Poravnava zvočnih posnetkov s transkripcijami narečnega govora in petja

Matija Marolt, Mark Žakelj, Alenka Kavčič, Matevž Pesek

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
Večna pot 113, 1000 Ljubljana
matija.marolt@fri.uni-lj.si

1 Uvod

V povzetku predstavljamo sistem za poravnavo zvočnih posnetkov slovenskega govora s pripadajočimi transkripcijami na nivoju besed. Pri razvoju sistema nas je še posebej zanimala njegova uporabnost pri poravnavi narečnega govora in petja, saj avtomatska razpoznava govora v tovrstnih posnetkih deluje nezanesljivo, z veliko napakami. Natančna avtomatska poravnava posnetkov in transkripcij nam tako lahko pomaga pri analizi narečnih korpusov in pripravi novih anotiranih podatkov za učenje razpoznavalnikov. V povzetku predstavimo sistem za poravnavo in primerjamo kvaliteto poravnave nenarečnih in narečnih govorcev. Analiziramo tudi kvaliteto poravnave narečnega petja z uporabo sistema, ki je učen zgolj na govoru. Ker se petje lahko zelo razlikuje od govora (dodatna spremljava, večglasno petje, dolgi toni, ...), se v nalogi osredotočimo zgolj na enoglasno petje brez spremljave, ki je še najbolj podobno govoru.

2 Sistem za poravnavo

Sistem za poravnavo posnetkov in transkripcij je sestavljen iz treh glavnih komponent:

- segmentacija posnetka, s čimer razdelimo celoten posnetek na več krajših delov, hkrati pa odstranimo šum in tišino;
- razpoznava govora, s čimer iz avdio signala pridobimo približno tekstovno transkripcijo;
- poravnava, s čimer vsaki besedi v originalnem besedilu določimo mesto v pridobljeni transkripciji in s tem tudi čas pojavitve.

2.1 Segmentacija posnetka

Segmentacija je osnovana na Googlovem WebRTC–VAD algoritmu[1], ki je hiter, robusten in v praksi pogosto uporabljen. S tem algoritmom lahko klasificiramo posamezen časovni okvir kot govor ali ozadje. Algoritem robustne segmentacije je povzet po izvorni kodi, uporabljeni v sistemu DeepSpeech (Hilleman et al., 2018). WebRTC–vad ima nastavljiv parameter *aggresiveness*, ki lahko zasede vrednosti med 0 in 3. Parameter smo nastavili na vrednost 2, tako smo dobili dovolj kratke segmente, da proces dekodiranja pri razpoznavi govora ni trajal predolgo.

2.2 Razpoznava govora

Razpoznava govora je implementirana v dveh delih: 1) uporaba globokega akustičnega modela za pridobitev verjetnosti posameznih znakov za vsak časovni okvir in 2) dekodiranje izhoda modela za pridobitev končne transkripcije.

Podatki za učenje akustičnega modela so bili pridobljeni iz različnih virov: Gos (Zwitter et al., 2013), Gos VideoLectures (Videolectures, 2019), CommonVoice[2] , SiTEDx (Žgank et al., 2016), Sofes (Dobrišek et al., 2017) in narečni govor s portala narecja.si[3].

Akustični model je implementiran z uporabo ogrodja Nvidia NeMo, uporabili smo globoki model QuartzNet_15x5 (Kriman et al., 2019). Uporabili smo ga, ker lahko z njim kljub relativno majhnem številu parametrov (18,9 milijona) še vedno dobimo dokaj dobro natančnost razpoznave, primerljivo z večjimi modeli (več kot 100 milijonov parametrov). Primerjali smo dva modela: QuartzNet_15x5, učen zgolj na slovenskih podatkih, in QuartzNet_15x5, predučen na angleških podatkih, nato pa dodatno učen še na slovenskih podatkih. S slednjim modelom smo preverili kvaliteto prenosa znanja iz tujega jezika v slovenščino.

