Modelovanje trajanja govornih segmenata u sintezi govora

Sandra Sovilj-Nikić

***Sadržaj* – S obzirom na značaj trajanja govornih segmenata sa perceptivnog stanovišta, specijalizovani modul za određivanje potrebnog trajanja predstavlja komponentu TTS (eng. Text-to-Speech) sistema od izuzetne važnosti za proizvodnju sintetizovanog govora visokog kvaliteta. Modelovanje trajanja govornih segmenata u različitim jezicima jeste predmet mnogobrojnih do sada sprovedenih istraživanja u kojima su primenjivane različite tehnike modelovanja.**

U ovom radu je pored kratkog prikaza različitih modela trajanja, dat detaljniji opis CART (eng. Classification and Regression Trees) metode koja će u daljem istraživanju biti primenjena za razvoj modela trajanja glasova u sintezi govora na srpskom jeziku.

***Ključne reči* – CART metoda, modelovanje trajanja, kontekstualni faktori**

I. UVOD

Među najznačajnija prozodijska obeležja, sa stanovišta govornih tehnologija i njihove primene, ubrajaju se osnovna učestanost, glasnost i trajanje govornih segmenata. Iako su mnoga do sada sprovedena istraživanja neosporno pokazala da kretanje osnovne učestanosti glasa, odnosno f_0 kriva, predstavlja najznačajnije prozodijsko obeležje sa perceptivnog stanovišta, takođe postoje mnoga istraživanja koja pokazuju da trajanja govornih segmenata imaju neznatno manje značajnu ulogu od f_0 krive za razumevanje poruke upućene slušaocu [1]. Stoga je mogućnost automatske procene prirodnog trajanja glasova od izuzetne važnosti za postizanje prirodnosti sintetizovanog govora.

U prirodnom govoru trajanja glasova su izuzetno zavisna od konteksta u kojem se određeni govorni segment nalazi, pri čemu je ta zavisnost veoma kompleksna i uključuje mnogobrojne faktore [2]. Stoga je za proizvodnju sintetizovanog govora visokog kvaliteta veoma bitno da u okviru TTS sistema postoji specijalizovani modul čiji je zadatak da modeluje trajanja govornih segmenata iz prirodnog govora uzimajući u obzir različite faktore koji utiču na trajanje, a imajući u vidu prirodu problema to su oni faktori koje je moguće izvući iz samog teksta.

U ovom radu dat je prikaz različitih modela trajanja govornih segmenata, kao i detaljniji opis CART (eng. *Classification and Regression Trees*) metode. Takođe, ukazano je na značaj izbora faktora koji u najvećoj meri utiču na trajanje govornih segmenata.

Sandra Sovilj-Nikić, stipendista-doktorant Ministarstva nauke, Fakultet tehničkih nauka, Novi Sad, Srbija
(tel:0642302938, e-mail:sandrasn@eunet.yu)

II. MODELI TRAJANJA

Modeli za predviđanje trajanja mogu se podeliti u dve grupe: modeli zasnovani na primeni pravila (eng. *rule-based models*) i korpusno orijentisani modeli (eng. *corpus-based models*).

Jedan od najpoznatijih modela za predviđanje trajanja primenom niza uzastopnih pravila, ujedno i najstariji model, razvio je Denis Klatt [3]. Njegov rad predstavlja osnovu za razvoj modela trajanja govornih segmenata za nekoliko svetskih jezika kao što su američki engleski, švedski, francuski i brazilski portugalski. Kod ovakvog tipa modela pretpostavlja se da svaki fonem poseduje određeno inherentno trajanje koje predstavlja jedno od distinktivnih svojstava fonema. Svakom fonemu inicijalno se dodeljuje inherentno trajanje koje se zatim modifikuje primenom niza uzastopnih pravila. Primenom određenog pravila trajanje datog govornog segmenta produžava se ili skraćuje za određeni procenat, pri čemu nakon skraćenja trajanje ne može biti manje od minimalnog trajanja. Kod ovakvog tipa modela neophodna je *lookup* tabela koja sadrži minimalno i inherentno trajanje svakog fonema.

