Uticaj augmentacija trening podataka na performanse doobučenog Whisper modela

Siniša Suzić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
sinisa.suzic@uns.ac.rs
0000-0002-0511-6729

Tijana Nosek
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
tijana.nosek@uns.ac.rs
0000-0002-3707-0286

Nikola Simić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
nikolasimic@uns.ac.rs
0000-0002-0748-4672

Darko Pekar
Alfanum Ltd,Novi Sad
darko.pekar@alfanum.co.rs

Vlado Delić
Fakultet tehničkih nauka,
Univerzitet u Novom Sadu
vlado.delic@uns.ac.rs
0000-0002-4558-9918

***Apstrakt* - Whisper je jedan od najpopularnijih višejezičnih modela za automatsko prepoznavanje i transkripciju govora u poslednje vreme. Njegova popularnost zasnovana je na otvorenosti samog modela i podrški prepoznavanju velikog broja jezika. Međutim, pošto su količine materijala za obuku ovog modela za različite jezike bile veoma nesrazmerne i performanse prepoznavanja se za različite jezike dosta razlikuju. U ovom radu je prikazan pozitivan uticaj doobuke medium varijante Whisper modela na vrednosti verovatnoće greške na nivou reči, i to za slučaj srpskog jezika, koji predstavlja tipičan primer slabo zastupljenog jezika u obuci inicijalnog modela. Posebno je ispitan i uticaj dodavanja novih podataka za obuku korišćenjem sledećih tipova augmentacija – kompresija, dodavanje šuma, reverberacija. Pokazano je da dodavanje augmentovanih podataka u slučaju prepoznavanja javno dostupnih baza dovodi do smanjenja verovatnoće greške. Sa druge strane, u slučaju testiranja doobučenog modela na realnim primerima, uticaj augmentacije podataka za obuku na performanse modela nije konzistentan, što naglašava potrebu za dodatnim testiranjima korišćenjem i nekih drugih postupaka augmentacije.**

Ključne reči – ASR, doobuka, augmentacija.

I. UVOD

Automatsko prepoznavanje govora (ASR, engl. Automatic Speech Recognition) predstavlja jednu od ključnih tehnologija u domenu interakcije između čoveka i mašine. Osnovna funkcija ASR sistema jeste konverzija govornog signala u tekstualni zapis, čime se omogućava efikasnija komunikacija sa softverskim i hardverskim sistemima. Zahvaljujući brzom razvoju u oblasti veštačke inteligencije i obrade prirodnog jezika, ovi sistemi nalaze široku primenu u različitim kontekstima — od digitalnih asistenata i korisničke podrške, do automatske transkripcije sastanaka, medicinske dokumentacije i medijskog sadržaja [1].

Poseban značaj ASR sistema ogleda se u sve većim zakonskim i društvenim zahtevima za inkluzivnošću, kao što je obavezno titlovanje televizijskog programa u cilju pristupačnosti za osobe sa oštećenim sluhom [2]. Takođe, organizacije sve češće koriste transkripciju govora radi arhiviranja i pretrage sadržaja, što je znatno efikasnije kada su podaci dostupni u tekstualnom obliku.

Pored toga, razvoj interneta stvari (IoT) dodatno naglašava potrebu za prirodnim načinima interakcije sa uređajima, pri čemu se govor nameće kao najintuitivniji modalitet [3]. U tom kontekstu, ASR tehnologija igra ključnu ulogu u omogućavanju ovakve komunikacije.

Tradicionalni ASR sistemi sastoje se od 3 komponente: akustičkog modela, leksikona i gramatike/jezičkog modela. Akustički model uspostavlja vezu između akustičkih obeležja i fonema ili neke druge jedinice za prepoznavanje, npr. trifona. Leksikoni omogućavaju da se jedinice za prepoznavanje povezuju u reči, dok gramatike, odnosno modeli jezika omogućavaju povezivanje reči u veće celine, tj. cele rečenice. Ovakva struktura ASR sistema činila ih je jezički zavisnim i bilo je neophodno razvijanje posebnih sistema za svaki jezik.

