ALFANUM SISTEM ZA SINTEZU GOVORA NA OSNOVU TEKSTA NA SRPSKOM JEZIKU

Milan Sečujski, Radovan Obradović, Darko Pekar, Ljubomir Jovanov, Vlado Delić

Fakultet tehničkih nauka, Univerzitet u Novom Sadu
{secujski, tlk_delic}@uns.ns.ac.yu

SADRŽAJ

U ovom radu prikazani su osnovni principi za projektovanje i realizaciju prvog sintetizatora govora na srpskom jeziku orijentisanog na povezivanje raznovrsnih govornih segmenata. Detaljno je opisan generator prozodijskih obeležja koji je korišćen, s osvrtom na određene specifičnosti srpskog jezika koje su uticale na njegovu realizaciju. U radu je, osim toga, dat i pregled kriterijuma za *on-line* selekciju odgovarajućih segmenata iz obimne govorne baze.

1. UVOD

S obzirom na njenu perspektivnost kao govorne tehnologije, sintezi govora na osnovu teksta se već nekoliko godina posvećuje velika pažnja u okviru projekta AlfaNum, na Fakultetu tehničkih nauka u Novom Sadu. Jedan od glavnih ciljeva tog projekta je realizacija kvalitetnog sintetizatora govora na srpskom jeziku, orijentisanog na sintezu govora povezivanjem govornih segmenata. Tekuća verzija AlfaNum sintetizatora govora zasnovana je na TD-PSOLA algoritmu, sprovedenom nad segmentima koji se u realnom vremenu izdvajaju iz unapred snimljene baze koja sadrži kontinualan govor, u skladu sa unapred određenim kriterijumima. U ovom radu detaljno je opisan AlfaNum sintetizator govora, uključujući opis generatora prozodije zasnovanog na odgovarajućem rečniku, i opis implementacije *on-line* pretrage govorne baze u potrazi za odgovarajućim segmentima uz pomoć kojih bi se mogla sintetizovati tražena govorna celina. Biće reči i o određenim postupcima predobrade govorne baze, koji su neophodan preduslov za implementaciju TD-PSOLA algoritma.

2. O NEKIM SPECIFIČNOSTIMA SRPSKOG JEZIKA

Postoje određene specifičnosti srpskog jezika koje u mnogome utiču na sintezu govora na osnovu teksta. Jedna od najistaknutijih je svakako njegova jednostavna fonetizacija. Izuzeci od pravila "jedno slovo-jedan glas" su veoma retki, pojavljuju se po pravilu na granicama između reči i lako ih je identifikovati. Određeni konsonanti poseduju svoje alofone, ali retki od njih imaju uticaja na razumljivost govora, tako da je zadatak fonetizacije na srpskom jeziku relativno trivijalan.

Zadatak generatora prozodije jeste da obezbedi da izgovorena govorna celina zvuči što prirodnije. Osim što je sintetizovan govor koji ima prirodnu intonaciju prijatniji za slušanje, on je i daleko jednostavniji za razumevanje, s obzirom da je slušaocu lakše da u govornom toku koji je u praksi neprekidan identifikuje granice između reči. Problem sintetizatora sa konstantnom f_0, jeste zapravo teškoća s kojom slušalac prati tok misli onoga ko emituje poruku, jer velike napore ulaže u to da otkrije gde se završava jedna, a gde počinje sledeća reč. U srpskom jeziku značaj ispravne intonacije rečenice je još izraženiji, s obzirom da položaj akcenta u okviru reči ponekad utiče na značenje same reči ili nosi poruku o različitoj morfološkoj kategoriji, tako da pogrešno akcentovana reč može uneti zabunu kod slušaoca. U prilog tvrdnji da ispravna intonacija rečenice u mnogome doprinosi pravilnom razumevanju govora, pogotovo u otežanim uslovima, pored brojnih istraživanja na tu temu sprovedenih u svetu, govori i eksperiment izvršen u okviru ovog rada, o kome će u daljem tekstu biti više reči.

