Postprocesing metode za validaciju prepoznavanja AlfaNum ASR sistema

Dragiša Mišković, Mirjana Zindović, Darko Pekar

***Sadržaj* — U radu je opisan način za poboljšanje kvaliteta prepoznavanja sistema za automatsko prepoznavanje kontinualnog govora (eng. Continuous Automatic Speech Recognition) na srpskom jeziku. Osnovna ideja je da se u govornim aplikacijama (govorni automati, pozivni centri...) poveća robusnost sistema primenom metoda validacije izlaznih rezultata. Prva metoda se odnosi na praćenje energije prepoznatih reči a druga na modelovanje njihovog trajanja. U oba slučaja radi se o postprocesing metodama koje koriguju vrednosti dobijene na izlazu Viterbijevog algoritma.**

***Ključne reči* — modelovanje trajanja, prepoznavanje govora.**

I. Uvod

NAJPOPULARNIJA komercijalna primena govornih tehnologija je u okviru govornih automata. Oni predstavljaju aplikacije računarske telefonije, koje omogućavaju korisnicima da bez učešća operatera (osim kada to sami zatraže) pristupe velikoj količini informacija preko običnog telefona, kao i da istim putem iniciraju određene radnje, vršenje rezervacija, pokretanje i kontrola transakcija i slično. Uspešnost govorne CTI (eng. Computer Telephony Integration) aplikacije pored adekvatnog ASR (eng. Automatic Speech Recognition) algoritma zavisi i od pažljivo osmišljenog dijaloga između mašine i čoveka čime se skup reči koje treba prepoznati svodi u realne okvire.

Međutim, veoma su česte situacije kada se korisnici zabune i ne znaju šta se od njih očekuje kao odgovor. Tada oni koriste reči koje su izvan definisanog rečnika za prepoznavanje, odgovaraju sa celim rečenicama ili postavljaju pitanja automatu. U tim slučajevima ocena kvaliteta prepoznavanja određuje dalji tok govorne aplikacije. Kada je prepoznavanje nepouzdano korisnik se jednostavno zamoli da ponovi odgovor na prethodno postavljeno pitanje.

Ovaj rad je realizovan u okviru istraživačkog projekta koga finansira Ministarstvo za nauku države Srbije pod nazivom "Razvoj govornih tehnologija za srpski jezik i primena u 'Telekomu Srbija' " (TR-6144A).

Dragiša M. Mišković, Fakultet tehničkih nauka u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-4852521; faks: 381-21-4752997; e-mail: dragisa@uns.ns.ac.yu).

Mirjana J. Zindović, stipendista Ministarstva za nauku na Fakultetu tehničkih nauka u Novom Sadu, Srbija (telefon: 381-21-4750012; faks: 381-21-4752997; e-mail: mirjanazindovic@yahoo.com).

Darko J. Pekar, Alfanum d.o.o. u Novom Sadu, Trg Dositeja Obradovića 6, 21000 Novi Sad, Srbija (telefon: 381-21-4750080; faks: 381-21-4750080; e-mail: darko.pekar@alfanum.co.yu).

Metode koje su opisane u ovom radu spadaju u grupu postprocesing metoda tj. primenjuju se nakon što je završeno prepoznavanje govora.

Rad je organizovan po delovima. Nakon uvodnog dela, specifikacije govorne baze i načina zadavanja gramatike, četvrti deo rada odnosi se na osnovne parametre obuke i modela. Faktori koji se koriste pri određivanju ocene kvaliteta prepoznavanja su opisani u petom delu. Šesti deo rada se odnosi na rezultate dobijene prilikom testiranja sistema.

II. Govorna baza

U radu je korišćen deo SpeechDat(E) govorne baze [1] koji obuhvata samo muške govornike. SpeechDat(E) govorna baza na srpskom jeziku je dizajnirana za potrebe sistema sa prepoznavanjem kontinualnog govora preko fiksne telefonske mreže. Snimci su u 8kHz *A-law* formatu sa 8 bita po odmerku.

III. Gramatika

U govornim aplikacijama zasnovanim na Alfanum sistemu za prepoznavanje govora gramatike se zadaju u vidu EBNF forme (eng. Extended Backus Naur Form) pri čemu se interakcija između čoveka i mašine deli na segmente sa tačno definisanim skupom reči za prepoznavanje.