Prilikom razvoja modela zasnovanih na primeni pravila neophodno je znanje stručnih lingvista za dati jezik, odnosno njihovo učestvovanje u sastavljanju određenih pravila. Pisanje pravila često može biti veoma naporan i vremenski zahtevan posao, a takođe veoma je teško formirati dovoljan broj pravila kojima bi bile obuhvaćene sve moguće situacije u nekom jeziku. Stoga, kod ovakvog tipa modela pojava izuzetaka najčešće predstavlja problem jer su pravila uglavnom takva da često dovode do prevelikog uopštavanja. Međutim, pored niza prethodno spomenutih nedostataka modela zasnovanih na primeni pravila oni poseduju i određene prednosti. Osnovna prednost ovakvih modela leži u činjenici da ne zahtevaju obiman govorni korpus, što je bilo od izuzetne važnosti u vreme njihovog nastanka kada računarski resursi za generisanje i analizu obimnih govornih korpusa nisu bili dostupni kao danas.

Međutim, razvojem računarske tehnologije korpusno orijentisani modeli postaju sve zastupljeniji. Korpusno orijentisani statistički modeli zahtevaju obiman korpus snimljenog govora jer se modelovanje trajanja vrši primenom neke od metoda automatskog učenja na obimnom govornom korpusu. U zavisnosti od metode automatskog učenja koja se primenjuje u svrhu

modelovanja trajanja van Santen [4] razlikuje tri tipa modela:

- linearni statistički modeli
- modeli dobijeni primenom neuralnih mreža
- modeli dobijeni primenom CART metode.

Van Santen, kao primer linearnog statističkog modela, navodi aditivni model koji je Kaiki razvio za japanski jezik. Kod ovakvog modela trajanje govornog segmenta u datom kontekstu dobija se sumiranjem niza parametara kojima se modeluje uticaj različitih kontekstualnih faktora (identitet fonema, fonetsko okruženje, tj. prethodni i naredni fonem, itd.) na trajanje govornog segmenta. Estimacija ovih parametara vrši se primenom neke od standardnih statističkih metoda [5]. Takođe, postoje i multiplikativni modeli kod kojih se umesto sumiranja parametara vrši njihovo množenje.

U svojim radovima van Santen predlaže model sume proizvoda koji predstavlja generalizaciju aditivnog i multiplikativnog modela. Van Santen kao osnovne prednosti predloženog modela ističe mogućnost opisivanja međusobnog uticaja faktora jednostavnim aritmetičkim operacijama sabiranja i množenja, kao i sposobnost modela za estimaciju parametara u slučaju pojave retkih vektora. Ovakav model primenjen je za modelovanje trajanja govornih segmenata u engleskom, nemačkom i japanskom jeziku.

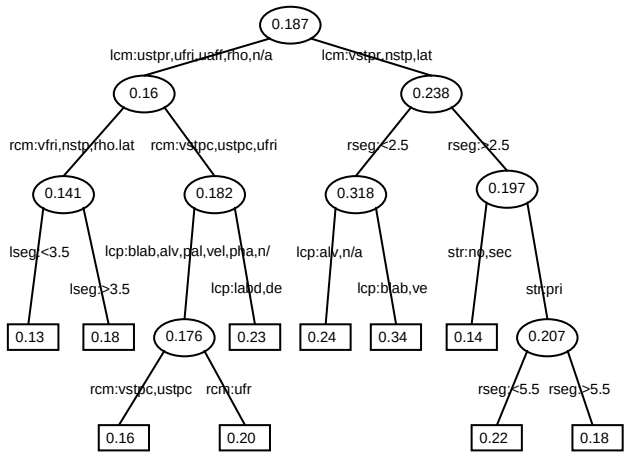
Druga mogućnost za obučavanje sistema u cilju sticanja određenih znanja koja će se kasnije koristiti za predviđanje trajanja govornih segmenata jeste primena neuralnih mreža. U svom radu Campbell prvi primenjuje neuralne mreže za predviđanje trajanja sloga u engleskom jeziku i predlaže modelovanje u dva koraka [6]. U prvom koraku on primenjuje neuralne mreže sa tri nivoa i propagacijom unazad (eng. *three-level back-propagation neural networks*) za predviđanje odstupanja trajanja pojedinačnog sloga od srednje vrednosti. U cilju pronalaženja faktora koji utiču na trajanje sloga Campbell primenjuje analizu kategorijskog faktora. Trajanje svakog segmenta u slogu određuje se u drugom koraku modelovanja rešavanjem sledeće jednačine po k:

$$\Delta = \sum_{i=1}^{n} \exp(\mu_i + k \cdot \sigma_i) \qquad (1)$$

gde je: Δ trajanje sloga određeno u prethodnom koraku
n broj segmenata u slogu
μ_i i σ_i srednja vrednost i standardna devijacija trajanja datog fonema

Nakon primene iterativnog postupka za rešavanje eksponencijalne jednačine (1) odgovarajućem govornom segmentu i dodeljuje se trajanje $\exp(\mu_i + k \cdot \sigma_i)$.

U treću grupu statističkih modela spadaju modeli zasnovani na primeni stabala odluke. Prvi takav model za predviđanje trajanja govornih segmenata u američkom engleskom jeziku razvio je Riley [7] koristeći CART tehniku. Primenom CART tehnike formira se binarno stablo koje u svakom čvoru sadrži da/ne pitanje o nekom obeležju, odnosno faktoru koji utiče na trajanje govornog segmenta. Predviđanje trajanja govornog segmenta vrši se prolaskom kroz stablo odluke, od korena do lista stabla, prolazeći kroz unutrašnje čvorove stabla onom putanjom koja se formira u zavisnosti od zadovoljenja određenog uslova o vrednostima obeležja u svakom od unutrašnjih čvorova. List stabla sadrži predviđenu vrednost trajanja datog govornog segmenta (slika 1 [7]).



Sl.1 Regresiono stablo za predviđanje trajanja glasa [æ]

Modelovanje trajanja govornih segmenata primenom CART metode realizovano je za mnoge jezike, među koje spadaju češki, grčki [8], litvanski, mandarinski, britanski engleski, vijetnamski, koreanski, indijski jezici hindi i telugu, turski [9].

III. CART METODA

CART metodu razvio je 1984. godine Leo Breiman [10]. Ova tehnika razvijena kao spoj statistike i veštačke inteligencije poseduje niz prednosti i kao takva danas predstavlja jednu od najčešće primenjivanih metoda za modelovanje trajanja govornih segmenata. Jedna od osnovnih prednosti CART algoritma jeste mogućnost validacije razvijenog modela, što se u praksi najčešće vrši procenom performansi modela na podacima koji nisu korišćeni u fazi obuke. Takođe, CART algoritam je relativno robustan u slučaju manjka podataka [10], omogućava jednostavnu interpretaciju i obradu dobijenih rezultata, statistički selektuje najznačajnija obeležja i omogućava kombinovanje kategorijskih (npr. identitet fonema) i numeričkih vrednosti (npr. trajanje fonema) obeležja.

Modelovanje trajanja govornih segmenata primenom CART tehnike podrazumeva upotrebu regresionog stabla za predviđanje trajanja datog govornog segmenta koji je u bazi predstavljen preko odgovarajućeg vektora obeležja. Formiranje pomenutog stabla sastoji se od nekoliko koraka: formiranje seta pitanja i izbor najboljeg pitanja na osnovu kojeg se vrši podela u datom čvoru; izbor kriterijuma za prestanak podele u nekom čvoru, odnosno proglašenje datog čvora za terminalni čvor (list) stabla; procena vrednosti u datom čvoru.

