Validacija zvočnih posnetkov pri izdelavi podatkovne zbirke za učenje razpoznavalnika slovenščine

Janez Križaj, Simon Dobrišek

Fakulteta za elektrotehniko, Univerza v Ljubljani
E-pošta: janez.krizaj@fe.uni-lj.si

Audio validation of candidate training data for automatic speech recognition of Slovene

This paper presents a support tool for the validation of audio recordings of Slovene speech. The tool allows the user to verify compliance with the quality requirements regarding the correct audio format, adequate speech volume, compatibility of the read speech with the corresponding text, suitability of initial/final pauses and then it removes recordings that do not adhere to the named requirements. By doing so, it is possible to quickly and efficiently verify audio recordings that can eventually be used as training data for learning the automatic speech recognition models. We demonstrate the tool's efficiency by building the Slovene speech corpus from the recordings collected under the RSDO project (Development of Slovene in a Digital Environment).

1 Uvod

Področje samodejnega razpoznavanja govora je v zadnjih letih doživelo velik napredek predvsem zahvaljujoč uporabi globokih nevronskih mrež [1, 2], ki pa za svoje učenje potrebujejo veliko učnega materiala. Žal pa trenutno za slovenski jezik ni na voljo ustreznih prosto dostopnih govornih baz, ki bi jih lahko uporabili za izgradnjo razpoznavalnika slovenskega jezika in na njem temelječih komercialnih produktov. Obstoječe baze so večinoma bodisi nedostopne bodisi plačljive bodisi na voljo samo za nekomercialno rabo. Poleg tega po obsegu ne zadoščajo za izgradnjo razpoznavalnika po sodobnih standardih. V okviru projekta RSDO smo si zato kot enega izmed ciljev zadali zgraditi govorno bazo, ki bo prosto dostopna in jo bo moč uporabljati tako za nekomercialne kot tudi komercialne namene.

Za uspešno učenje samodejnih razpoznavalnikov govora morajo učni podatki praviloma izpolnjevati določene zahteve, ki se nanašajo predvsem na ustrezno kakovost zvočnega posnetka in skladnost posnetega govora z referenčnim besedilom. Aplikacija, predstavljena v tem članku, omogoča preverjanje izpolnjevanja omenjenih zahtev in zavrnitev posnetkov, ki teh zahtev ne izpolnjujejo in bi posledično lahko negativno vplivali na strojno učenje razpoznavalnika. Poglavitni namen udejanjitve aplikacije je olajšati preverjanje ustreznosti zvočnega gradiva pri izgradnji govorne baze v projektu RSDO. Govorna baza, bo omogočala razvoj boljših razpoznavalnikov, kot je to mogoče s trenutno razpoložljivimi viri in orodji. Trenutne govorne baze slovenskega jezika omogočajo izgradnjo razpoznavalnika z WER (angl. Word Error Rate) med približno 25 % in 30 %, omejeno na akustične situacije, ki so zajete v teh bazah. Z novo govorno bazo bo po naši oceni možno doseči bistveno boljši WER.

Delo, ki je predstavljeno v članku vsebuje naslednje doprinose:

1. Izdelava aplikacije za validacijo govornih posnetkov, ki olajša izločanje posnetkov neustrezne kakovosti.

2. Uporaba izdelane aplikacije za izgradnjo govorne baze, ki bo osnova za izdelavo razpoznavalnika slovenskega jezika.

V nadaljevanju sledi podrobna predstavitev razvite aplikacije in preliminarnih rezultatov, ki smo jih pridobili na posnetkih, zbranih v okviru projekta RSDO.

2 Aplikacija za validacijo posnetkov

To poglavje vsebuje predstavitev gradnikov in opis delovanja razvitega orodja za validacijo govornih posnetkov.