IV. Osnovni parametri obuke i modela

Sistem za automatsko prepoznavanje govora koji je korišćen u ovom radu je baziran na primeni fonetskog prepoznavača govora nezavisnog od govornika. On koristi skrivene Markovljeve modele (eng. Hidden Markov Models) za predstavljanje fonema. Karakteristike fonema zavise od susednih fonema, tako da je osnovna jedinica prepoznavanja fonem (monofon) u određenom kontekstu odnosno trifon.

Svaki fonem je sastavljen od određenog broja subfonema u zavisnosti od trajanja monofona iz kog je trifon izveden. Subfonemi odgovaraju različitim stanjima u okviru HMM modela. Na osnovu akustičkih obeležja, dobijenih obradom govorne baze, svako stanje (subfonem u kontekstu) se opisuje sumom Gausovih raspodela (GMM) koje sa centroidom, varijansom i težinom (ponderom te raspodele u sumi) predstavljaju dato stanje u prostoru akustičkih obeležja [2]. Ono što se modeluje pomoću GMM je verovatnoća da je stanje generisalo neku observaciju. Broj Gausovih raspodela je srazmeran varijabilnosti obeležja kojim je opisano dato stanje.

Modeli fonema na kojima se zasniva AlfaNum ASR formiraju se na osnovu tzv. *semi continuous HMM* pristupa (SCDHMM) koji podrazumeva da je celokupni skup smeša nastao u toku obuke indeksiran a svakom modelu fonema (subfonema) u kontekstu se pridružuju određeni indeksi smeša koji mu pripadaju. Ovo znači da određene smeše ne moraju pripadati samo jednom modelu ali se stepen pripadnosti određuje na osnovu težina za svaku smešu pridruženu modelu.

Pored toga za svaki model se čuva raspodela njegovog trajanja, koja se kao i preostali parametri estimira na osnovu podataka sadržanih u bazi za obuku. Na osnovu ovog podatka se modeluje trajanje modela pri prepoznavanju.

Raspodela trajanja stanja (eng. state duration distribution), koja se čuva u vidu histograma, predstavlja takođe polaznu osnovu za naknadnu verifikaciju rezultata prepoznavanja i korekciju verodostojnosti kao ocene pouzdanosti da je prepoznavanje ispravno.

Skup obeležja čine statička i dinamička obeležja. U statička obeležja spada 12 kepstralnih koeficijenata, normalizovana energija i log energija. Dinamička obeležja čine prvi i drugi izvod ovih statičkih obeležja. Ovako formiran vektor obeležja, dimenzije 42, podeljen je u dva strima. Prvi strim čine koeficijenti koji opisuju energiju (dimenzije 6) a drugi strim koeficijenti koji opisuju obvojnicu spektra (kepstralni koeficijenti, dimenzije 36).

V. Faktori ocene kvaliteta prepoznavanja

A. Akustička pouzdanost

Na izlazu iz prepoznavača poznata je verovatnoća (metrika) za svaki frejm datog zvučnog fajla. Akustička pouzdanost prepoznate reči se određuje kao prosek 50% najlošijih metrika frejmova koji pripadaju datoj reči.

Pored akustičke pouzdanosti se računa i prosečna akustička metrika. Prosečna akustička metrika se određuje kao prosek metrika svih frejmova na nivou jedne prepoznate reči.

B. Praćenje energije

Normalizovana energija (u nastavku rada samo energija) kao jedno od obeležja govornog signala se određuje na nivou frejma. U našem sistemu frejm je segment govornog signala čija je dužina 10ms.

Za svaku prepoznatu reč u sistemu određuje se energija reči kao suma energija svih frejmova koji pripadaju datoj reči. Takođe, potrebno je odrediti ukupnu energiju datog zvučnog fajla. Ona se računa kao suma energija svih frejmova koji pripadaju tom fajlu.

Na osnovu odnosa energije svih prepoznatih reči i energije u preostalom delu zvučnog fajla formira se vrednost koja je označena kao koeficijent energije (k_e).

U zavisnosti od vrednosti ovog koeficijenta primenjuje se procedura koja je dobijena kao kompromis između težnje da se ocena kvaliteta prilikom lošeg prepoznavanja dodatno umanji a sa druge strane da to smanjenje ne bude preveliko.