A. *Formiranje seta pitanja Q i izbor kriterijuma podele*

Neka je svaki od N podataka za obuku u bazi predstavljen preko odgovarajućeg vektora obeležja u formi:

$$X = (x_1, x_2, ..., x_M) \quad (2)$$

Za slučaj predviđanja trajanja x_1 može biti na primer način artikulacije prethodnog konsonanta, x_2 redni broj datog segmenta od kraja reči, itd. Treba zapaziti da elementi vektora obeležja mogu biti kategorijskog tipa, tj. mogu uzeti jednu od vrednosti iz konačnog neuređenog skupa (npr. način artikulacije konsonanta) ili numeričkog tipa, vrednost koju uzimaju je proizvoljan realan broj (npr. broj segmenata od kraja reči). U zavisnosti od tipa promenljive x_i formira se set pitanja Q:

1. Ako je nezavisna promenljiva x_i kategorijskog tipa, odnosno $x_i \in \{c_1, c_2, ..., c_K\} = C$ tada Q sadrži sva pitanja sledećeg oblika:
   $\{da\ li\ x_i \in A?\}$, $\forall A \subset C$
2. Ako je nezavisna promenljiva x_i numeričkog tipa, odnosno $-\infty < x_i < \infty$ tada Q sadrži sva pitanja sledećeg oblika:
   $\{da\ li\ x_i \le k?\}$, $\forall k$

Nakon formiranja celokupnog seta mogućih pitanja Q potrebno je pronaći najbolje pitanje za dati čvor, odnosno ono pitanje koje daje najbolju podelu podataka u datom čvoru.

Kod primene regresionih stabala, odnosno problema čije se rešavanje svodi na predviđanje kontinualne vrednosti najčešće primenjivani kriterijum podele jeste srednja kvadratna greška. Neka je Y stvarna vrednost trajanja nekog govornog segmenta u bazi predstavljenog preko vektora obeležja X, tada se ukupna greška predikcije u čvoru t definiše kao:

$$E(t) = \sum_{X \in t} \left| Y - \overline{d(X)} \right|^2 \quad (3)$$

gde je $\overline{d(X)}$ predviđena vrednost trajanja.

U skupu mogućih pitanja Q potrebno je pronaći pitanje koje najviše umanjuje kvadratnu grešku, odnosno pitanje q^* koje maksimizuje:

$$\Delta E_t(q) = E(t) - (E(l) + E(r)) \quad (4)$$

gde su l i r čvorovi koji se dobijaju nakon podele čvora t.

Za čvor t očekivana kvadratna greška definiše se kao:

$$V(t) = E\left\{ \sum_{X \in t} \left| Y - \overline{d(X)} \right|^2 \right\} = \frac{1}{N(t)} \sum_{X \in t} \left| Y - \overline{d(X)} \right|^2 \quad (5)$$

gde je $N(t)$ ukupan broj podataka koji se nalaze u datom čvoru t.

Može se zapaziti da $V(t)$ zapravo predstavlja varijansu procene trajanja ukoliko je $\overline{d(X)}$ srednja vrednost trajanja svih govornih segmenata koji se nalaze u čvoru t.

Ponderisana kvadratna greška $\overline{V(t)}$ u čvoru t definiše se kao:

$$\overline{V(t)} = V(t) \cdot P(t) = \left(\frac{1}{N(t)} \sum_{X \in t} \left| Y - \overline{d(X)} \right|^2 \right) \cdot P(t) \quad (6)$$

gde je $P(t)$ odnos broja podataka u čvoru t i ukupnog broja podataka.

Konačno, kriterijum podele u nekom čvoru t može se napisati kao:

$$\Delta \overline{V}_t(q) = \overline{V(t)} - (\overline{V(l)} - \overline{V(r)}) \quad (7)$$

pri čemu je potrebno pronaći pitanje q koje minimizuje varijansu predikcije nakon podele čvora t na čvorove l i r.

B. *Formiranje stabla*

Za dati skup pitanja Q i kriterijum podele $\overline{V}_t(q)$ proces formiranja stabla počinje od stabla koje sadrži samo koren, odnosno čvor u kome se nalazi svih N podataka za obuku iz baze. U svakom čvoru stabla za svaki element vektora obeležja x_i $i = 1, ..., M$ algoritam pronalazi najbolje pitanje iz datog skupa Q koristeći odgovarajući kriterijum podele. Nakon toga, od ukupno M izabranih pitanja bira se najbolje među njima, odnosno vrši se izbor najznačajnijeg obeležja u datom čvoru. Opisani postupak se ponavlja za svaki novodobijeni čvor sve dok ne bude zadovoljen jedan od sledećih uslova:

1. dalja podela u datom čvoru nije moguća jer svi podaci koji se nalaze u tom čvoru pripadaju istoj klasi
2. nakon podele maksimalno smanjenje varijanse je ispod unapred utvrđenog praga β, tj.:

$$\max_{q \in Q} \Delta \overline{V}_t(q) < \beta \quad (8)$$

3. broj podataka koji se nalaze u datom čvoru t je manji od unapred utvrđenog praga α.