Kao i u mnogim drugim oblastima, rast popularnosti dubokih neuronskih mreža doveo je i do promena u pristupu prepoznavanju govora. Akustički modeli bazirani pre svega na skrivenim Markovljevim modelima [4], kao i jezički modeli bazirani na N-gramima [5], bivaju zamenjeni modelima na bazi neuronskih mreža [6, 7]. Međutim, možda najveću promenu u ASR sistemima u poslednje vreme predstavlja uvođenje tzv. end-to-end modela (E2E) [8]. U poređenju sa standardnim modelima E2E modeli integrišu sve različite komponente ASR sistema u jednu celinu i omogućavaju zajedničku obuku i predikciju [8]. Među E2E modelima u poslednje vreme veliku popularnost stekao je model pod nazivom Whisper [9], zahvaljujući činjenici da podržava veliki broj različitih jezika, kao i zahvaljujući dostupnosti za neograničeno korišćenje i doobuke. Iako Whisper podržava veliki broj jezika, performanse modela se značajno razlikuju za različite jezike, što je uglavnom u korelaciji sa količinom podataka dostupnih za obuku za određeni jezik. Takođe, performanse modela su obično merene na bazama na kojima je model i obučavan, iako ne nužno da istim rečenicama, jezički sadržaj je često ograničen temom, a govor sniman u ujednačenim akustičkim uslovima. Zbog toga se peformanse sistema u praktičnim primenama često razlikuju od onih koje su navedene u literaturi. U ovom radu biće prikazani rezultati adaptacije Whisper modela na srpski jezik, kao i uticaj primene odgovarajućih postupaka augmentacije na rezultate prepoznavanja.

Ostatak rada je organizovan na sledeći način. U Poglavlju II opisan je korišćeni ASR model. U poglavlju III dat je pregled korišćenih augmentacija, dok je u poglavlju IV dat pregled korišćenih baza i predstavljeni su dobijeni rezultati. Potom slede zaključak i pravci daljeg istraživanja.

II. WHISPER MODEL

Whisper model predstavlja E2E ASR sistem obučen na 680.000 sati višejezičnog audio materijala, što omogućava prepoznavanje, odnosno, prepoznavanje govora na 97 jezika. Očekivano, najveća količina materijala za obuku je na engleskom jeziku (preko 400.000 sati). Za srpski jezik korišćeno je svega 28 sati javno dostupnog audio materijala, uz dodatnih 91 sat na hrvatskom jeziku. Posledica neravnomerne distribucije podataka za obuku ogleda se u činjenici da model ne ostvaruje jednake performanse za sve jezike.

Pored prepoznavanja, odnosno, transkripcije govora ovaj model podržava i prevođenje, pri čemu je podržana samo opcija prevođenja na engleski. Odnosno, model za govor na srpskom jeziku može direktno da generiše rezultujući tekst na engleskom. Obrnuti pravac u ovom trenutku nije moguć, tj. nije moguće za govor na engleskom jeziku dobiti transkripciju na srpskom. Iako model nudi mogućnost generisanja vremenskih odrednica za segmente, ova opcija nije dovoljno pouzdana pa se u literaturi javilo nekoliko nadogradnji Whisper modela koje nude i automatsko poravnanje [10].

Obuka se sprovodi na parovima audio-tekst, pri čemu trajanje pojedinačnih audio snimaka ne prelazi 30 sekundi, uz učestanost odabiranja od 16 kHz. Kraći snimci se dopunjavaju tišinom. Tokom predikcije, model takođe obrađuje isključivo segmente trajanja do 30 sekundi, što može predstavljati izazov pri obradi dužih audio-zapisa i dovesti do grešaka na prelazima segmenata.

Whisper model zasnovan je na transformer arhitekturi u obliku koder-dekoder strukture. Ulaz kodera predstavljaju logmel-spektrogrami dobijeni iz audio-snimka. Izlaz kodera prosleđuje se dekoderu, koji generiše niz tokena. Ovi tokeni, koji mogu predstavljati foneme ili cele reči, preuzeti su iz ChatGPT-a [11]. Model stoga ne sadrži specifične jezičke modele za sve podržane jezike, što sa sobom nosi i prednosti i nedostatke. U slučaju srpskog jezika, jedan token može predstavljati jedan ili više fonema (uglavnom do tri), što može rezultirati generisanjem reči koje nisu deo srpskog jezika.