Za pridobitev transkripcij smo primerjali tri različne metode dekodiranja CTC: 1) požrešna metoda največjih verjetnosti (*greedy*), kjer za vsak časovni korak v CTC izberemo najbolj verjeten znak, nato združimo sosednje

[1]Webrtc google repository.
https://chromium.googlesource.com/external/webrtc/+/branch-heads/43/webrtc/common_audio/vad
[2]Mozzilla Common Voice website. https://commonvoice.mozilla.org/sl/datasets.
[3] https://narecja.si/

ponovitve; 2) iskanje v snopu z besednim jezikovnim modelom (*word*) in iskanje v snopu z znakovnim jezikovnim modelom (*char*).

Za jezikovni model smo uporabili N–gram jezikovni model KenLM (Heafield, 2011). Ker se model uporablja zgolj med dekodiranjem CTC za posamezen primer poravnave, smo za gradnjo modela uporabili kar originalno besedilo posameznega primera. Tako dobimo model, ki ni posplošen za slovenski jezik, temveč je prilagojen posamezni poravnavi. Testi so pokazali, da red jezikovnega modela ne vpliva bistveno na rezultat, na koncu smo uporabili model četrtega reda.

2.3 Poravnava in iterativno združevanje

S pomočjo razpoznavalnika govora iz posnetka pridobimo približno transkripcijo govora. Le-to moramo v zadnjem koraku poravnati z originalnim besedilom posnetka. Za osnovno poravnavo uporabimo algoritem povzet po orodju DeepSpeech. Izkaže se, da z uporabo tega algoritma ne zagotovimo poravnave vseh besed originalnega besedila. Krajše besede pogosto nimajo nujno dovolj konteksta ali pa so slabo transkribirane. Da zagotovimo poravnavo vseh besed, smo razvili algoritem iterativnega združevanja besed.

Glavna ideja algoritma je naslednja: besede, ki niso poravnane, združimo s sosednjo besedo v besedilu (odstranimo presledek in tvorimo enoten niz znakov). Osnovni algoritem poravnave ponovno poženemo, tokrat z modificiranim seznamom besed. Ta dva koraka ponavljamo, dokler niso vse besede (oziroma skupki besed) poravnani, nato lahko vsaki besedi originalnega besedila pripišemo začetni in končni čas glede na približno transkripcijo.

3 Evalvacija

Natančnost sistema smo ovrednotili na testni množici s primerjavo z ročno izdelanimi poravnavami. Za oceno kvalitete poravnave uporabljamo tri mere: povprečje (MAE) in standardni odklon (STD) absolutnih napak začetnih časov besed ter delež absolutnih napak, manjših od 0,5 sekunde (< 0,5s).

3.1 Testna množica

Testno množico sestavlja 26 primerov: 7 primerov nenarečnega govora, 13 primerov narečnega govora in 6 primerov narečnega enoglasnega petja brez spremljave. Najkrajši posnetek je dolg 21 sekund, najdaljši 219, povprečna dolžina posnetkov je 89 sekund. Primeri so pridobljeni iz naslednjih virov: Slovenske ljudske pesmi V (Kaučič et al., 2007), portal narecja.si, terenski posnetki GNI ZRC SAZU. Pravilne poravnave so bile narejene ročno z orodjem Praat.

Tip posnetka	Število besed	Dolžina (min)
narečni govor	2428	18,7
nenarečni govor	1394	11,0
narečno petje	508	8,7
skupaj	4330	38,4

Tabela 1: Testna množica.

3.2 Primerjava modelov in metod dekodiranja

Primerjali smo osnovni akustični model (*base*), ki je grajen zgolj na slovenskih podatkih, ter model, ki je učen na angleških podatkih, nato pa doučen na slovenskih (*transfer*). Ob tem smo primerjali tri metode dekodiranja: požrešna metoda (*greedy*), iskanje v snopu z jezikovnim modelom na nivoju znakov (*char*), iskanje v snopu z jezikovnim modelom na nivoju besed (*word*). Primerjavo smo opravili za vsak tip testnih podatkov posebej. Rezultati so podani v Tabeli 2.

Iz tabele je razvidno, da pri nenarečnem govoru ne glede na metodo uporaba modela *transfer* prinese manjšo povprečno napako. Razlika je sicer majhna (0,06 do 0,07 sekunde), vendar je približno enaka za različne metode. Pri uporabi požrešne metode ima *transfer* sicer večji standardni odklon in manjši delež napak pod 0,5s, vendar je razlika minimalna. Različne metode dajejo zelo podobne rezultate. Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,12s, standardnim odkonom 0,10s in 99,4% deležem napak pod 0,5s.

