PD.TVZ.HR
POLYTECHNIC & DESIGN Vol. 9, No. 2, 2021.
Mrežna inačica: ISSN 2459-6302
Tiskana inačica: ISSN 1849-1995
DOI: 10.19279/TVZ.PD.2021-9-2-01

METODA AUTOMATSKE ANALIZE BRZINE GOVORA

A METHOD FOR AUTOMATIC ANALYSIS OF SPEECH TEMPO

Aleksandar Stojanović
Tehničko veleučilište u Zagrebu

SAŽETAK

U ovom radu opisana je metoda analize brzine govora ili *tempa* na osnovu uzoraka govora dobivenih s televizijskih kanala koji sadrže tekst izgovorenog u obliku titlova. Za prepoznavanje govora korištena je nepovratna neuronska mreža (engl. *feed-forward neural network*) trenirana s oko 160 sekundi govora. Da bi se odredile granice pojedinačnih riječi napravljena je komponenta za poravnavanje govora s tekstom koja pronalazi prihvatljivo podudaranje slova teksta s fonemima koje je klasificirala neuronska mreža. Komponenta za poravnavanje uzima u obzir kategorije fonema za koje neuronska mreža ima veću preciznost klasifikacije. Preliminarni rezultati pokazuju prosječne promašaje poravnavanja od jednog do tri fonema, zavisno od govornika, sadržaja izgovorenog i kvalitete snimke.

Ključne riječi: *prepoznavanje govora, poravnavanje, tempo, neuronska mreža.*

ABSTRACT

This paper describes a method for analysing speed of speech or *tempo* using speech recordings from Croatian TV news channels with subtitles. A feed-forward neural network was used for phoneme classification, trained with 160 seconds of recorded speech. To determine individual word positions a component for speech-to-text alignment was created which finds aproximate alignments of text from the subtitles and phonemes classified by the neural network. The alignment component relies on the fact that the neural network recognizes some groups of phonemes better than others.

Preliminary results showed an average alignment offset of one to about three phonemes, depending on the recording quality, speaker and the content.

Keywords: *speech recognition, alignment, tempo, neural network.*

1. UVOD
1. INTRODUCTION

Dvije često korištene mjere brzine govora su broj riječi i broj slogova u jedinici vremena. Broj slogova omogućava bolju granularnost jer su slogovi manje ovisni o duljini riječi. Na primjer, možemo izgovoriti dvije kratke riječi polako i tri dugačke riječi brzo unutar jednakog vremenskog intervala i mjerenjem broja riječi u tom vremenskom intervalu izgledalo bi da nema razlike u brzini govora jer su u oba slučaja izgovorene tri riječi. Ako ignoriramo pauze u govoru, trajanje riječi ovisi isključivo i trajanju vokala. Prema tome, važno je utvrditi ne samo koji fonem je izgovoren u određenom trenutku na snimci nego i njegovo trajanje. Dvije glavne komponente sustava za mjerenje brzine govora su klasifikator fonema i komponenta za poravnavanje govora s tekstom dostupnim u titlovima. Klasifikator fonema može se napraviti na više načina zavisno od toga treba li nam potpuno automatsko prepoznavanje govora ili ono za koje je dovoljno djelomično prepoznavanje. Pod „djelomičnim“ prepoznavanjem u ovom radu se podrazumijeva sustav koji ne može automatski konvertirati govor u tekst, ali može ispravno klasificirati dovoljno fonema da omogući poravnavanje govora s tekstom. Svrha poravnavanja govora s tekstom je da se utvrdi gdje (otprilike) svaka riječ na snimci počinje i završava.

