логит вредности примењује *softmax* функција, која их претвара у вероватноће за сваку класу. Обука додатог слоја се изводи минимизацијом CTC губитка (*Connectionist Temporal Classification loss*) [7], која користи транскрипте како би модел био прилагођен за препознавање говора.

Након почетне обуке додатог линеарног слоја, препоручљиво је дообучити и све остале слојеве модела са малом брзином учења, такође користећи CTC губитак. Изузетак је кодера обележја који није потребно додатно обучавати. Овим процесом се постиже фино прилагођавање целокупног модела за конкретан задатак препознавања говора на изабраном језику.

4. ДООБУКА МОДЕЛА ЗА ПРЕПОЗНАВАЊЕ ГОВОРА НА СРПСКОМ ЈЕЗИКУ

У овом поглављу детаљно су описане коришћене базе података на српском језику, кораци предобраде података и параметри тренинга модела за препознавање говора на српском језику. Такође су приказани су резултати кроз метрике као што су *CER* (*Character Error Rate*) и *WER* (*Word Error Rate*).

4.1. Базе података

За дообуку модела за препознавање говора на српском језику коришћене су три различите базе. Прва база је *Common Voice Corpus 18.0* [2], која садржи 7 сати валидираних аудио података на српском језику, при чему се подаци одлично поклапају са транскриптима, друга је Јужне Вести [3], која има 50.55 сати аудио података и трећа и највећа база је S*erbian ParlaSpeech-RS 1.0* [4], која садржи 896 сати аудио снимака. У друге две базе постоје одређена непоклапања између транскрипта и аудио снимака, што отежава тренирање модела. У зависности од тога који модел је коришћен (модел са основном архитектуром или вишејезички *Wav2vec* модел), предобрада база се може мало разликовати због ограничених ресурса који су били доступни.

Common Voice Corpus 18.0 база садржи 2183 аудио снимака за тренинг, валидациони скуп садржи 1645 и тест садржи 1385 аудио снимака.

База Јужне вести садржи 40 сати аудио снимака за тренинг, док валидациони и тест скуп садрже по 5 сати аудио снимака. Најдужи аудио снимак има 30 секунди, а најкраћи 2.3 секунде. За тренинг модела са основном архитектуром коришћена је цела база, док су у случају тренинга вишејезичког модела из базе уклоњени предуги аудио снимци који су доводили до недостатка меморије. Након тог филтрирања базе престало је 25 сати у тренинг сету, 3 сата у валидационом и 3 сата у тест сету, а најдужи аудио снимак има 18.7 секунди.

У бази *ParlaSpeech-RS 1.0* подела на тренинг, валидациони и тест скуп није унапред доступна. Најдужи аудио снимак има 214.5 секунди, а најкраћи 0.1 секунду. Уклоњени су сви аудио снимци дужи од 20 секунди и краћи од 5 секунди. Такође, постојао је још један значајан проблем са базом. Сви бројеви који се изговарају су у транскрипту наведени као цифре уместо као речи. То представља проблем јер модел зна да тумачи само токенизоване транскрипте, а као токени се користе само слова из српске азбуке и

неколико специјалних токена. Одлучено је да се сви аудио снимци у којима се изговара било који број избаце из базе. Пошто ови снимци чине око 15% целокупне базе, таква одлука није значајно умањила обим доступних података. Након свих филтрирања, тренинг скуп садржи 316.2 сати аудио снимака, валидациони 56.6 сати и тест скуп садржи 22.4 сати аудио снимака.. Ова база је коришћена само за тренинг модела са основном архитектуром. Разлог је велика количина података због које је тренинг основног модела трајао око 37 сати, док би тренинг вишејезичког модела трајао 2 до 3 пута дуже. Због ограниченог времена доступног за израду овог рада, одлучено је да се вишејезички модел не тренира на бази *ParlaSpeech-RS*.

Како би се базе података адекватно припремиле за тренинг модела било је потребно извршити предобраду аудио снимака и транскрипта. Сваки аудио снимак је ресемплован на брзину која је коришћена за тренинг оригиналног модела (16000 Hz). Затим је сваки аудио снимак стандардизован на нулту средњу вредност и јединичну варијансу. Из транскрипта су отклоњени сви знаци интерпункције као и велика слова. Затим је транскрипт токенизован тако да сваком слову из азбуке одговара по један број. Поред токена за 30 слова азбуке и за размак између речи, користе се и специјални токени.