Vokali srpskog jezika mogu biti naglašeni na četiri različita načina, što zapravo znači da postoji isto toliko tipičnih kombinacija kretanja osnovne učestanosti i trajanja naglašenog sloga. Još jedna od specifičnosti srpskog jezika jeste da je akcenat u velikom broju slučajeva određen kretanjem osnovne učestanosti glasa ne samo u naglašenom slogu, već u slogu posle njega. Četiri krive osnovne učestanosti karakteristične za četiri postojeća tipa akcenta prikazane su na slici 1, na primeru reči koje je izgovorila govornica čiji je glas korišćen pri snimanju govorne baze.

U zavisnosti od tipa akcenta, boja određenih vokala takođe varira, što se naročito može osetiti u severnijim krajevima govornog područja srpskog jezika, pri čemu je za kratke akcente po pravilu vezana otvorenija, a za duge zatvorenija boja vokala. Modifikacija boje vokala zahteva korišćenje složenijih algoritama parametarske sinteze

govora, i da bi se to izbeglo, pojedine klase vokala s različitom bojom su u okviru realizovanog sintetizatora od početka tretirane kao sasvim različiti vokali, što znači da nije dozvoljeno koristiti jedan tip vokala ukoliko u okviru govorne celine treba sintetizovati onaj drugi. O ovome je trebalo voditi računa kako prilikom labeliranja govorne baze, tako i prilikom samog odabira segmenata za sintezu.

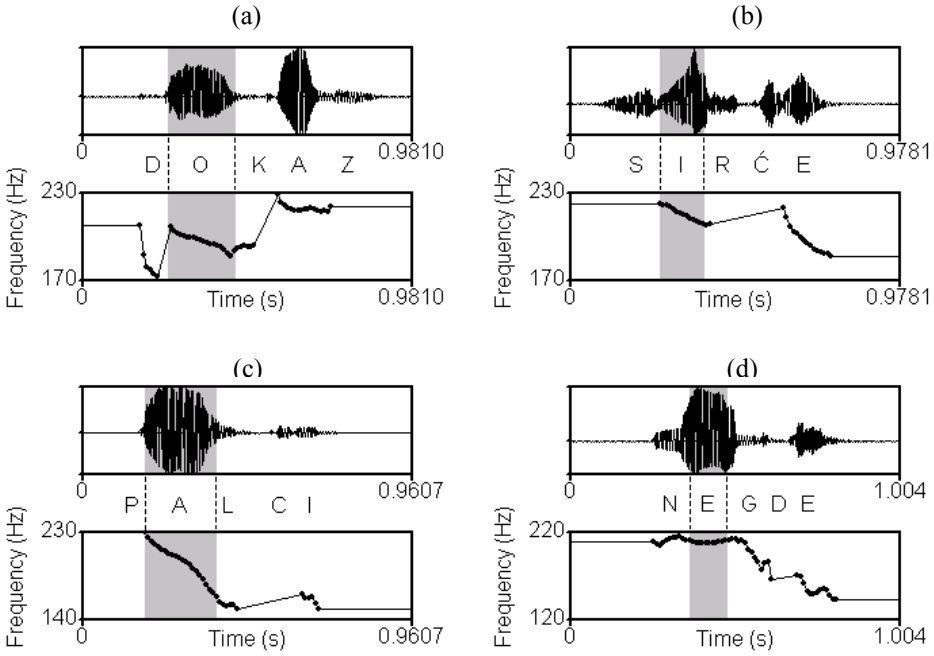
3. GOVORNA BAZA

3.1. Sadržaj

Govorna baza snimljena za potrebe ovog TTS sistema obuhvata nešto manje od dva sata kontinualnog govora, pri čemu je korišćen ženski glas. Imajući na umu moguće primene ovakvog sistema, kao i činjenicu da spajanje dužih govornih segmenata daje razumljiviji govor nego u slučaju kraćih, odlučeno je da govorna baza obuhvati fraze kao što su: lična imena i prezimena, adrese, nazivi preduzeća, gradova i država, novčani iznosi, nazivi valuta, vremenske fraze, izveštaji o vremenskoj prognozi i horoskopski izveštaji, tipične rečenice korišćene u interaktivnim govornim automatima, *e-mail* porukama i slično. Nad govornom bazom su obavljene određene operacije neophodne za implementaciju sinteze govora, kao što su labeliranje i postavljanje tzv. *pitch*-markera.