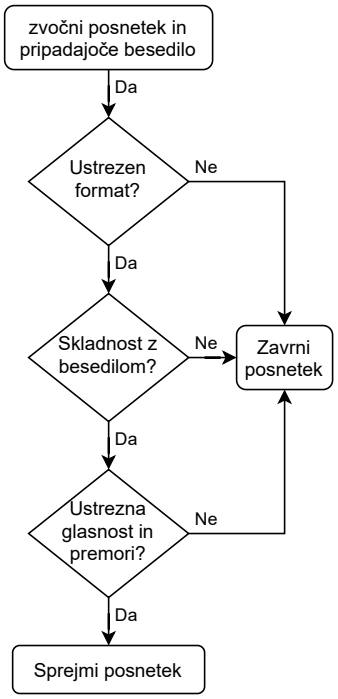
2.1 Postopek zbiranja posnetkov

V grobem obstajata dva pristopa k izdelavi govornih zbirk. Prvi pristop obsega snemanje posameznih besedilnih povedi, ki jih berejo govorci, pri čemer posamezen posnetek vsebuje le eno poved. Pri drugem pristopu pa dolgo zvočno datoteko in pripadajoče besedilo naknadno razdelimo na posamezne povedi [3]. Vsled enostavnejše zagotovitve poravnave zvoka in besedila smo se poslužili prvega pristopa.

Vsak govorec dobi seznam povedi, ki so bile predhodno zbrane s postopkom spletnega luščenja podatkov iz spletnih novičarskih portalov. Med snemanjem povedi se mora govorec držati predpisanih zahtev: *i*) vsaka posamezna poved naj bi bila shranjena v samostojni zvočni datoteki s končnico WAV, *ii*) na začetku in na koncu povedi mora biti vsaj pol sekunde in ne več kot ena sekunda originalno posnetega premora oziroma tišine, *iii*) posnetek se mora izvirno zajemati in shraniti v enokanalnem (mono) formatu s frekvenco vzorčenja 44,1 kHz.

2.2 Opis validacijskega postopka

Razvita aplikacija omogoča preverjanje zgoraj navedenih zahtev, ki naj bi se jih morali držati snemalci in govorci pri zbiranju posnetkov. Shematski prikaz validacije posnetkov prikazuje slika 1 iz katere je razvidno, da se test skladnosti posnetka s predpisanimi zahtevami vrši preko treh odločitvenih blokov, ki obsegajo preverjanje ustreznosti formata zapisa posnetka, preverjanje skladnost govora s pripadajočim besedilom in preverjanje ustreznost premorov in glasnosti posnetka. Če se izkaže, da dani posnetek ne izpolnjuje katerega izmed pogojev ga zavedemo v seznam zavrnjeni posnetkov.



Slika 1: Shematski prikaz procesa validacije govornega posnetka.

2.3 Grafični vmesnik

Interakcija uporabnika z aplikacijo je mogoča preko grafičnega vmesnika, ki ga prikazuje slika 2. Okno vmesnika je razdeljeno na več okvirjev. V levem zgornjem kotu je okvir za vnos vhodnih parametrov, levo spodaj se nahaja okvir za preverjanje skladnosti posnetka z referenčnim besedilom, sredinski del zaseda modul za preverjanje ustreznosti začetnega/končnega premora in glasnosti, v desnem delu okna pa se nahajata seznama sprejetih in zavrnjenih posnetkov in okvir s statističnimi podatki zvočnega posnetka.

Za zagon validacijskega procesa je potrebno najprej vnesti vhodne parametre, ki vključujejo *i*) pot do mape s posnetki WAV, *ii*) pot do datoteke XLSX s pripadajočim besedilom, *iii*) številko posnetka pri kateri želimo pričeti (ali nadaljevati) z validacijo in *iv*) način delovanja, kjer je na voljo

1. *samodejni* način, kjer se preverjanje vseh pogojev pri testu ustreznosti posnetka izvede samodejno in program od uporabnika, po vnosu vhodnih parametrov, ne zahteva več nobene interakcije.

2. *polsamodejni* način, ki zahteva uporabnikovo posredovanje le, če kateri izmed pogojev ni izpolnjen. V tem primeru ima uporabnik preko ustreznih gumbov možnost ročno odgovoriti ali posnetek res ne izpolnjuje dotičnega pogoja.

3. *ročni* način, kjer mora uporabnik na vsakega od pogojev odgovoriti s pritiskom na ustrezno tipko v grafičnem vmesniku.

Če izberemo samodejni ali polsamodejni način, je potrebno vnesti še prag WER nad katerim se smatra, da se posnetek ne sklada z besedilom. Po vnosu vhodnih parametrov zaženemo validacijo s pritiskom na tipko "Zaženi". S tem se sproži proces validacije v katerem je vsak obravnavani posnetek podvržen preverjanju treh omenjenih pogojev in je zaveden med zavrnjene, če ne izpolnjuje katerega od teh pogojev. Hitrost validacije je odvisna od izbranega načina delovanja, saj ročna interakcija precej upočasni preverjanje, vendar pa hkrati izboljša uspešnost validacije, ker je samodejno preverjanje pogojev podvrženo napakam, ki izhajajo predvsem iz pogreškov samodejnega razpoznavalnika govora in napak v oceni dolžine premorov.