Jedna poteškoća u korištenju djelomičnog prepoznavanja govora je u tome što mnogi fonemi mogu biti pogrešno klasificirani (što zavisi o količini podataka za treniranje i modela klasifikatora) tako da je svrha poravnavanja ta da se svako slovo teksta „poveže" s odgovarajućim fonemom na snimci. Postoji nekoliko radova u tom području. U radu [1] autori opisuju algoritam za poravnanje prema principu dinamičkog programiranja s matricom koja sadrži aproksimaciju poklapanja fonema s tekstom. U radu [2] opisana je metoda poravnavanja upotrebom skrivenih markovljevih modela (HMM ili *Hidden Markov Models*). Automatsko poravnavanje s detekcijom pogrešaka u ručno napravljenom transkriptu opisan je u [3]. Neke alternativne metode poravnavanja teksta s govorom opisane su u [4], [5], [6] i [7].

Automatsko prepoznavanje govora upotrebom neuronskih mreža opisana je u [8] i [9]. Postoje i neki prethodni radovi vezani za proučavanje hrvatskog govora. U [10] autori opisuju metodu za modeliranje osnovnih intonacijskih oblika upotrebom analize sintezom. U [11] autori analiziraju kako naglašavanje vokala i trajanje utječu na frekvencije formanata. U [12] opisan je model strukture vremena standardnog hrvatskog govora. U [13] opisani su rezultati akustičke analize izgovorenih vokala i vokala sintetiziranih u trodimenzionalnom prostoru F1xF2xF3 hrvatskog standardnog govora. U [14] opisana je metoda procjene varijacija govornog tempa.

2. CILJ
2. GOAL

Cilj ovog istraživanja je odgovoriti na sljedeća pitanja: 1) Kolika se preciznost poravnavanja teksta s govorom na snimci može postići bez upotrebe sustava za potpuno automatsko prepoznavanje govora? Pod takvim sustavom ovdje se podrazumijeva sustav koji je napravljen kao korisničko sučelje za govor s velikim rječnikom i koji je nezavisan od govornika. Iako postoje takvi sustavi za engleski jezik u ovom radu upotrebljava se hrvatski jezik pa su jezični i akustički modeli drugačiji. Za ovo istraživanje pristup prepoznavanju govora je pojednostavljen jer nije bilo potrebe za ustanovljavanjem šta je bilo izgovoreno nego gdje je na snimci nešto bilo izgovoreno. 2) Koristeći ovakav pristup je li moguće mjeriti brzinu govora ako imamo tekst onoga što je izgovoreno? Pretpostavka je da je moguće dobiti dovoljno informacija o fonemima pomoću klasifikatora, pogotovo o vokalima, kao što je njihova temporalna pozicija na snimci i trajanje, da bi se odredilo trajanje pojedinih govornih segmenata pomoću kojih je moguće rekonstruirati trajanje i izmjeriti brzinu govora.

3. Materijali i metode
3. Materials and methods

U ovom dijelu opisani su alati i metode korišteni za ovo istraživanje. Za analizu zvuka korišten je Praat sustav [15], a a neuronsku mrežu biblioteka Scikit-Learn za Python i njen MLPClassifier modul. Postupak se sastojao od četiri općenita koraka:

1. Segmentacija govora po pojedinačnim fonemima i na osnovu toga treniranje neuronske mreže.
2. Prepoznavanje govora.
3. Poravnavanje teksta s fonemima.
4. Analiza brzine govora.

Tri su vrste ulaznih podataka za ovaj postupak:

- LTAS[1] vrijednosti dobivene iz snimke govora gdje svakoj vrijednosti odgovara duljina snimke od oko 20 do 30 sekundi.
- Tekst segmenta govora koji analiziramo (jedna ili dvije rečenice).
- Aproksimacija intervala snimke na kojoj taj tekst počinje i završava.

Za prepoznavanje govora korištena je nepovratna neuronska mreža s tri skrivena sloja veličine 120, 120 i 160. Za treniranje neuronske mreže korišteni su samo muški glasovi, ali za prepoznavanje uzeti su i muški i ženski glasovi. Za snimke govora korišteni su televizijski kanali za vijesti, a većina govora snimljena je u studiju.