3.2 Labeliranje i postavljanje *pitch*-markera

Labeliranje baze podrazumeva postavljanje granica između jedinica koje pripadaju unapred definisanom skupu jedinica kao što su glasovi. U suštini, ono se svodi na smeštanje informacija o tim jedinicama, kao što su početni i završni trenuci, u posebnu bazu podataka. Labeliranje AlfaNum govorne baze zasnivalo se po pravilu na glasovima, mada je u nekim slučajevima usvojena bolja alternativa, zahvaljujući određenim fonetskim svojstvima pojedinih grupa fonema, kao i određenim specifičnostima srpskog jezika. Primera radi, neke klase fonema, kao što su plozivi i afrikati, bile su obeležavane kao parovi polufonema, koje su, u slučaju ploziva, činile okluzija i eksplozija, a u slučaju afrikata okluzija i frikcija. Vokali sa izraženom otvorenom ili zatvorenom bojom bili su tretirani kao posebni vokali, kao što je već pomenuto. S obzirom na način izgovaranja glasa R u srpskom jeziku (periodičan niz uzastopnih okluzija i eksplozija izazvanih oscilacijama jezika naspram tvrdog nepca), sve te okluzije i eksplozije bile su evidentirane. Osim postavljanja granica između glasova, bio je identifikovan i deo svakog fonema fonološki najnezavisniji od okoline i najmanje ugrožen koartikulacijom, da bi se naknadne modifikacije prozodije učinile što manje primetnim. S obzirom da govorna baza sadrži kontinualan govor sa svim njegovim nedostacima kao što su oštećeni ili sasvim progutani glasovi, svi



Slika 1. Tipovi akcenata u srpskom jeziku i odgovarajuće f_0 konture: (a) dugouzlazni, (b) kratkouzlazni, (c) dugosilazni, (d) kratkosilazni, kako ih je izgovorila govornica čiji je glas korišćen pri snimanju baze

takvi slučajevi bili su evidentirani, zajedno sa podatkom o stepenu oštećenja. Ovo je urađeno da bi se izbeglo korišćenje tih glasova u kontekstu u kom su tipično dobro artikulisani, kao na primer u okviru naglašenih slogova. Ovo se u podjednakoj meri odnosi i na vokale i na konsonante.

Labeliranje je izvršeno automatski, uz korišćenje AlfaNum prepoznavača kontinualnog govora [7], a provera obavljenog posla je izvršena ručno, na osnovu izgleda signala, njegovog spektrograma i auditorne percepcije. Pri tome je korišćen program SpeechVis, razvijen u okviru projekta AlfaNum [8].

Implementacija TD-PSOLA algoritma zahteva da u bazi prethodno budu postavljeni tzv. *pitch*-markeri, odnosno, da se u okviru glasova detektuju pozicije najpogodnije za centriranje preklopljenih prozorskih funkcija, u slučaju da se koristi algoritam TTS sinteze koji zahteva postavljanje frejmova usklađeno sa trenutnom osnovnom učestanošću govora. Tako je za vreme zvučnih frejmova postavljen jedan marker po periodi, a za vreme bezvučnih, markeri su postavljani u skladu sa prosečnom osnovnom učestanošću u čitavoj bazi. Da bi se izbegli čujni efekti prilikom naglih prelaza iz zvučnog u bezvučni segment i obrnuto, ova strategija je donekle modifikovana u blizini tih prelaza, a brzina kojom se mogu menjati uzastopna rastojanja između markera je značajno smanjena.

Postupak postavljanja markera u okviru zvučnih frejmova je sproveden u dve faze. Prvo je izvršena preliminarna procena f_0 krivih korišćenjem AMDF metode. Svaki segment je prethodno filtriran NF filtrom sa f_c=900 Hz. U sledećem koraku lociran je frejm sa najvišim stepenom zvučnosti i u okviru njega globalni maksimum signala. Tu je postavljen inicijalni *pitch*-marker. Nakon toga, potraga za ostalim markerima sprovedena je na osnovu preliminarnih procena f_0, čime se postiglo da frejmovi centrirani oko tih pozicija obuhvate najveći deo energije signala, tako da stepen izobličenja signala izazvanog upotrebom prozorskih funkcija bude što niži.