2.4 Preverjanje skladnosti s pripadajočim besedilom

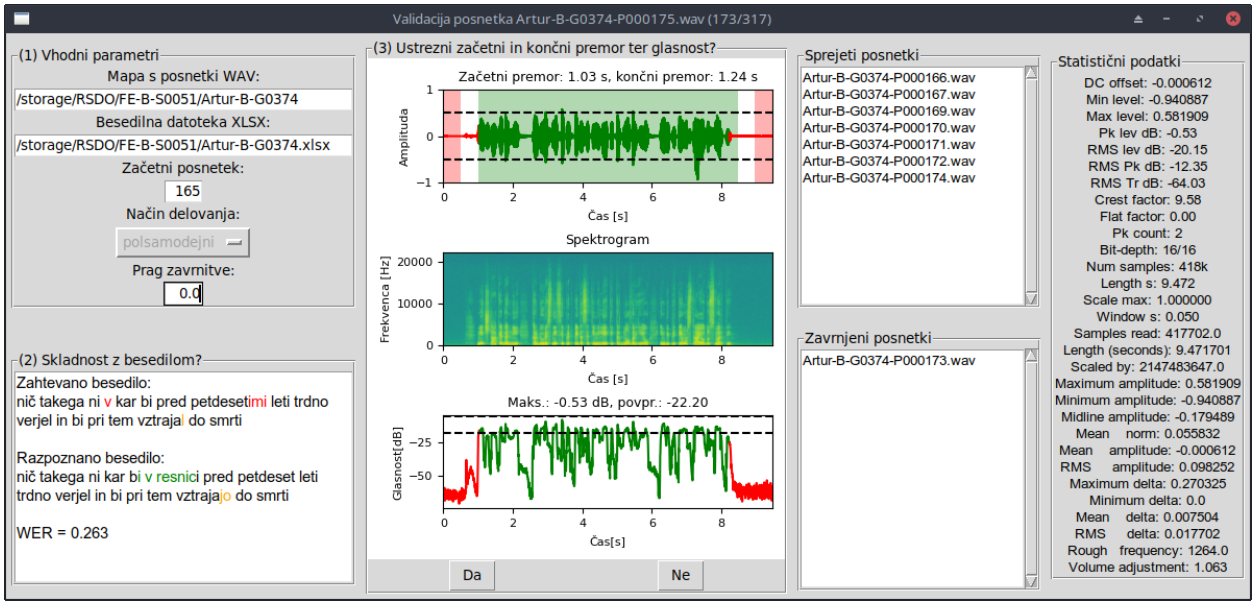
Preverjanje skladnosti posnetka s pripadajočim besedilom se vrši na dva načina. V ročnem načinu delovanja se v grafičnem vmesniku začne predvajati posnetek in izpiše pripadajoče besedilo, uporabnik pa s poslušanjem oceni ujemanje in sprejme/zavrne posnetek s pritiskom na ustrezno tipko. Pri (pol)samodejnem načinu pa se skladnost z besedilom oceni (pol)samodejno s pomočjo razpoznavalnika slovenskega govora, pri čemer smo se poslužili Googlovega razpoznavalnika. Razpoznavalnik temelji na uporabi rekurentnih nevronskih mrež [4] in deluje v obliki oblačne storitve, omogoča pa tudi uporabo privzetih poverilnic, zato je mogoča uporaba tudi brez prijave v oblačno storitev.

Besedilni niz, ki ga vrne razpoznavalnik se nato primerja s pripadajočim besedilom na podlagi metrike WER (ang. Word Error Rate), ki se najpogosteje uporablja za ocenjevanje kakovosti samodejnih razpoznavalnikov [5] in je definirana kot

$$WER = \frac{S + D + I}{N}, \tag{1}$$

kjer je S število zamenjav, D število izbrisov, I število vstavkov in N število vseh besed. Ker so v referenčnem besedilu vse številke zapisane z besedo, smo pred izračunom WER po (1) v izhodnem nizu Googlovega razpoznavalnika morebitne številke pretvorili v zapis z besedo. Razpoznano besedilo prav tako ne vsebuje ločil, zato smo jih pred izračunom WER izločili tudi iz referenčnega besedila.