1 LTAS je akronim za Long Term Average Spectrum.

Segmentacija govora i treniranje neuronske mreže

Snimke odabrane za treniranje neuronske mreže podijeljene su na manje dijelove nad kojima je napravljena segmentacija govora. Ukupna duljina zvuka za treniranje je oko 160 sekundi. Iako to nije velika količina podataka za treniranje, kao što je rečeno, cilj nije bio napraviti sustav za automatsko prepoznavanje govora nego sustav za poravnavanje govora s tekstom za koji je korišten kasnije opisani algoritam. Za treniranje neuronske mreže za svaki fonem izdvojen segmentacijom izračunat je njegov LTAS i spremljen u datoteku zajedno sa slovom koje označava. Širina pojasa za LTAS bila je 100 Hz gdje je izdvojeno samo prvih 120 vrijednosti (da bi odgovaralo gornjem pragu frekvencije od 12 kHz). Te su vrijednosti bile normalizirane upotrebom *feature scaling* metode da bi ih se svelo na interval [0, 1]. Nadalje, te su vrijednosti bile transformirane da bi se spriječilo to da intenzitet utječe na efikasnost treniranja gdje fonemi sličnog spektralnog oblika, ali različitog intenziteta izgledaju previše različito neuronskoj mreži. Ti su podaci onda korišteni kao skup za treniranje i propušteni kroz neuronsku mrežu gdje je svaki LTAS segment od 100 Hz predstavljao jedan ulazni parametar (od ukupno 120) za ulazni sloj mreže. Tijekom treniranja pauze u govoru ispod zadanog praga intenziteta bile su preskočene.

Prepoznavanje govora

Prepoznavanje govora odvijalo se u dva koraka. Prvo, ulazni je zvuk bio podijeljen u uzastopne segmente od 10 ms i za svaki taj segment bio je izračunat LTAS. U drugom koraku na mjestu na snimci na kojem su prisutni glotalni impulsi (odnosno na kojima ima zvuka) bio je izdvojen segment od 10 ms (5 ms ispred i iza te pozicije) i za svaki taj segment bio je izračunat LTAS. Ovaj drugi korak je važan da bi se izbjeglo slučajno preskakanje važnog dijela zvuka iz prvog koraka. Ovdje je važno napomenuti da ovaj drugi korak nije obuhvaćao frikative jer oni se ne sastoje od glotalnog zvuka tako da se za taj tip fonema oslanjamo isključivo na prvi korak. Podaci dobiveni iz ova dva koraka su nakon toga bili spojeni na osnovu vremena u kojem se pojavljuju (to jest, temporalne pozicije segmenta zvuka) i korišteni kao ulazni skup za postupak klasifikacije neuronskom mrežom. Postupak prepoznavanja (klasifikacije fonema) rezultirao je time da su se mnogi fonemi uzastopce ponavljali (zbog toga što većina fonema traje duže od 10 ms i spajanja segmenata zvuka iz dva prethodna koraka). Takvi su fonemi bili grupirani po jednakosti i klasi fonema kojoj pripadaju. Fonemi su grupirani u četiri klase prikazane u tablici 1. Rezultat klasifikacije neuronske mreže je, prema tome, bio transformiran u niz grupa fonema uređen prema vremenu pojavljivanja na snimci.

Tablica 1. *Četiri klase fonema.*

Table 1. *Four classes of phonemes.*

Vokali	*a, e, i, o, u*
Frikativi 1	*š, č*
Frikativi 2	*z, s, c*
Frikativi 3	*ž, đ*
Ostalo	Svi ostali fonemi

Ispis na slici 1 prikazuje jedan primjer gdje je na intervalu 1 neuronska mreža klasificirala fonem kao *i* u tri uzastopna segmenta od 10 ms, a u intervalu 2 klasificirala je fonem kao *s* na isti način. Međutim, na intervalima 3 i 5 klasificirala je segmente kao dva različita fonema, *o* i *a*, i s obzirom da ti fonemi pripadaju istoj kategoriji stavljeni su u istu grupu. Također, na intervalu 4 fonemi koji ne pripadaju niti jednoj od prve četiri kategorije stavljeni su u istu grupu.