Tudi v primeru narečnega govora uporaba modela *transfer* izboljša rezultate. Razlika v povprečnih napakah je majhna (0,04 do 0,09 sekunde), vendar je med akustičnima modeloma opazna razlika tudi v standardnem odklonu in deležu napak manjših od 0,5s. Z uporabo modela *transfer* so rezultati za različne metode poravnave zelo podobni, pri čemer se metoda *word* izkaže za najbolj robustno, saj ima najmanšo napako in standardni odklon pri obeh modelih. Pri modelu *transfer* ima metoda *greedy* sicer nekoliko večji delež napak pod 0,5s,

vendar je razlika majhna (0,4%). Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,14s, standardnim odkonom 0,24s in 97,3% deležem napak pod 0,5s. V primerjavi z najboljšim rezultatom nenarečnega govora se povprečna napaka poveča za 0,02s, standardna deviacija za 0,13s, delež napak pod 0,5s se zmanjša za 2,1%. Razlika ni velika in je približno podobna za ostale kombinacije metod in modelov.

tip testnih podatkov	metoda	model	MAE	STD	< 0,5s
Nenarečni govor	greedy	base	0,20	0,13	99,1%
		transfer	0,14	0,15	98,5%
	char	base	0,21	0,09	99,0%
		transfer	0,14	0,10	98,9%
	word	base	0,19	0,10	98,6%
		transfer	**0,12**	0,11	99,4%
Narečni govor	greedy	base	0,22	0,39	94,9%
		transfer	0,15	0,27	97,7%
	char	base	0,21	0,32	95,7%
		transfer	0,15	0,28	97,1%
	word	base	0,18	0,28	97,2%
		transfer	**0,14**	0,24	97,3%
Narečno petje	greedy	base	0,59	0,82	70,2%
		transfer	1,28	2,49	63,9%
	char	base	0,82	1,66	66,7%
		transfer	0,44	0,41	73,4%
	word	base	0,48	0,58	73,4%
		transfer	**0,37**	0,30	79,9%

Tabela 2: Rezultati

Pri narečnem petju je napaka poravnave opazno večja. Pri metodah *word* in *char* akustični model *transfer* deluje bolje. Z metodo *char* je povprečna napaka prepolovljena, standardni odklon je štirikrat manjši, delež napak pod 0,5s se izboljša za 6,7%. Z metodo *transfer* je povprečna napaka za 0,11s manjša, standardni odklon za 0,28s, delež napak pod 0,5s se izboljša za 6,5%. Pri metodi *greedy* je boljši model *base*, kar je edini tak primer v rezultatih. Rezultati različnih metod dekodiranja med seboj niso podobni. Pri obeh modelih metoda *word* bistveno izboljša rezultat. Kombinacija modela *transfer* in metode *word* da najboljši rezultat s povprečno napako 0,37s, standardnim odkonom 0,30s in 79,9% deležem napak pod 0,5s. V primerjavi z najboljšim rezultatom nenarečnega govora se povprečna absolutna napaka poveča za 0,25s, standardna deviacija za 0,19s in delež napak pod 0,5s se zmanjša za 19,5%. Razlika je velika in je vidna tudi pri ostalih kombinacijah metod in modelov. Povprečna absolutna napaka se poveča za faktor vsaj 2,5, standardni odklon za faktor vsaj 2,7 in delež napak pod 0,5s se zmanjša za vsaj 19,5%.

3.3 Ugotovitve

Kvaliteta poravnav na nenarečnem govoru se izkaže za dobro in je primerljiva s podobno delujočimi sistemi, npr. (Malfrère et al, 2003). Tudi pri narečnem govoru je kvaliteta poravnav dobra. Napaka je nekoliko večja kot pri nenarečnem govoru, kar je pričakovano, saj je večina učnih podatkov za akustični model nenarečnih. V splošnem ocenjujemo, da sistem dobro deluje na slovenskem govoru in je zato uporaben za večino aplikacij. Vredno je omeniti, da v primeru kratkih posnetkov in popolnih transkripcij za učenje akustičnih modelov obstajajo potencialno boljše tehnike poravnave (Brognaux in Drugman, 2015).