Whisper model je dostupan u više varijanata, koje se razlikuju po veličini arhitekture i broju parametara. Postoje modeli od najmanjeg („tiny“) sa 39 miliona parametara, do najvećeg („large“) sa 1550 miliona parametara. Veći modeli su robustniji i obučeni na većoj količini podataka, ali zahtevaju značajno snažniji hardver ne samo za obuku već i za primenu (engl. *inference)*. U ovom radu korišćen je srednji („medium“) model sa 769 miliona parametara. Za rad sa ovim modelom dovoljna je grafička kartica sa 10 GB video memorije.

III. METODE AUGMENTACIJE

Sistemi za automatsko prepoznavanje govora u slučaju jezika sa ograničenim resursima, poput srpskog jezika, suočavaju se sa nedostatkom podataka za obuku (audio snimaka praćenih odgovarajućim transkriptima), raznovrsnošću akustičkog okruženja i kanala, kao i varijabilnošću govora. Kako bi se unapredili postojeći modeli za ovakve jezike i razvili novi, neophodno je snimiti nove, pažljivo osmišljene baze podataka, što predstavlja proces zahtevan u pogledu potrebnih resursa. Povećanje raznovrsnosti postojećih skupova podataka metodama augmentacije može značajno uticati na performanse dodatno obučenog ASR modela, jer i relativno mali skupovi postaju reprezentativniji za realne uslove.

U tom kontekstu, kompresija audio-zapisa predstavlja efikasan način za unošenje dodatne varijabilnosti u postojeće podatke. Ideja je da se audio-zapis dekoduje i ponovo koduje u različitim formatima i sa različitim kvalitetima kompresije, čime se simuliraju realni uslovi u kojima korisnici koriste različite uređaje, aplikacije i mrežne protokole koji često komprimuju govor radi uštede protoka i prostora [12]. U slučaju finog podešavanja modela za prepoznavanje govora obučenog na čistim audio-podacima postoji velika verovatnoća da neće davati dobre rezultate na stvarnim korisničkim snimcima koji su prošli kroz različite oblike kompresije. Uvođenjem audio-fajlova koji su prethodno komprimovani u različitim formatima i pri različitim bitskim brzinama, model se izlaže raznovrsnijim akustičkim uslovima, čime se povećava njegova robustnost, odnosno, otpornost na degradaciju zvuka. Proces augmentacije obično podrazumeva konverziju audio-fajlova iz nekomprimovanog WAV formata u formate poput MP3 i AAC. Jedan od danas najpopularnijih alata za ovu vrstu manipulacije je FFMPEG [13], koji omogućava jednostavno kodovanje i dekodovanje audio-fajlova. Iako je moguće proizvesti više verzija istog zvučnog zapisa sa različitim bitskim brzinama, u ovom radu nasumično je birana po jedna vrednost bitske brzine iz unapred definisanog skupa za svaki audio-fajl koji se augmentujeme, kako ne bi bilo ponavljanja u skupu podataka za obuku. Korišćen je MP3 format sa sledećim bitskim brzinama: 20 kbps, 24 kbps, 32 kbps, 36kbps, 38 kbps i 64 kbps.

Pored kompresije, još jedan važan aspekt akustičke varijacije koji treba uzeti u obzir jeste reverberacija. Dok kompresija simulira degradaciju signala kroz digitalne kanale, reverberacija modeluje uticaj fizičkog prostora na kvalitet govora. Pojava vremenskog i spektralnog zamućenja snimljenog audio-signala usled višestrukih refleksija od različitih površina u prostoriji u kojoj se vrši snimanje naziva se reverberacija [14]. U mnogim realnim okruženjima poput kancelarija, učionica i holova, reverberantni govor je česta pojava. ASR sistemi obučeni isključivo na čistom govoru ne postižu visoke performanse u takvim uslovima pa se pribegava odgovarajućoj augmentaciji skupa podataka za obuku [15]. Augmentacija pomaže bolju generalizaciju u različitim akustičkim okruženjima i poboljšava otpornost u slučaju udaljenog izvora zvuka i u slučaju prisustva šuma. Kako je prikupljanje podataka koji sadrže reverberaciju kompleksan i vremenski zahtevan proces, korišćenje simuliranih impulsnih odziva prostorija predstavlja čest pristup u literaturi [15-16]. Ovi odzivi se konvoluiraju sa čistim audio-signalom da bi se simuliralo ponašanje zvuka u fizičkom prostoru. Za potrebe simulacije koristili smo skup predefinisanih vrednosti parametara koji se odnose na slabljenje i veličinu odgovarajuće prostorije (‘*damping factor*’ i ‘*room size*’).