1. iiiii	[i]
2. sssssssss	[s]
3. aaaaaaaaaaao	[a, o]
4. ndf	[n, d, f]
5. ooaaaaaooo	[o, a]
6. mnnmmmn	[m, n]
7. oaooa	[a, o]

Slika 1 *Grupiranje fonema koje je klasificirala neuronska mreža.*

Figure 1 *Grouping of phonemes classified by the neural network.*

Jedan važan aspekt ovakve klasifikacije fonema je taj da ona sadrži trajanje fonema (vidljivo kroz duljinu niza), posebno vokala, tako da je nakon poravnanja moguće mjeriti trajanje slogova i riječi.

Na primjer, u tablici 2 na intervalu [20.05, 20.18] fonem je bio klasificiran kao *a* ili *o*, a njegovo je trajanje bilo malo duže od vokala na intervalu [18,94, 19,33].

Tablica 2. *Primjer ispisa niza grupa fonema na segmentu zvuka između 18.94 i 20.22.*

Table 2. *An example of a sequence of groups of phonemes on a sound segment between 18.94 and 20.22.*

SEGMENT ZVUKA	GRUPE FONEMA
18.94-19.33	aooooooaaaaaaao
19.34-19.35	mmml
19.36-19.38	aaao
19.39-19.49	ooooooooooooooooo
19.51-19.51	n
19.52-19.56	**ouuouoooouoou**
19.56-19.56	**v**
19.58-19.62	**uoooaaooaaaaoa**
19.63-19.65	**aaooaaeaeo**
19.66-19.67	**ššččš**
19.68-19.69	**sszs**
19.76-19.87	**aaaaaaaaaaaaaaaa**
19.88-19.90	**aoaaa**
19.91-19.92	**ooaa**
19.93-19.94	tv
19.95-20.04	zcszsscccss
20.05-20.18	aaaaaaaaaaaaaaooooooaaaoooo
20.20-20.20	m
20.20-20.22	aaoaoa

Poravnavanje

S obzirom da neuronska mreža mnoge foneme nije ispravno klasificirala napravljen je algoritam za poravnavanje klasificiranih fonema s tekstom. Cilj tog algoritma je procjena početka i kraja intervala zvuka na kojem se nalazi neka riječ. Na primjer, za niz fonema *zcsonmeaoe* i riječ *soba* algoritam je poravnao s-s, o-o, m-b i a-a kako je pokazano na slici 2. Nadalje, algoritam je bio napravljen tako da razmak između dva poravnata fonema ne može biti veći od prosječnog trajanja dva fonema. To je napravljeno zbog toga da se izbjegnu neka nemoguća poravnanja gdje bi razmak između dva fonema bio previše nerealističan za normalni govor.

Niz fonema	z	c	s	o	n	m	e	a	o	e
Tekst			s	o		b		a		

Slika 2 *Primjer poravnanja.*

Figure 2 *An alignment example..*

Algoritam kao ulazne podatke prima tekst (oko 15 do 20 riječi) i listu fonema koje je klasificirala neuronska mreža (kao u tablici 2) i vraća aproksimaciju poravnanja svih slova teksta s grupom fonema. Ovaj algoritam ne vraća uvijek optimalno poravnanje, što zavisi od kombinacije slova i fonema. Jedan detalj koji je bio napravljen prije upotrebe ovog algoritma je transformacija teksta tako da bolje odgovara govoru što se tiče koartikulacije i asimilacije. Jedan problem kod ovog algoritma je preciznost poravnanja kada je manje fonema nego što je slova teksta. U tom slučaju algoritam „posuđuje“ foneme iduće riječi tako da nakon tog mjesta poravnanje može značajno odstupati.

Analiza brzine govora

Vremenski interval koji zauzima svaka riječ dobiven je poravnanjem teksta s fonemima. Da bi analizirali brzinu govora te su riječi bile spojene i podijeljene u segmente (u redoslijedu u kojem su se pojavljivale u tekstu) tako da je svaki segment sadržavao najmanje pet slogova. S obzirom da su vremenke granice riječi bile utvrđene poravnanjem, brzina je izračunata kao broj slogova po vremenskom intervalu segmenta. Slogovi umjesto riječi su odabrani za određivanje brzine govora jer oni imaju bolju granularnost od riječi. Na primjer, unutar intervala od dvije sekunde moguće je izgovoriti tri kratke riječi sporo ili tri dugačke riječi brzo tako da brojanje izgovorenih riječi unutar intervala ne bi bilo dovoljno precizno.