4.2. Токенизација

Токенизација је један од кључних корака у процесу креирања модела за препознавање говора. Током овог процеса, свако слово српске азбуке мапира се на одређени број, а користи се и токен за размак између речи. Поред ових 31 токена, користе се и специјални токени <*BLANK*> и <*UNK*>. *Blank* токен игра важну улогу у рачунању *CTC* губитка, и пожељно је да се мапира на 0. *UNK* токен се користи за означавање непознатих симбола који се могу јавити у транскрипту. Да би се омогућила обука модела кроз *batch* величине веће од 1, потребно је изједначити дужину свих аудио снимака и транскрипата унутар једног *batch*-а. <*BLANK*> токен може да се користи и као токен за *CTC* функцију, али и за обуку модела кроз *batch*-еве. У процесу дообуке мапирање токена је урађено на следећи начин: <*BLANK*> токен на 0, <*UNK*> токен на 1, слова азбуке на бројеве од 2 до 31 и размак између речи на 32.

Када се за обуку користе базе Јужне Вести и *ParlaSpeech-RS, batch* величина је увек постављена на 1 због дугих аудио снимака који трају до 20 и 30 секунди. Такви снимци онемогућавају коришћење већег *batch*-а због ограничених меморијских ресурса. У случају обуке на *Common Voice Corpus* бази користи се *batch* једнак 4.

4.3. Метрике

У процесу процене перформанси модела за препознавање говора користе су две основне метрике *Word Error Rate* (*WER*) и *Character Error Rate* (*CER*). *WER* је најчешће коришћена метрика у области препознавања говора. *WER* представља однос броја погрешних речи у препознатом тексту у односу на укупан број речи у оригиналном транскрипту. *CER* је метрика која мери проценат грешака на нивоу карактера. Ниже вредности метрика указује на боље перформансе модела.

4.4. Тренинг акустичког модела

У овом раду су испитане перформансе 2 модела. Први има основну архитектуру и претрениран је на 960 сати нелабелираних аудио података. Други коришћени модел има проширену архитектуру и обучаван је на вишејезичкој бази са 56000 сати нелабелираних аудио података. Први модел је треniран на све 3 описане базе, док је други модел треniран на базама Јужне вести и *Common Voice Corpus 18.0*.

Тренинг је рађен на следећи начин. На претренирани модел је додат један на случај иницијализовани линеарни слој, а затим је он треniран минимизацијом *CTC* губитка. Користи се *Adam* оптимизатор. После тога су искључане све претрениране тежине осим кодера обележја и додатно трениране са малом брзином учења, такође минимизацијом *CTC* губитка. Оба модела су тренирана одвојено на свакој од база. Када је тренинг на једној бази завршен, односно, када је постигнут минимални валидациони губитак, модел је сачуван и тренинг је настављен на следећој бази.

4.5. Резултати на тест сету

У овом поглављу је дат упоредни приказ резултата три модела са основном архитектуром, модел треniран само на *ParlaSpeech-RS* бази, затим тај модел дотрениран на бази Јужне вести и на крају модел дотрениран на бази *Common Voice Corpus 18.0*. Такође су приказани резултати коришћењем *greedy* декодовања и декодовања са 5-грам језичким моделом. Коришћење *greedy* декодовања заправо представља резултате које постиже сам акустички *Wav2Vec* модел, без икаквог знања о контексту које даје примена језичког модела. Као тест сет података коришћена је унија свих података из тест сетова три коришћене базе за обуку. Поред тога, посебно су приказати резултати на тест сету базе *Common Voice Corpus 18.0* јер је то једина база које има добро поравнате све аудио снимке и транскрипте, чиме обезбеђује најрелевантнији тест модела. Коришћене су метрике *CER* и *WER*. Примена језичког модела има првенствено утицај на *WER* метрику. Сви резултати су приказани у табели 1. *Common Voice Corpus 18.0* је у табелама наведена као *CV18.*

У случају коришћења уније сва три тест скупа постоји небалансираност података јер тест скуп из базе *ParlaSpeech-RS* има значајно више података у односу на друге 2 базе. Због тога су резултати сва три модела на таквом тест сету слични, односно, чини се да додатна обука модела на базама Јужне вести и *Common Voice Corpus 18.0* није унапредила резултате. Међутим, уколико се погледају резултати које модели постижу на тест сету базе *Common Voice Corpus 18.0* (која је и најквалитетнија), видимо да додатна обука модела на бази Јужне вести (после обуке на *ParlaSpeech-RS*) знатно унапређује резултате. Очекивано, после тренинга на *Common Voice Corpus*

18.0 резултати су још бољи. Такође, примена језичког модела значајно унапређује вредности *WER* метрике, поготово после теста на *Common Voice Corpus* бази.