V (pol)samodejnem načinu se skladnost z referenčnim besedilom potrdi/zavrne, če je vrednost WER



Slika 2: Grafični vmesnik aplikacije za validacijo govornih posnetkov.

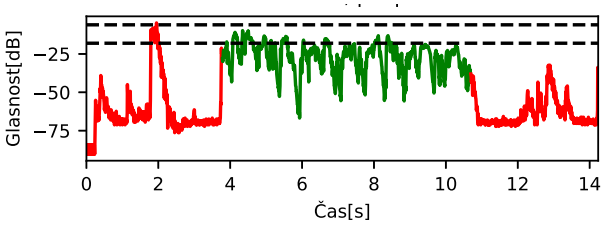
manjša/večja od vnaprej nastavljenega praga. V ročnem načinu in v primeru zavrnitve pri polsamodejnem načinu se skladnost posnetka z besedilom preveri s poslušanjem posnetka. Tu so v pomoč barvno kodirane razlike med referenčnim in razpoznanim besedilom, pri čemer so zamenjave označene z oranžno, vstavki z zeleno in izpuščeni deli povedi z rdečo barvo (glej spodnji levi kot slike 2).

2.5 Preverjanje premorov in glasnosti

Ustreznost začetnega in končnega premora ter glasnosti se prav tako izvaja bodisi samodejno bodisi ročno, glede na izbiro načina validacije. Pri samodejnem izračunu premorov izrabljamo zahtevo, da morata začetni in končni del posnetka vsebovati od 0,5 s do 1,0 s premora. Iz začetnega in končnega dela posnetka tako lahko izračunamo povprečno vrednost glasnosti, ki pripada tihemu delu, iz preostanka posnetka pa povprečno glasnost govornega dela posnetka. Prag med negovornimi in govornimi odseki posnetka izračunamo kot povprečje teh dveh glasnosti. Izločitev morebitnih krajših odsekov v negovornem delu, kjer šum preseže vrednost praga dosežemo z morfološkimi operacijami erozije. Primer uspešne določitve začetnega in končnega premora v prisotnosti šuma je povzema graf glasnosti na sliki 3, kjer glasnost šuma v negovornem delu, mestoma celo preseže povprečno glasnost govornega odseka.

Glasnost govora v posnetku se izračuna kot povprečna glasnost govornega dela signala. Če je povprečna glasnost manjša od vnaprej nastavljenega praga, se posnetek pri samodejnem načinu delovanja zavrne.

Pri ročnem načinu validacije se v grafičnem vmesniku izrišejo trije grafi (sredinski del slike 2), ki so v pomoč pri odločanju o ustreznosti premorov in glasnosti posnetka. Zgornji graf prikazuje amplitudo zvočnega posnetka, srednji grafu zavzema spektrogram, pod njim pa je graf glasnosti zvočnega posnetka. Grafa amplitude in glasnosti imata barvno označene negovorne (rdeča) in govorne odseke (zelena) in na ta način omogočata enostaven pregled ustreznosti premorov. Črtkani horizontalni črti na grafu glasnosti pa omejujeta območje med -18 dbFS in -6 dBFS, znotraj katerega naj bi bil signal pri običajni glasnosti govora. Iz spektrograma pa med drugim lahko opazimo ali so bili premori v posnetek dodani naknadno, kar je tudi razlog za zavrnitev posnetka.



Slika 3: Samodejna določitev začetnih in končnih premorov na pošumljenem posnetku govora. Zeleno obarvani del predstavlja govor, rdeči del pa pokriva začetni in končni premor.

2.6 Implementacijski detajli

Aplikacija je udejanjena v programskem jeziku Python 3. Posamezne naloge v procesu validacije posnetkov smo izvedli z uporabo že obstoječih knjižnic, ki so navedene v Tabeli 1. Programska koda je prosto dostopna na https://github.com/jan3zk/audio_validation. Poleg Python skripte je za Windows okolje na voljo tudi samostojna izvršna datoteka .exe, ki smo jo tvorili s pomočjo knjižnice PyInstaller in omogoča enostaven zagon aplikacije brez predhodne namestitve Python knjižnic.