Ovakva procedura ipak nije bila u potpunosti nepogrešiva, jer NF filtriranje ponekad može da izmeni vršne vrednosti signala do te mere da maksimumi u okviru određenih segmenata mogu da se ne podudare sa maksimumima u okviru nekih drugih. Ovo je moguće u slučaju onih zvučnih glasova čiji talasni oblici poseduju dva maksimuma približno iste veličine. Postoji još jedan razlog zbog kog može doći do grešaka. Na mestu granica između pojedinih glasova ponekad dolazi do nepravilnosti u radu glotisa, što izaziva odstupanja u položajima glotalnih impulsa. Ako se usvoji strategija postavljanja *pitch*-markera zasnovana na glasovima, to jest, ako se potraga za njima vrši u okviru svakog glasa posebno, a ne u okviru segmenata koji sadrže više od jednog glasa, većina ovakvih grešaka je automatski eliminisana.

Baza koja obuhvata informacije o položajima *pitch*-markera pogodna je za sintezu govora bilo kojom tehnikom orijentisanom na povezivanje segmenata, čak i onom koja ne zahteva eksplicitno poznavanje položaja *pitch*-markera, zato što se f_0 krive mogu jednostavno odrediti na osnovu njih.

4. GENERISANJE PROZODIJE

U okviru AlfaNum TTS sistema, akustički parametri, među kojima najvažnije mesto zauzima f_0 kriva, bili su izračunati u dva koraka. Prvo je rečenica analizirana reč po reč, da bi se došlo do informacija o akcenatskoj konfiguraciji svake reči pojedinačno, kao i osnovnih gramatičkih informacija. One su neophodne za korektnu sintezu prozodije, jer naglasak reči u srpskom jeziku može da varira s promenom morfološke kategorije reči, a ponekad te varijacije mogu da označe da je u pitanju sasvim druga leksička reč. U nekim slučajevima čak ni sintaksna analiza nije dovoljna, i nekoliko interpretacija iste rečenice, od kojih sve zadovoljavaju sintaksu, može biti naglašeno na različite načine, i imati različito značenje. Ako sintaksnu i semantičku analizu vrši čovek, kod njega obično nema zabune, jer nedoumicu može da otkloni oslanjajući se na kontekst.

Imajući sve ovo na umu, dolazi se do osnovne strukture modula za generisanje prozodije, koga čine modul za akcentuaciju rečenice i modul za generisanje f_0 krive na osnovu poznate akcentuacije, kako je prikazano na slici 2. U tekućoj verziji programa radi pojednostavljenja nije uzeta u obzir promena relativne glasnosti pojedinih slogova u zavisnosti od akcentuacije, jer se pošlo od pretpostavke da ona manje utiče na razumljivost od intonacije rečenice, odnosno f_0 krive. Ova pretpostavka potvrđena je i u literaturi, a u okviru ovog rada biće opisan i eksperiment kojim je potvrđen značaj f_0 krive kao najvažnijeg elementa rečenične prozodije.

Akcentuacija rečenice obavlja se u nekoliko koraka. Prvo se za svaku reč u rečenici pronađu sve mogućnosti za njenu akcenatsku konfiguraciju (vodeći računa o vrsti reči koja bi u pojedinim slučajevima bila u pitanju, kao i o vrednostima njenih morfoloških kategorija). Na osnovu toga se formiraju hipoteze o akcentuaciji čitave rečenice, kombinovanjem pojedinačnih hipoteza za svaku reč. Nakon toga pristupa se analizi konteksta, odnosno, za svaku hipotezu se proverava u kolikoj meri se slažu morfološke kategorije susednih reči, i