Табела 1. Приказ резултата добијених обуком модела са основном архитектуром.

Модел трениран на базама	Тест сет	*greedy*		језички	
		CER	WER	CER	WER
ParlaSpeech-RS	Унија	0.386	0.587	0.379	0.556
	CV18	0.214	0.508	0.205	0.411
ParlaSpeech-RS + Јужне вести	Унија	0.381	0.593	0.372	0.556
	CV18	0.147	0.385	0.125	0.278
ParlaSpeech + Јужне вести + CommonVoice	Унија	0.379	0.591	0.368	0.553
	CV18	0.085	0.264	0.055	0.143

За вишејезички модел су приказани резултати по истом принципу као за модел са основном архитектуром. Разлика је у томе што за обуку овог модела није коришћена *ParlaSpeech-RS*, већ обука полази од базе Јужне вести. Упоредни приказ резултата дат је у табели 2.

Табела 2. Приказ свих резултата добијених обуком вишејезичког модела.

Модел трениран на базама	Тест сет	*greedy*		језички	
		CER	WER	CER	WER
Јужне вести	Унија	0.427	0.725	0.413	0.615
	CV18	0.110	0.390	0.084	0.210
Јужне вести + CommonVoice	Унија	0.415	0.710	0.392	0.600
	CV18	0.068	0.263	0.034	0.103

Резултати се могу интерпретирати слично као за модел са основном архитектуром. Резултати који овај модел постиже на *Common Voice Corpus 18.0* бази су бољи од резултата основног модела. То су уједно и најбољи постигнути резултати у овом истраживању. *CER* од 0.034 значи да модел погрешно транскрибује тек 3.4% свих карактера, док *WER* од 0.103 значи да модел погрешно предвиди око 10% речи. Ипак у случају теста на унији три базе, вишејезички модел даје лошије резултате у поређењу са основним. То је и очекивано јер у тој унији доминирају узорци из *ParlaSpeech-RS* базе.

5. ЗАКЉУЧАК

Wav2Vec 2.0 модел показује велики потенцијал за препознавање говора на језицима са ограниченим ресурсима као што је српски. Резултати су показали да вишејезички модел постиже боље резултате на *Common Voice Corpus 18.0* бази у поређењу са основним моделом. Уз примену језичког модела, стопа грешке у препознавању карактера (*CER*) је 0.034, док је стопа грешке у препознавању речи (*WER*) 0.103. Будућа истраживања треба усмерити на побољшање квалитета података, прецизније поравнање транскрипата и аудио снимака, као и на примену напреднијих језичких модела ради даљег унапређења перформанси система.

6. ЛИТЕРАТУРА

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Oct. 22, 2020, *arXiv*: arXiv:2006.11477. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/2006.11477

[2] "Common Voice." Accessed: Sep. 24, 2024. [Online]. Available: https://commonvoice.mozilla.org/en

[3] "ASR training dataset for Serbian JuzneVesti-SR v1.0." Accessed: Sep. 24, 2024. [Online]. Available: https://www.clarin.si/repository/xmlui/handle/11356/1679

[4] "Parliamentary spoken corpus of Serbian ParlaSpeech-RS 1.0." Accessed: Sep. 24, 2024. [Online]. Available: https://www.clarin.si/repository/xmlui/handle/11356/1834

[5] A. Vaswani *et al.*, "Attention Is All You Need," 2017, *arXiv*. doi: 10.48550/ARXIV.1706.03762.

[6] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," Jan. 22, 2019, *arXiv*: arXiv:1807.03748. Accessed: Sep. 24, 2024. [Online]. Available: http://arxiv.org/abs/1807.03748

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 369–376. doi: 10.1145/1143844.1143891.

Кратка биографија:

Тамара Бановац рођена је у Сремској Митровици 1999. год. Мастер рад на Факултету техничких наука из области Електротехнике и рачунарства – Обрада сигнала одбранила је 2024. године.
Контакт: tamarabanovac3@gmail.com

Владо Делић рођен је 1964. год. Докторирао је на Факултету техничких наука 1997. год., а од 2013. је у звању редовни професор. Област интересовања су говорне технологије.

UDK: 4.9
DOI: https://doi.org/10.24867/30BE40Banovac