3 Eksperimenti

Delovanje aplikacije za validacijo posnetkov je preizkušeno na posnetkih, zbranih v okviru projekta RSDO, z namenom izgradnje govorne baze za učenje razpoznavalnika slovenskega jezika. Rezultati, ki so bili prido-

Tabela 1: Uporabljene programske knjižnice

Opravilo	**Knjižnica**
grafični vmesnik	tkinter
preverjanje formata zapisa posnetka	soundfile
skladnost z besedilom	Google STT API, SpeechRecognition, difflib, jiwer, num2words
test premorov in glasnosti	scipy, pydub, matplotlib, sox
pretvorba iz .py v .exe	pyinstaller

Tabela 2: Deleži najdenih neustreznih posnetkov.

	Vzrok zavrnitve		
Način delovanja	**a**[*]	**b**[†]	**c**[§]
ročni	0.001	0.091	0.078
polsamodejni	0.001	0.098[‡]	0.079
samodejni	0.001	0.624[‡]	0.107

[*] neustrezen format

[†] neskladje z referenčnim besedilom

[§] neustrezni premori in glasnost

[‡] delež pri pragu $WER = 0.0$

bljeni na delu do sedaj zbranih posnetkov, so prikazani v tabeli 2.

Delež zavrnjenih posnetkov zaradi neustreznega formata ni odvisen od načina delovanja, ker se ustreznost format zapisa pri vseh načinih delovanja preveri samodejno.

Delež zavrnitev zaradi neskladanja z besedilom je pri ročnem načinu primerljiv z deležem, dobljenem pri polsamodejnem načinu, medtem ko je delež te vrste zavrnitev precej večji v samodejnem načinu delovanja. Razlog za večji delež zavrnitev v samodejnem načinu je, da preverjanje skladnosti posnetka z referenčnim besedilom na podlagi ujemanja referenčnega besedila z razpoznanim besedilom pri $WER = 0.0$ vnese veliko napačnih zavrnitev. Pri polsamodejnem načinu, ki ponudi ročno preverjanje vsake zavrnitve, pa smo napačne zavrnitve razveljavili.

Deleži zavrnitev zaradi neustreznih premorov ali glasnosti so prav tako nekoliko višji v samodejnem načinu validacije, kar je večinoma posledica nezanesljive ocene trajanja premorov zaradi šuma v posnetku.

Rezultati v tabeli 2 nakazujejo, da je zaradi nezadostne zanesljivosti samodejnega razpoznavalnika govora smotrno izbrati polsamodejni način validacije, ki pri nizkih vrednostih WER naredi malo napačnih samodejnih potrditev, vendar pa pohitri validacijo napram ročnemu načinu delovanja.

4 Zaključek

V članku je predstavljena aplikacija za validacijo govornih posnetkov, ki smo jo razvili za lažje preverjanje ustreznosti posnetkov pri izdelavi govorne baze za učenje samodejnega razpoznavalnika slovenščine. Aplikacija vključuje preverjanje posnetka z referenčnim besedilom, kakor tudi preverjanje ustreznosti formata, začetnega in končnega premora ter glasnosti posnetka. Uporabnost aplikacije smo pokazali z validacijo posnetkov, zbranih v okviru projekta RSDO.

Bodoče delo vključuje uporabo različnih razpoznavalnikov pri preverjanju skladnosti z referenčnim besedilom in primerjavo WER razpoznavalnika, naučenega na novi zbirki, z razpoznavalnikom, naučenim na obstoječih prosto dostopnih zbirkah slovenskega govora kot je npr. Mozilla Common Voice [6].

Zahvala

Raziskovalno delo, ki je pripeljalo do predstavljenih rezultatov, je bilo delno financirano s strani programa RSDO (Razvoj slovenščine v digitalnem okolju), financiranega s strani Ministrstva za kulturo in Evropskega sklada za regionalni razvoj.

Literatura

[1] S. Kriman *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," v *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, str. 6124–6128.

[2] Y. Kong *et al.*, "Multi-channel automatic speech recognition using deep complex unet," v *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, str. 104–110.

[3] E. Bakhturina, V. Lavrukhin, in B. Ginsburg, "Nemo toolbox for speech dataset construction," *ArXiv*, zv. abs/2104.04896, 2021.

[4] W. Chan, N. Jaitly, Q. Le, in O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," v *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, str. 4960–4964.

[5] A. Ali, W. Magdy, P. Bell, in S. Renais, "Multi-reference wer for evaluating asr for languages with no orthographic rules," v *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, str. 576–580.

[6] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," v *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, Francija: European Language Resources Association, maj 2020, str. 4218–4222.