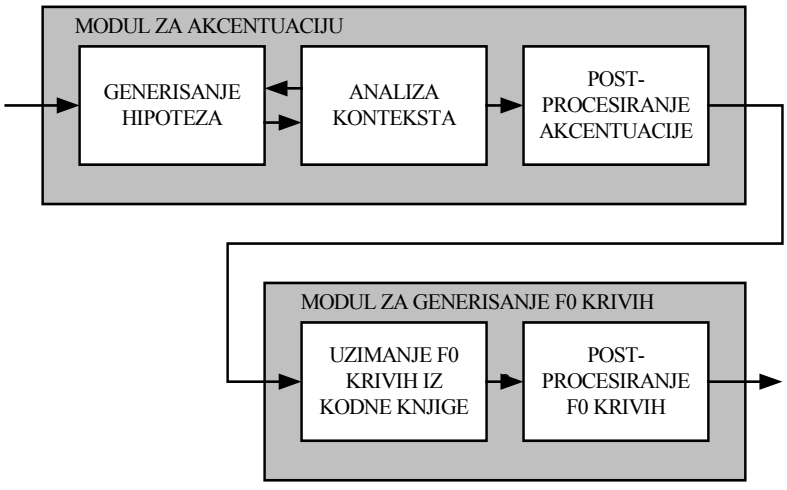
na osnovu toga se hipotezama dodeljuju odgovarajući bodovi. Međutim, ne kreiraju se odmah sve moguće hipoteze, a da se tek nakon toga pristupi bodovanju, jer bi u slučaju predugih rečenica broj hipoteza mogao neograničeno narasti. Hipoteze se kreiraju reč po reč, pri čemu se po dodavanju nove reči u hipotezu odmah vrši i bodovanje. Na taj način se postiže da se proširuju samo hipoteze koje su ostvarile relativno visok parcijalni skor, dok se one sa niskim parcijalnim skorom odbacuju pre nego što se do kraja formiraju, pa se na taj način ukupan broj formiranih hipoteza može održati u određenim granicama.

Umesto opisane analize konteksta, koja se zasniva na heuristikama dobijenim na osnovu proučavanja morfologije reči i sintakse rečenica u srpskom jeziku, efikasnije i pouzdanije rešenje bila bi kompletna sintaksna analiza rečenice zasnovana na formalnim gramatikama, ali je u ovoj fazi projekta usvojeno daleko jednostavnije rešenje, koje nije i mnogo manje pouzdano.

Po nalaženju najbolje hipoteze potrebno je nad njom izvršiti i određeno post-procesiranje akcentuacije, zahvaljujući činjenici da u srpskom jeziku naglasak u nekim situacijama može preći na reč koja je inače nenaglašena, te dolazi do tzv. *kliza* (npr. *spâvam* → *nè spāvam*). Drugim rečima, akcentuacija rečenice nije u opštem slučaju potpuno određena akcentuacijom svih reči u njoj. Zadatak modula za post-procesiranje akcentuacije je upravo u tome da na osnovu određenih heuristika detektuje takve slučajeve i da na odgovarajući način modifikuje akcentuaciju.

Rad modula za generisanje prozodije biće ukratko prikazan na primeru rečenice *„Sam idem iz knjižare“*. Prva faza je određivanje akcentuacije svake reči pojedinačno. U toj fazi se saznaje da *sam* može biti pomoćni glagol u prezentu prvog lica jednine, i u tom slučaju je nenaglašen (enklitički) – [*sām*], a može biti i pridev u nominativu ili akuzativu jednine muškog roda, tada ima dugosilazni akcenat – [*sâm*]. *Idem* je neprelazni glagol u prezentu prvog lica jednine, i ima kratkouzlazni akcenat – [*ìdem*], a *iz* je predlog, nepromenljiv i nenaglašen (proklitički) – [*ĭz*], koji zahteva imenički konstituent u genitivu. *Knjižare* može biti genitiv jednine ili nominativ, akuzativ ili vokativ množine imenice *knjižara*, i tada ima kratkouzlazni akcenat na prvom slogu – [*knjìžare*], ali može biti i akuzativ množine imenice *knjižar*, i tada ima dugouzlazni akcenat na drugom slogu – [*knjižáre*]. Zadatak modula za analizu konteksta jeste da odredi koje su od ovih mogućnosti prave.