Ukoliko je u nekom čvoru t zadovoljen jedan od prethodno navedenih uslova ne vrši se dalja podela u tom čvoru, odnosno čvor se proglašava za terminalni čvor stabla. Algoritam se završava kada je svaki čvor stabla proglašen za terminalni. Po završetku faze formiranja stabla zadovoljenjem nekih od prethodno navedenih uslova obično se dobija veliko stablo $T_{\max}$ koje može biti formirano striktno prema podacima koji su korišćeni u fazi obuke i takvo stablo nema sposobnost generalizacije, odnosno neće pokazati dobre performanse u slučaju primene nad podacima koji nisu korišćeni u fazi obuke. Stoga, potrebno je pronaći stablo optimalne veličine i izbeći *overfitting* podataka. U literaturi se navodi da je bilo niz pokušaja za prevazilaženje ovog problema među kojima se kao najbolje rešenje izdvaja Breiman-ov postupak koji se sastoji od nekoliko koraka: 1) formira se sekvenca podstabala $T_{\max} \supseteq ... \supseteq T_k \supseteq ... \supseteq T_K = t_1$ 2) za svako podstablo procenjuje se stopa greške 3) bira se stablo sa najmanjom stopom greške, odnosno stablo optimalne veličine [10]. Opisani postupak naziva se *cost-complexity pruning*. Prilikom formiranja sekvence podstabala koja se dobijaju odstranjivanjem pojedinih

grana parametar kompleksnosti α varira od 0 (za $T_{\max}$) do ∞ (za podstablo koje sadrži samo koren) tako da je zadovoljen uslov:

$$\min_T \left\{ \sigma^2(T) + \alpha \cdot |T| \right\} \quad (9)$$

gde je: $\sigma^2(T)$ varijansa greške predikcije za dato podstablo

$|T|$ broj terminalnih čvorova podstabla

U cilju procene stope greške podstabla se testiraju na podacima koji nisu korišćeni u fazi obuke. Procedura koja se najčešće primenjuje za procenu naziva se *cross-validation*. Naime, ukupna količina raspoloživih podataka podeli se na deset međusobno disjunktnih podskupova na kojima se vrši testiranje podstabala koja su formirana na osnovu preostalih 9/10 podataka. S obzirom da se postupak testiranja ponavlja deset puta za svako podstablo se izračunava prosečna varijansa. Ukoliko se varijansa dobijena na ovaj način posmatra kao funkcija veličine stabla, tada će za stablo određene veličine biti dostignut minimum varijanse i takvo stablo smatra se stablom optimalne veličine, jer dalje povećanje veličine stabla povećava varijansu.

IV. IZBOR NAJUTICAJNIJIH FAKTORA NA TRAJANJE

Trajanje fonema zavisi od mnogobrojnih faktora među koje se ubrajaju:

- identitet fonema;
- fonološki faktori: vrsta fonema, fonetsko okruženje, naglašenost sloga, položaj nenaglašenog sloga u naglasnoj celini (stopi), položaj fonema u reči, struktura sloga
- fiziološki faktori: minimalno trajanje određenog fonema;
- sintaksni faktori: položaj stope u okviru rečenice – ispred pauze, iza pauze, između dve pauze;
- veličina stope koju određuje broj slogova u stopi – da li je stopa jednosložna, dvosložna, itd.;
- brzina izgovaranja – normalno, brzo, sporo.

Uticaj nabrojanih faktora na trajanje fonema je u različitim jezicima različit. S obzirom na zavisnost trajanja glasova u prirodnom govoru od mnogobrojnih faktora, mnoga istraživanja fokusirana su na proučavanje uticaja ovih faktora na trajanje govornih segmenata u određenom jeziku jer izbor neadekvatnog ili nepotpunog skupa atributa može rezultovati velikom greškom prilikom procene trajanja. Stoga, identifikovanje najuticajnijih faktora predstavlja krucijalan korak u procesu modelovanja trajanja.