4. REZULTATI
4. RESULTS

U ovom dijelu prikazani su rezultati poravnanja govora s tekstom i detekcije brzine govora upotrebom snimki tri muška i tri ženska glasa s hrvatskih televizijskih kanala za vijesti. Trajanje snimki za muške glasove bilo je 6.2, 6.1 i 5.3 sekundi, a za ženske 6, 5.3 i 7.4 sekundi. Niti jedan od tih glasova nije bio upotrijebljen za treniranje neuronske mreže. Nadalje, kako je prethodno spomenuto, neuronska mreža bila je trenirana samo s muškim glasovima.

Poravnanje

Polazna pretpostavka bila je da poravnanje neće biti u potpunosti precizno jer neuronska mreža nije ispravno prepoznavala sve foneme. Također, prepoznavala je neke grupe fonema bolje od ostalih. Primjerice, foneme *s*, *c*, *š* i *č* klasificirala je puno preciznije od drugih klasa fonema kao što su okluzivi i nazali. Algoritam poravnanja pronalazio je podudaranja slova teksta s grupom fonema upotrebom sustava „bodovanja“ koji je uzimao u obzir činjenicu da neuronska mreža bolje prepoznaje prvu grupu fonema u tablici 1 od ostalih. Na primjer, algoritam poravnanja pridružio je više bodova kod poravnanja fonema *s* s grupom kao što je ['s', 'c', 'z'] nego s grupom kao što je ['d', 'm']. Primjer ispisa u tablici 2 ilustrira kako je ovaj algoritam procijenio gdje se nalazi riječ *oproštaja* u nizu fonema koje je klasificirala neuronska mreža. Niz grupa fonema za tu riječ prema procjeni algoritma za poravnanje istaknuta je debljim slovima.

U slučajevima s poravnanjem pojedinačnih riječi algoritam za poravnanje manje je precizan nego kod poravnavanja dužih segmenata. To se moglo očekivati jer kod poravnavanja kraćeg segmenta teksta kao što je jedna riječ vjerojatnije je da će postojati više segmenata grupa klasificiranih fonema koje bi dale prihvatljivo poravnanje jer ne postoji dovoljno elemenata kojima bi dotični segment bio jedinstven. Tablice 3 i 4 pokazuju rezultate poravnavanja na šest uzoraka (tri muška i tri ženska glasa), svaki s drugačijim govornikom. Za svakog govornika postoje dva stupca: *od* i *do*. Svaki redak s tim stupcima predstavlja jednu riječ izgovorene rečenice (drugačija rečenica za svakog govornika) s brojevima koji označavaju odstupanje poravnanja od onog ispravnog (utvrđenog slušanjem) kako je odredio algoritam poravnanja. Stupci *od* i *do* sadrže početno i završno vrijeme izgovorene riječi, to jest njeno poravnanje s tekstom. Negativna vrijednost (u milisekundama) znači da je poravnanje pomaknuto u lijevo (počinje prerano), a pozitivna da je pomaknuto u desno (završava prekasno). Ovi podaci pokazuju da je na ovih šest uzoraka prosječno odstupanje bilo od jednog do tri fonema pod pretpostavkom da je prosječno trajanje fonema za hrvatski jezik oko 76 milisekundi ([16] i [17]).

Jedan od faktora koji su utjecali na ova odstupanja je taj da je neuronska mreža slabo prepoznavala okluzive (kao *p* i *t*). Iako neuronska mreža nije bila trenirana sa ženskim glasovima rezultati pokazuju da prepoznavanje fonema može biti relativno dobro s takvim govornicima, zavisno od kvalitete snimke i jasnoće izgovora. To je bilo za očekivati jer je prosječan spektar ženskog glasa lagano pomaknut prema višim frekvencijama i malo više zaobljen, ali njegov je oblik isti kao i kod muških glasova. U slučaju uzorka 2 u tablici 4 prosječno odstupanje bilo je manje od jednog fonema, iako je više izraženo na uzorcima 1, 2 (tablica 3) i 3 (tablica 4).