Još jedna standardna metoda augmentacije podataka u cilju poboljšanja ASR sistema jeste dodavanje šuma [17], odnosno simulacija situacije u kojoj je govor snimljen u prisustvu različite pozadinske buke. Baza uzoraka šumova, odnosno, pozadinske buke dobijena je od preduzeća AlfaNum [18], a čini je oko 70.000 različitih šumova trajanja u proseku 6s, prikupljanih na različite načine (samostalno snimanje i preuzimanje sa interneta). U bazi se nalaze različite vrste šumova poput vetra, saobraćajne buke, buke sa gradilišta, do žamora u kafićima, pozadinskog zvuka sa televizije, i dr. Šumovi su dodati na pojedinačne audio-snimke nasumičnim

odabirom sa konstantnim SNR od 15dB tokom celog trajanja audio-snimka.

U ovom radu primenjene su tehnike augmentacije koje se odnose na kompresiju signala, reverberaciju i dodavanje šuma, i to simultano nad svim podacima iz baze, tako što su vrednosti relevantnih parametara za svaku od primenjenih tehnika nasumično odabrane iz unapred definisanih skupova.

IV. EKSPERIMENTI

A. Baza za obuku

U procesu doobuke modela korišćena je baza od 1500 sati transkribovanog audio-materijala na srpskom i hrvatskom jeziku, nastala manuelnom transkripcijom audio-materijala koji potiče iz audio-knjiga i radio-televizijskih emisija, koju je za potrebe ovog istraživanja takođe obezbedilo preduzeće AlfaNum. Prosečna dužina rečenice je 5 sekundi, a svi fajlovi imaju frekvenciju odabiranja 16 kHz.

Korišćenjem postupaka augmentacije opisanih u prethodnim odeljcima kreirano je još dodatnih 3000 sati audio materijala. Od ove količine materijala 1500 sati je dobijeno dodavanjem šuma, a preostalih 1500 kombinacijom kompresije i reveberacije.

B. Test podaci

Za potrebe testiranja korišćene su dve javno dostupne baze na srpskom jeziku – *Common Voice* [19] i *Fleurs* [20]. Pored ovih baza korišćen je i skup podataka koji potiče iz realnih primera iz prakse, a koji uključuju sastanke, emisije i telefonske razgovore. Konkretno, 25 snimaka, po 5 iz grupe sastanci i emisije, trajanja od 3-30minuta, i 15 telefonskih razgovora trajanja po 10-ak sekundi.

C. Rezultati

Za testiranje performansi modela korišćena je greška prepoznavanja na nivou reči (WER, engl. *Word Error Rate*). Ova mera računa se na osnovu sledeće formule:

$$WER = \frac{S+I+D}{W}. \quad (1)$$

U jednačini (1) W je ukupan broj reči koji je izgovoren u test rečenici. Sa S je označen broj reči koje su pogrešno prepoznate (npr. izgovoreno „mama“, a prepoznato „tama“), sa I broj dodatih reči (npr. izgovoreno „ja sam voleo“, a prepoznato „ja sam je voleo“), dok je sa D označen broj reči koje postoje u test-rečenici, ali ne i u prepoznatoj (npr. izgovoreno „ja sam je voleo“, a prepoznato „ja sam voleo“).