S obzirom na direktnu zavisnost najuticajnijih faktora na trajanje govornih segmenata od konkretnog jezika, izbor skupa atributa, odnosno elemenata vektora obeležja kojim se predstavlja svaki fonem u govornoj bazi u procesu modelovanja razlikuje se od jezika do jezika. U dosadašnjim istraživanjima utvrđeno je da fonetsko okruženje utiče na trajanje vokala u srpskom jeziku, kao i da vokali traju duže u zvučnom nego u bezvučnom okruženju [11]. Takođe je primećen uticaj položaja u reči na trajanje vokala, odnosno efekat produženja trajanja vokala na finalnoj poziciji u reči [11]. Veličina stope, koju određuje broj slogova u stopi, takođe predstavlja jedan od faktora koji utiču na trajanje vokala u srpskom jeziku jer primećeno je da vokali srpskog jezika, bez obzira da li su naglašeni ili ne, traju kraće ukoliko je broj slogova u stopi veći [11, 12]. Prethodno navedeni faktori biće uzeti u obzir u nastavku istraživanja, odnosno prilikom razvoja modela trajanja glasova u sintezi govora na srpskom jeziku kao fonološki faktori koji utiču na trajanje vokala srpskog jezika.

LITERATURA

[1] I. Bulyko, M. Ostendorf, P. Price, "On the Relative Importance of Different Prosodic Factors for Improving Speech Synthesis", *In Proc. of ICPhS*, vol. 1, pp. 81-84, 1999.

[2] J. P. H. van Santen, "Contextual Effects on Vowel Duration", *Speech Communication*, pp. 513-546, 1992.

[3] D. H. Klatt, "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence", *Journal of the Acoustical Society of America*, (59), pp. 1209-1221, 1976.

[4] J. P. H. van Santen, "Timing in Text-to-Speech Systems", *in Proc. of EUROSPEECH*, pp. 1397-1404, 1993.

[5] N. Kaiki, K. Takeda, Y. Sagisaka, "Statistical analysis for segmental duration rules in japanese speech synthesis", *in Proc. of ICSLP '90*, pp. 17-20, 1990.

[6] W. N. Campbell, "Multi-level Speech Timing Control", PhD dissertation, University of Sussex, 1992.

[7] M. Riley, "Tree-based modeling of segmental durations", *Talking Machines: Theories, Models and Designs*, Elsevier, pp. 265-273, 1992.

[8] A. Lazaridis, P. Zervas, N. Fakotakis, G. Kokkinakis, "A CART Approach for Duration Modeling of Greek Phonemes", *in Proc. of SPECOM*, pp. 287-292, 2007.

[9] O. Ozturk, "Modeling Phoneme Durations and Fundamental Frequency Contours in Turkish Speech", PhD dissertation, Middle East Technical University, 2005.

[10] L. Breiman, J. H. Fredman, R. A. Olshen, C. J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Belmont, CA., 1984.

[11] S. Sovilj-Nikić, "Trajanje vokala kao jedan od prozodijskih elemenata u sintezi govora na srpskom jeziku", Magistarski rad, Fakultet tehničkih nauka, Novi Sad, 2007.

[12] M. ,Marković, T. Milićev, "Uticaj veličine stope na trajanje vokala", *DOGS2008*, pp. 79-81, 2008.

ABSTRACT

Taking the significance of segmental duration for understanding the spoken text into consideration, specialized module whose task is to model the duration pattern of natural speech is very important for the production of high quality synthesized speech. Duration modeling in different languages is the subject of many realized research in which different modeling techniques have been used.

In this paper, besides the short review of different duration models, more detailed description of CART (Classification and Regression Trees) method is given. This method will be used for developing the segmental duration model for the speech synthesis in the Serbian language.

MODELING OF SEGMENTAL DURATION IN THE SPEECH SYNTHESIS

Sandra Sovilj-Nikić