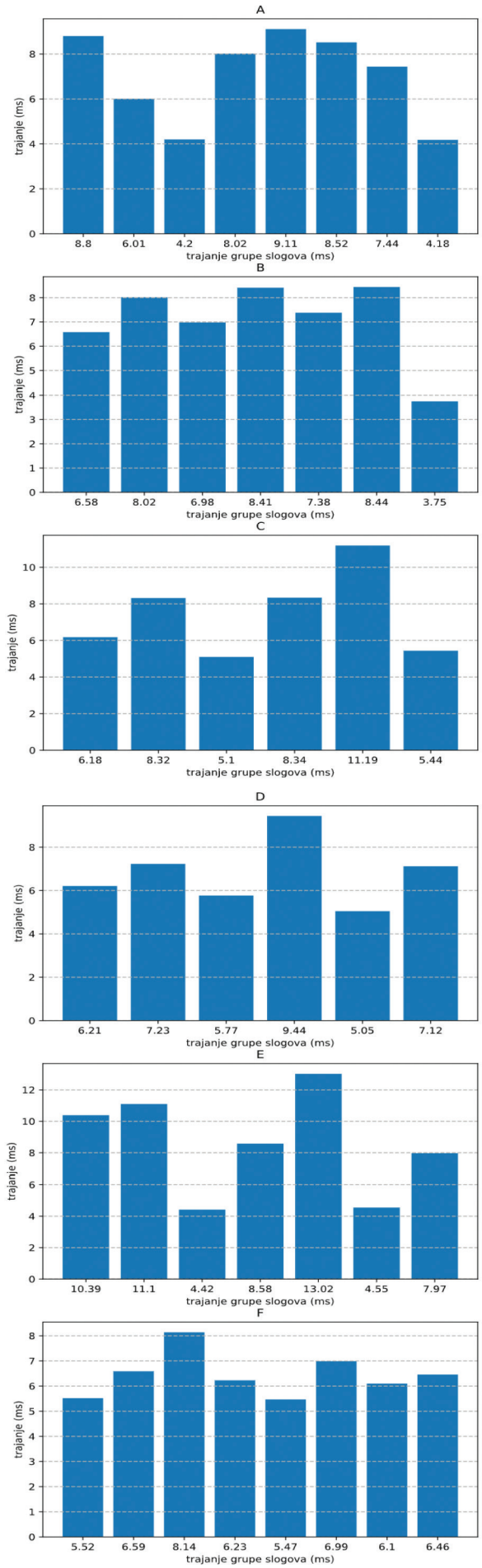
Detekcija brzine govora

Slika 3 pokazuje utvrđene varijacije u brzini govora za istih šest govornika. Svaki stupić predstavlja trajanje grupe riječi izračunato kao *broj slogova / dužina intervala*. Ovi dijagrami pokazuju da su većina govornika imali uravnoteženu brzinu govora na ovim uzorcima, s tim da dijagrami *a*, *e* i *c* pokazuju manje varijacije u sredini. Sve su te varijacije male i teško uočljive slušanjem.