Ako je koeficijent energije u intervalu od 0 do 2 (ne uzimajući u obzir 0 i 2), izlazna metrika prepoznate(ih) reči se smanjuje za vrednost *s* koja se određuje na osnovu relacije:

$$s = k_s \cdot (2 - k_e) \tag{1}$$

pri čemu k_s predstavlja koeficijent skaliranja. U ovom radu vrednost koeficijenta skaliranja je postavljena na 10.

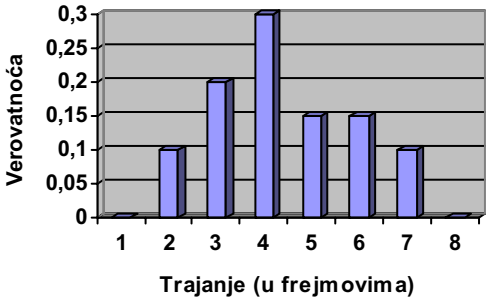
Ukoliko je dobijeni koeficijent energije izvan navedenog intervala od nula do dva izlazna metrika prepoznate(ih) reči ostaje nepromenjena.

C. Skorovanje trajanja

Skorovanje trajanja fonema (i reči) predstavlja postprocesing metodu koja uzima u obzir dva pristupa prilikom modelovanja trajanja:

- modelovanje trajanja svakog fonema pojedinačno odnosno modelovanje nezavisno od konteksta (MTNK – modelovanje trajanja nezavisno od konteksta)
- modelovanje koje ocenjuje trajanje svakog fonema u donosu na ostatak reči (MTZK – modelovanje trajanja zavisno od konteksta).

Osnova za prvi vid modelovanja trajanja (MTNK) ostvarena je u toku obuke modela na osnovu inicijalne raspodele vektora akustičkih obeležja po stanjima [3]. Ova raspodela se čuva u vidu 2D matrice čijom normalizacijom se dobija histogram trajanja. Sl. 1 predstavlja primer histograma trajanja jednog modela.



Sl. 1. Raspodela trajanja određenog stanja.

Pristup modelovanju koji se oslanja samo na histogram praktično je neprimenjiv usled čestih pojava nedovoljnog broja instanci pojedinih modela u bazi. Usled toga se trajanje svakog modela modeluje gama raspodelom koja je data sa:

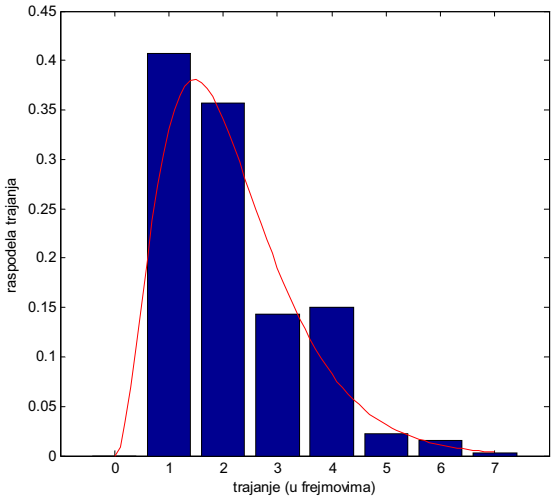
$$\rho(\tau) = K \cdot e^{-\alpha\tau} \cdot \tau^{\rho-1} \qquad \tau = 0, 1, 2, \ldots \tag{2}$$

pri čemu *K* predstavlja normalizacioni koeficijent odnosno normalizacionu sumu. Koeficijenti α i ρ se određuju na osnovu očekivanog trajanja i varijanse koji su dobijeni na osnovu histograma iz [4].

$$\alpha = \frac{E\{\tau\}}{VAR\{\tau\}} \tag{3}$$

$$\rho = \frac{E^2\{\tau\}}{VAR\{\tau\}} \tag{4}$$

Sl. 2 prikazuje rezultat primene gama raspodele na model histograma.



Sl. 2. Raspodela trajanja na osnovu podataka iz baze (histogram) i gama aproksimacija (kriva).