Vrednosti za WER dobijene na javnim bazama prikazane su u tabeli 1, dok su za test-rečenice koje potiču iz primera iz prakse prikazane u tabeli 2. Iz tabele 1 jasno se vidi u kojoj meri doobuka modela za specifičan jezik utiče na krajnje performanse. Za *Common Voice* bazu WER je sa 85.6% kod polaznog modela pao na svega 8.06% kod doobučenog, dok je kod *Fleurs* baze taj uticaj bio manji, ali i dalje veoma značajan – WER je kod doobučenog modela pao sa polaznih 44.9% na 10.61%. Kod obe baze uključivanje augmentovanog materijala je dovelo do dodatnog, ali ne tako drastičnog smanjenja WER. U tumačenju dodatnog poboljšanja vrednosti WER pod uticajem augmentovanih podataka u obzir treba uzeti i samu prirodu test-baza. I *Common Voice* i *Fleurs* sastoje se od veoma kratkih rečenica koje se često sastoje od 3 ili 4 reči. Kod takvih rečenica greška i u jednom slovu podiže vrednosti WER na 25% ili 33% (u konkretnim primerima, a uzimajući jednačinu 1 u obzir, W ima vrednost 3 ili 4, S je jednako 1, a preostale vrednosti su 0). Tako je u slučaju *Common Voice* baze ukupan broj sasvim tačno prepoznatih rečenica dobijenih korišćenjem modela doobučenog nad originalnom bazom 1282, dok je taj broj u slučaju modela dobijenog doobukom nad augmentovanom bazom 1305. Drugim rečima, ukupan broj sasvim tačno prepoznatih rečenica porastao je za 1.7%. Slično je kod *Fleurs* baze, gde je ukupan broj tačno prepoznatih rečenica zahvaljujući augmentaciji podataka za obuku porastao sa 226 na 240, tj. za oko 5.8%.

Rezultati prikazani za test-rečenice iz realnih situacija (tabela 2) ipak pokazuju manju konzistentnost u odnosu na rezultate dobijene nad javnim bazama. Naime, samo u slučaju telefonskih razgovora augmentacija podataka dovela je do poboljšanja rezultata, dok se u slučaju snimaka sastanaka i televizijskih emisija dobijaju nešto lošiji rezultati. Ovo se može objasniti činjenicom da degradacija signala korišćenjem MP3 kodovanja niskog bitskog protoka odgovara degradacijama koje nastaju prenosom signala telefonskim kanalom. Problemi koji su uočeni kod snimaka sastanaka i televizijskih emisija nisu eksplicitno vezani za sam kvalitet signala, nego su potpuno druge prirode – česta preklapanja govornika, nepotpune rečenice, reči ili čak i veće rečenične celine koje su slabo artikulisane, a čija simulacija nije obuhvaćena predloženim augmentacijama.

Tabela 1. WER za različite modele dobijen na standardnim javno dostupnim bazama

	Osnovni model	**Doobuka**	**Doobuka+ augmentacija**
CommonVoice	85.6	8.07	7.24
Fleurs	44.9	10.61	10.12

Tabela 2 WER za različite modele dobijen na internim test-podacima

	Doobuka	**Doobuka+ augmentacija**
Sastanci	23.53	24.89
Emisije	13.17	14.13
Tel. razgovori	7.83	5.8

V. ZAKLJUČAK

U radu je ispitivan uticaj augmentacije podataka na performanse doobučenog *Whisper* modela. Pokazano je da u slučaju javno dostupnih baza doobuke sa korišćenim augmentovanim materijalom utiču na poboljšanje performansi modela. Sa druge strane, doprinos predloženih tehnika augmentacije u slučaju snimaka dobijenih u realnim uslovima, kao što je transkripcija sastanaka ili TV emisija, nekonzistentan je i ograničen jer se kod ovakvih snimaka javljaju problemi koji nisu isključivo vezani za degradaciju kvaliteta signala. Na osnovu toga može se zaključiti da je potrebno ispitati i druge specifične tehnike augmentacije, koje bi, primera radi, uključivale preklapanje originalnog signala sa pozadinskim govorom ili simulaciju promene rastojanja od mikrofona.

ZAHVALNICA

Istraživanje sprovedeno uz podršku Fonda za nauku Republike Srbije, za projekat br. 7449 „Multimodalna višejezička komunikacija između čoveka i mašine, AISPEAK“.