Tablica 3. *Rezultat poravnanja na tri uzorka govora s muškim govornicima.*

Table 3. *Alignment on three samples of speech with different male speakers.*

1 (muški)		*2 (muški)*		*3 (muški)*	
OD	*DO*	*OD*	*DO*	*OD*	*DO*
16	-59.0	1	214.4	21	54.3
-24.9	-116	221.9	167	131	135.5
-118.8	-287.1	193.8	132.3	187.9	177.5
-274	-186.9	91	27.6	181	319
-284.6	15.1	38.8	-11.8	319	351
16	-4	101	1	121.0	-0.3
13.9	-4	11	61.6	101	66.9
14	-17.9	81	54	78	21
-36.1	-48.9	37.5	-16.2	71	4.3
-36.9	-31.1	21	78	11	71
-27.9	304	28.3	183.1	11	-66.1
212.1	411.9	232	227.8	1	-3.6
71	10.9	233	66.5	5.2	14.1
34	-11.0	42.9	204.4	24.3	31
25	-8.7	141	277.1	17.2	-26.9
4.1	-82.0	299	396	11	-0.5
-84	-211.0	378	61	7.6	-23
-197.1	-185.1			-13.3	86.0
-193.1	-219.8			91	334.3
-206.9	-244				
-270	-263.1				
-114	5.9				
104.38	**125.6**	**127.12**	**128.44**	**72.95**	**94.13**



Tablica 4. *Rezultat poravnanja na tri uzorka govora sa ženskim govornicima.*

Table 4. *Alignment on three samples of speech with different female speakers.*

1 (ženski)		*2 (ženski)*		*3 (ženski)*	
OD	*DO*	*OD*	*DO*	*OD*	*DO*
1	108.9	9.4	137.5	29.2	234.5
124.3	106.1	132	150.1	237.1	-4
111.3	88.2	197.8	2.1	-51	58.9
11	27.5	86.1	-4.2	181	260.1
1	-13.0	1	-3.3	213.9	261
-82.7	-8.2	-43.1	-51	262	205.2
8.9	-206.9	57.2	1	137.1	195.2
-192.7	-3.2	19.9	58	142.1	-114.8
54.3	-14.9	62.8	42.1	-119.1	-26.3
-109.5	131	41	41	17.9	51
133.1	162.5	11	5	-361	-6.2
155.9	-14.7	21	11	21	-5.8
-3.8	213.9	12.1	-13	-48.4	-22.2
151	-85.1	72.1	134.9	11	8
-81	-109.1	56.0	-5.2	11	-38.8
-16.3	78.2	-20.0	-7.8	-42	-229
83.2	19.8	71	-9.3		
-67.8	-61.1	-52.1	-4.9		
		15.7	-93.2		
		-91.5	51		
		129.5	11		
		11.1	-55.1		
77.01	**81.05**	**55.2**	**39.05**	**118.11**	**106.92**

Slika 3 *Dijagrami varijacija u brzini govora na uzorcima tri muška (a, b, c) i tri ženska (d, e, f) govornika.*

Figure 3 *Diagrams showing speech tempo variations on three samples of male (a, b, c) and three samples of female (d, e, f) speakers.*

5. ZAKLJUČAK
5. CONCLUSION

Rezultati pokazuju da kada je na raspolaganju tekst izgovorenog za mjerenje brzine govora moguće je napraviti sustav za djelomično prepoznavanje govora za hrvatski jezik s relativno malom količinom podataka za treniranje. To može biti korisno za jezike za koje ne postoje komercijalni sustavi za prepoznavanje govora kao što postoje, primjerice, za engleski jezik. Algoritam za poravnavanje govora s tekstom može se upotrijebiti da bi se aproksimirale granice riječi na osnovu rezultata prepoznavanja fonema od strane neuronske mreže. Iako se mogu upotrijebiti razni algoritmi za parcijalno poravnavanje teksta s nizom fonema (kao što je *Edit distance*), algoritam koji uzima u obzir to da neuronska mreža prepoznaje neke grupe fonema bolje od drugih vjerojatno daje bolje rezultate, iako u ovom istraživanju nije napravljena direktna usporedba s drugim algoritmima.

6. REFERENCE
6. REFERENCES

[1.] X. Anguera, J. Luque and C. Gracia, "Audio-to-text alignment for speech recognition with very limited resources," in INTERSPEECH 2014, ISSN 1990-9770, Singapore, 2014.

[2.] S. Hoffmann and B. Pfister, "Text-to-Speech Alignment of Long Recordings Using Universal Phone Models," in INTERSPEECH 2013, ISSN 2308-457X, Lyon, 2013.