Kvaliteta poravnav enoglasnega petja brez spremljave je v primerjavi z govorom opazno slabša, kar smo tudi pričakovali, saj je v splošnem poravnava petja in besedila težji problem. V primerjavi z nenarečnim govorom je povprečna napaka približno trikrat večja in veliko več je napak večjih od pol sekunde. Povprečna napaka je sicer primerljiva s podobno delujočim sistemom za poravnavo petja (Stoller et al., 2019), vendar naši testni podatki ne vključujejo večglasnega petja ali petja s spremljavo, zato ta primerjava ne pove veliko. Domnevamo, da bi se kvaliteta poravnav bistveno izboljšala, če bi učna množica akustičnega modela vsebovala petje.

V veliki večini primerov se akustični model *transfer* izkaže bolje od modela *base*. Edini obraten primer je v primeru petja in metode *greedy*, kjer model *base* doseže boljši rezultat, vendar ker ta kombinacija metode in modela ne da najboljšega rezultata pri petju, ni bistvena za oceno kakovosti. Na podlagi rezultatov potrjujemo domnevo, da prenos znanja z modelom *transfer* pozitivno vpliva na kvaliteto poravnave tako pri govoru kot pri petju.

Čeprav je v primeru govora najboljša metoda za dekodiranje *word*, ostali dve metodi nimata bistveno večjih napak. V primeru nenarečnega govora z modelom *transfer* je povprečna napaka z metodo *word* manjša za 0,02s, v primeru narečnega govora pa za 0,01s. V aplikacijah, ko zelo natančna poravnava govora ni ključna, je pa pomemben čas računanja, je bolj smiselno uporabiti metodo *greedy,* saj le-ta ne zahteva iskanja v snopu ter uporabe jezikovnega modela in je zato bistveno hitrejša. Pri petju metoda *greedy* da bistveno slabše rezultate od metode *word*, zato je smiselno uporabiti slednjo.

Zahvala

Raziskave, opisane v prispevku, so bile opravljene v okviru temeljnega raziskovalnega projekta »Misliti folkloro: folkloristične, etnološke in računske perspektive in pristopi k narečju« (J7-9426, 2018-2022), programske skupine »Digitalna humanistika: viri, orodja in metode« (P6-0436, 2022-2027), oba financira ARRS, in raziskovalne infrastrukture DARIAH-SI.

Literatura

Sandrine Brognaux in Thomas Drugman. *Hmm-based speech segmentation: Improvements of fully automatic approaches*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24:1–1, 01 2015.

Simon Dobrišek, Jerneja Žganec Gros, Janez Žibert, France Mihelič, in Nikola Pavešić. *Speech database of spoken flight information enquiries SOFES 1.0*, 2017. Slovenian language resource repository CLARIN.SI.

Kenneth Heafield. *KenLM: Faster and smaller language model queries*. In Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics.

Ryan Hilleman, Tilman Kamp in Tobisas Bjornsson. *Dsalign*. https://github.com/mozilla/DSAlign, 2018.

Marjetka Golež Kaučič, Marija Klobčar, Zmaga Kumer, Urša Šivic, and Marko Terseglav. *Slovenske ljudske pesmi V*. 2007.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li in Yang Zhang. Quartznet: *Deep automatic speech recognition with 1d time-channel separable convolutions*, 2019.

F. Malfrère, O. Deroo, T. Dutoit, in C. Ris. *Phonetic alignment: speech synthesis-based vs. viterbi-based*. Speech Communication, 40(4):503–515, 2003.

Daniel Stoller, Simon Durand, in Sebastian Ewert. *End-to-end lyrics alignment for polyphonic music using an audio-to-character recognition model*, 2019.

VideoLectures.NET. *Spoken corpus gos VideoLectures 4.0 (audio)*, 2019. Slovenian language resource repository CLARIN.SI.

Ana Zwitter Vitez, Jana Zemljarič Miklavčič, Simon Krek, Marko Stabej in Tomaž Erjavec. *Spoken corpus gos 1.0*, 2013. Slovenian language resource repository CLARIN.SI.

Andrej Žgank, Mirjam Sepesy Maučec in Darinka Verdonik. *The SI TEDx-UM speech database: a new Slovenian spoken language resource*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4670–4673, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).