LITERATURA

[1] V. Delić, D. Pekar, M. Sečujski, B. Popović, E. Pakoci and S. Suzić, „Development of speech technology for Serbian and its applications“, 1st Serbian International Conference on Applied Artificial Intelligence (SICAAI), Kragujevac, Srbija, Maj 2022

[2] A. Coy, P.S. Mohammed and P. Skerrit, “Inclusive Deaf Education Enabled by Artificial Intelligence: The Path to a Solution”, International Journal of Artificial Intelligence in Education, str.1-39, 2024

[3] H. Younis, and J.H. Hansen, “Challenges in real-time-embedded IoT command recognition”, 7th World Forum on Internet of Things (WF-IoT), pp. 848-851, jun 2021.

[4] S.E. Levinson, L.R. Rabiner and M.M. Sondhi, “An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition”, BellSystem Technical Journal 62, str. 1035–1074, 1983.

[5] F. Jelinek, “Statistical methods for speech recognition”, MIT press, 1998.

[6] G.E. Dahl, D. Yu, L. Deng, L. and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition”, IEEE Transactions on audio, speech, and language processing, 20(1), str. 30-42, 2011.

[7] T. Mikolov, M.Karafiát, L. Burget, J. Cernocký, J. and S. Khudanpur, “Recurrent neural network based language model”, 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, str. 1045-1048, Japan, Septembar 2010.

[8] R. Prabhavalkar, T. Hori, T.N. Sainath, R. Schlüter and S. Watanabe, “End-to-end speech recognition: A survey”. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32, pp.325-351, 2023

[9] A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey and I. Sutskever, “Robust speech recognition via large-scale weak supervision”, International conference on machine learning, str. 28492-28518, Jul 2023.

[10] L. Wagner, B. Thallinger and M. Zusag, “CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions, arXiv preprint arXiv:2408.16589, 2024

[11] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.L. Han, Q.L. and Y. Tang, “A brief overview of ChatGPT: The history, status quo and potential future developmen”, IEEE/CAA Journal of Automatica Sinica, 10(5), pp.1122-1136, 2023

[12] M. P. Fernández-Gallego, and D. T. Toledano, “A Study of Data Augmentation for ASR Robustness in Low Bit Rate Contact Center Recordings Including Packet Losses”, Applied Sciences, 12(3), 1580, 2022.

[13] S. Tomar, “Converting video formats with ffmpeg”, Linux journal, 2006(146):10, 2006.

[14] H. Malik, and H. Farid, "Audio forensics from acoustic reverberation," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2010), Dallas, TX, USA, 2010, pp. 1710-1713

[15] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017), New Orleans, LA, USA, 2017, pp. 5220-5224.

[16] J. Malek,, and J. Zdansky, “On Practical Aspects of Multi-condition Training Based on Augmentation for Reverberation-/Noise-Robust Speech Recognition”. In: Ekštein, K. (eds) Text, Speech, and Dialogue. TSD 2019. Lecture Notes in Computer Science(), vol 11697. 2019, Springer, Cham.

[17] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Ng, “Deep Speech: Scaling up end-to-end speech recognition,” in arXiv, 2014.

[18] www.alfanum.co.rs

[19] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers and G. Weber, “Common voice: A massively-multilingual speech corpus”. arXiv preprint arXiv:1912.06670, 2020

[20] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, “Fleurs: Few-shot learning evaluation of universal representations of speech”, IEEE Spoken Language Technology Workshop (SLT), str.. 798-805, 2023

The Impact of Training Data Augmentations on the Performance of a Fine-Tuned Whisper Model

Siniša Suzić, Tijana Nosek, Nikola Simić, Darko Pekar, Vlado Delić

ABSTRACT

Whisper is one of the most popular speech recognition models in recent times. Its popularity is based on the openness of the model itself and its support for recognition in a large number of languages. However, since the amount of data used for training varied significantly across languages, the recognition performance also differs greatly. This paper analyses the positive impact of fine-tuning the medium variant of the Whisper model on word error rate values for the Serbian language, which is an example of an under-represented language in the initial model's training data. The influence of adding new training data through augmentation by compression, noise addition, and reverberation was also examined. It was shown that adding augmented data in the case of publicly available datasets leads to WER improvement. On the other hand, when testing the fine-tuned model on real-world examples, the impact of data augmentation on model performance was not consistent, which highlights the need for additional testing using other augmentation techniques as well.