[3.] T. J. Hazen, "Automatic alignment and error correction of human generated transcripts for long speech recordings," in INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, ISSN 1990-9772, Pittsburgh, USA, 2006.

[4.] G. Bordel, S. Nieto, M. Penagarikano, L. J. Rodriguez-Fuentes and A. Varona, "A simple and efficient method to align very long speech signals to acoustically imperfect transcriptions," in INTERSPEECH 2012, ISSN 1990-9770, Portland, 2012.

[5.] P. J. Moreno, C. Joerg, J.-M. V. Thong and O. Glickman, "A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments," in 5th Int. Conf. on Spoken Language Processing, Sydney, 1998.

[6.] A. Stan, P. Bell and S. King, "A Grapheme-Based Method for Automatic Alignment of Speech And Text Data," in Spoken Language Technology Workshop (SLT), 2012.

[7.] P. J. Moreno and C. Alberti, "A Factor Automaton Approach for the Forced Alignment of Long Speech Recordings," in ICASSP, ISSN 1520-6149, Taipei, 2009.

[8.] D. Dhanashri and S. Dhonde, "Speech Recognition Using Neural Networks: A Review," International Journal of Multidisciplinary Research and Development, ISSN 2349-4182, vol. 2, no. 6, pp. 226-229, 2015.

[9.] R. P. Lippmann, "Review of Neural Networks for Speech Recognition," Readings in Speech Recognition, A. Waibel and K. F. Lee, Editors, Morgan Kaufmann Publishers, ISBN 9780080515847, pp. 374-392, 1990.

[10.] J. Bakran, V. Erdeljac and N. Lazić, "Modeliranje temeljnih intonacijskih oblika," Govor, ISSN 0352-7565, no. 2, pp. 105-111, 2001.

[11.] J. Bakran, "Djelovanje naglaska i dužine na frekvencije formanata," Govor, ISSN 0352-7565, vol. VI, no. 2, pp. 1-12, 1989.

[12.] J. Bakran, Model vremenske organizacije hrvatskog standardnoga govora, Zagreb: PhD thesis, unpublished, 1984.

[13.] J. Bakran and M. Stamenković, "Formanti prirodnih i sintetiziranih vokala hrvatskoga standardnoga govora," Govor, ISSN 0352-7565, vol. VII, no. 2, pp. 119-138, 1990.

[14.] A. Stojanović and N. Lazić, "A Method for Estimating Variations in Speech Tempo from Recorded Speech," MIPRO 2019, ISBN 978-1-5386-9296-7, pp. 1277-1282, 2019.

[15.] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed 15 5 2018].

[16.] J. Bakran, Zvučna slika hrvatskoga govora, Zagreb: IBIS grafika, ISBN 978-9539657725, 1996.

[17.] S. Babić, D. Brozović, M. Moguš, S. Pavešić, I. Škarić and S. Težak, Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika, Zagreb: Hrvatska akademija znanosti i umjetnosti, Globus, Nakladni zavod, ISBN 86-407-0014-1, 1991.

AUTOR · *AUTHOR*

● Aleksandar Stojanović
Predavač je na Tehničkom veleučilištu u Zagrebu i izvodi nastavu na predmetima iz područja programskog inženjerstva, objektno-orijentiranog programiranja i naprednog programiranja. Godine 1996. diplomirao je informacijske i komunikacijske znanosti na Filozofskom fakultetu sveučilišta u Zagrebu, a 1999. godine magistrirao je računarstvo u SAD-u na sveučilištu Midwestern State University te je radio kao programski inženjer u području telekomunikacija, financija i energetike. Godine 2019. doktorirao je na Filozofskom fakultetu sveučilišta u Zagrebu s disertacijom pod naslovom "Metoda automatske detekcije naglašenih riječi u zvučnom zapisu". Njegova područja interesa uključuju programske jezike, prevodioce i principe programiranja. Autor je knjige “Elementi računalnih programa s primjerima u Pythonu i Scali”.

Korespondencija · *Correspondence*
aleksandar.stojanovic@tvz.hr