Ako se predpostavi da se korelacija između trajanja subfonema u okviru jednog fonema može zanemariti, vrednosti očekivanog trajanja (*E*) i varijanse (*VAR*) za fonem koji se sastoji od n substanja se dobijaju na osnovu jednačina (5) i (6). Na taj način gama raspodela pored dobre aproksimacije pruža i matematički alat koji se može upotrebiti za proračun odstupanja i penalisanje trajanja svakog fonema u okviru reči.

$$E[Ph] = E[Ph_0] + E[Ph_1] + \ldots + E[Ph_n] \qquad (5)$$

$$VAR[Ph] \approx VAR[Ph_0] + VAR[Ph_1] + VAR[Ph_2] + \ldots + VAR[Ph_n] \qquad (6)$$

Drugi pristup u modelovanju (MTZK) zasniva se na korelaciji trajanja izmedju pojedinih fonema u okviru jedne reči. Takođe na osnovu očekivane vrednosti (*Ed*) i varijanse (*VAR*) se prate relativna odstupanja prema jednačini (7):

$$f_{jk} = \frac{d_j - Ed_j}{VAR_j} - \frac{d_k - Ed_k}{VAR_k} \qquad (7)$$

gde f_{jk} predstavlja odnos relativnih odstupanja *j*-tog i *k*-tog fonema u reči.

Na ovaj način se grupna odstupanja trajanja u odnosu na neku statističku raspodelu tolerišu dok značajno odstupanje samo pojedinih fonema vodi ka negativnom skoru. Time se dozvoljavaju promene u načinu izgovora (brže ili sporije) ali se penališu slučajevi kada prepoznavač, isključivo na osnovu akustičkih obeležja, prepozna određenu reč samo na osnovu dobrog uklapanja jednog ili dva fonema (najčešće su to samoglasnici).

Ukupna pouzdanost na nivou reči se dobija kao suma akustičke pouzdanosti, koeficijenta energije i penalisanja trajanja.

VI. Rezultati

Fajlovi koji su korišćeni prilikom testiranja su nezavisni od fajlova koji su korišćeni za obuku ASR sistema. U prvoj grupi testova korišćeno je 25 zvučnih fajlova koji su dobijeni snimanjem preko javne telefonske mreže sa korisničkih servisa. Korišćena je gramatika u kojoj ne postoji ni jedna reč koja je izgovorena u zvučnim fajlovima. Cilj ovako postavljenog testa je simulacija situacije u kojoj se korisnik raspriča, umesto da da kratak odgovor.

Grafici na slikama 3 i 4 prikazuju modifikaciju izlaznih metrika prilikom primene metoda za validaciju. Očigledno je da je u ovom slučaju aktivirano penalisanje "raspričavanja" preko skorovanja energije pošto je najveći deo zvučnog fajla prepoznat kao govor u kome se ne nalaze reči definisane gramatikom ili šum (*int* je u okviru AlfaNum ASR-a simbol za šum i reči van gramatike).

Slični rezultati se dobijaju i pri aktiviranju skorovanja trajanja. Primenom ove metode se potiskuju izlazi iz prepoznavača koji su posledica dobrog akustičkog uklapanja samo dela fonema u reči (samoglasnika).

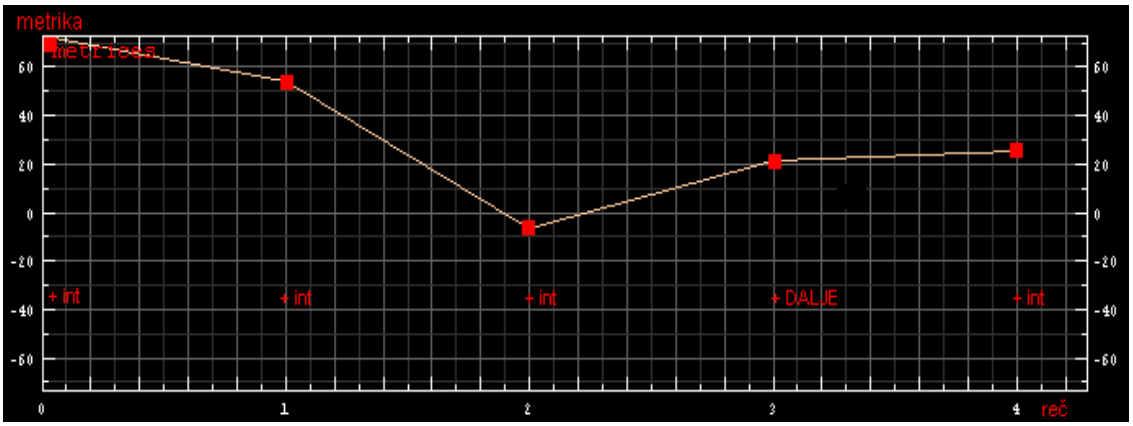
VII. Zaključak

Obe primenjene metode, praćenje energije i skorovanje trajanja preko gama raspodele, mogu se uspešno koristiti za validaciju krajnje ocene kvaliteta prepoznavanja i time značajno povećavati robusnost samog procesa.

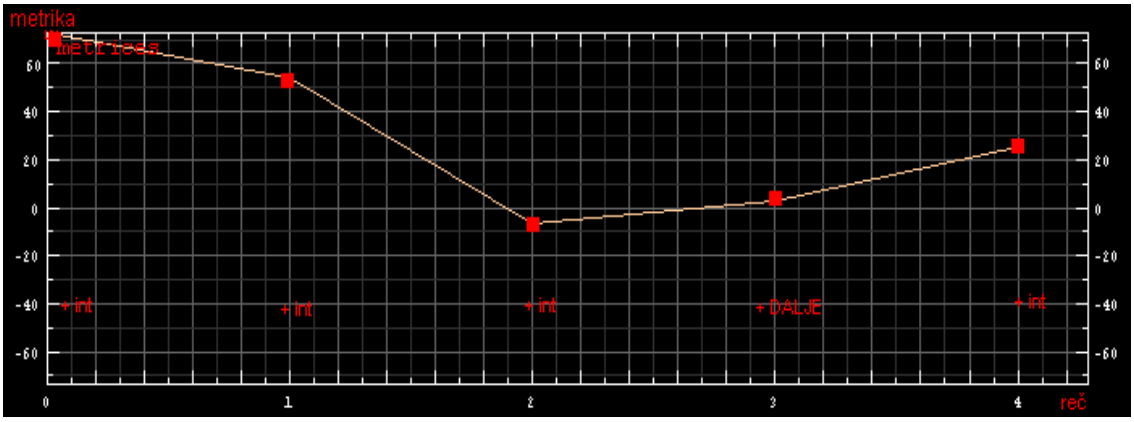
Izlaz iz Viterbijevog prepoznavača je metrika prepoznate reči tj. onog što se najbolje uklopilo sa sadržajem zvučnog fajla. Vrednost metrike je često visoka čak i pri pogrešnom prepoznavanju, tako da je neophodno svesti je na realnu meru.

Zahvaljujući primeni ovih metoda vršimo validaciju verodostojnosti na izlazu iz Viterbija što omogućava da se u krajnjoj aplikaciji postavi prag kao minimalna metrika za korektan rezultat prepoznavanja.

Sledeći korak u realizaciji ovog pristupa je implementacija skorovanja trajanja u vidu izmenjenog Viterbijevog algoritma.



Sl. 3. Izlaz iz ASR-a sa nekorektno prepoznatom reči DALJE bez primene postprocesing metoda.



Sl. 4. Izlaz iz ASR-a sa nekorektno prepoznatom reči DALJE sa primenom postprocesing metoda.

LITERATURA

[1] N. Đurić, D Pekar, Lj. Jovanov, "*Struktura srpske SpeechDat(E) govorne baze snimljene preko fiksne telefonske mreže*," *DOGS 2002*, Bečej 2002, pp 57-60.

[2] D. Pekar, R. Obradović, V. Delić, "*Programski paket AlfaNumCASR – sistem za prepoznavanje kontinualnog govora*," *DOGS 2002*, Bečej 2002, pp 49-56.

[3] D. Mišković, D. Pekar, N. Jakovljević, N. Vujnović, "*Vremensko poravnavanje govorne baze u toku obuke sistema za prepoznavanje govora*," *ETRAN 2006*, Beograd 2006, vol. II, pp 466-469.

[4] S. J. Park, M. W. Koo, C. S. Jhon, "*Context-dependent phoneme duration modeling with tree-based state tying*," *IEICE Trans. Inf & Syst.*, vol. E88-D, March 2005.

ABSTRACT

This paper describes a method for improving quality of recognition for continuous Automatic Speech Recognition sistem on Serbian language. Basic idea is to increase system robustness in speech applications (speech machines, call centres...) by applying validation methods on the final results. The first method relies on energy tracking of recognized words and the second one on modeling their duration. Both methods are considered as postprocessing methods which correct values resulted on the output of Viterby algorithm.

POSTPROCESSING METHODS FOR RECOGNITION VALIDATION OF ALFANUM ASR SISTEM

Dragiša Mišković, Mirjana Zindović, Darko Pekar