U sledećoj fazi prelazi se na kreiranje hipoteza, koje čine strukturu sličnu trelisu. Prvi korak je ubacivanje svih mogućnosti za prvu reč u trelis. Za tu reč postoje, kao što je navedeno, tri mogućnosti, ali pošto rečenica praktično nikad ne počinje enklitikom, hipoteza koja je pošla od pretpostavke da je u pitanju pomoćni glagol biće u startu daleko lošije ocenjena od ostalih, pa će, u slučaju da broj aktuelnih hipoteza u nekom trenutku premaši određenu granicu, takve hipoteze s mnogo većom verovatnoćom biti odbacivane. U svakom narednom koraku u trelis se ubacuje nova reč i ocenjuje se njeno slaganje sa prethodnim rečima. Primera radi, za poslednju reč *knjižare* ima pet mogućnosti, ali kad bude ubačena u trelis, biće utvrđeno njeno slaganje sa prethodnom rečju, a to je bio predlog *iz* koji zahteva genitiv. Samo jedna od pet mogućnosti za poslednju reč je imenica u genitivu, tako da će hipoteza koja obuhvata tu mogućnost dobiti mnogo više bodova od ostalih. Na ovaj način u većini slučajeva samo jedna od hipoteza će se izdvojiti kao najbolja. U ovom primeru još uvek ostaje dilema da li je prva reč pridev u nominativu ili akuzativu, ali s obzirom na sintaksnu strukturu rečenice može se zaključiti da je u pitanju aktuelni



Slika 2.Izgled modula za generisanje prozodije

kvalifikativ koji se odnosi na neiskazani subjekat, pa je, prema tome u nominativu, a ne akuzativu (jer pravog objekta nema, pošto je glagol neprelazni). Olakšavajuća okolnost jeste da je za potrebe određivanja ispravne akcentuacije rečenice nekad (kao u ovom slučaju) dovoljna sasvim pojednostavljena sintaksna analiza, jer i da smo se greškom opredelili za akuzativ, intonacija rečenice bi bila ista, jer je i akuzativ naglašen na isti način. Bitno je odlučiti se za pravu varijantu samo u slučajevima kada se akcenatske konfiguracije razlikuju.

Kada je nađena najviše ocenjena hipoteza, a ta je u ovom slučaju [*sâm ìdem ĭz_knjìžare*], ona se pretražuje u potrazi za kombinacijama reči u kojima nastupaju klize (npr. *ne* + glagol sa silaznim akcentom na prvom slogu), a kako takvih u ovom slučaju nema, akcentuacija rečenice je završena, i utvrđeno je da se rečenica sastoji od tri naglasne celine (stope), koje imaju akcenatske konfiguracije [ˆ], [\ ˇ] i [ˇ \ ˇ ˇ].

Po završenoj akcentuaciji, uz pomoć odgovarajuće kodne knjige u kojoj su smešteni oblici f_0 krivih za pojedine stope, kreira se i f_0 kriva čitave rečenice. Resursi potrebni za generisanje prozodije na opisan način jesu, dakle, akcenatsko-morfološki rečnik, baza gramatičkih pravila i kodna knjiga f_0 krivih.

4.1. Rečnik

Akcenatsko-morfološki rečnik korišćen u okviru AlfaNum sistema za sintezu govora detaljno je opisan u [6], a dat je osvrt i na principe njegove koncepcije i realizacije. Ovaj rečnik pokriva preko milion reči, uključujući tu i izvedene oblike reči kao posebne unose, mada se oni u pojedinim slučajevima ne nalaze fizički u rečniku, ali se preko unapred definisanih pravila mogu izvesti iz postojećih unosa. Za reči koje ne postoje u rečniku definisane su posebne strategije akcentovanja, detaljnije opisane u [6].

4.2. Kodna knjiga f_0 krivih

Koristeći informacije iz rečnika, sistem može da rekonstruiše određenu akcenatsku konfiguraciju grupe reči koje formiraju naglasnu celinu (stopu). U ovoj fazi projekta, svakoj naglasnoj celini dodeljeno je nekoliko f_0 krivih dobijenih empirijskim putem, pri čemu se odabir konkretne krive vrši na osnovu položaja naglasne celine u rečenici (početak, neutralni položaj, ispred zareza, kraj), a rezultantna f_0 kriva dobija se spajanjem odabranih krivih i ublažavanjem prelaza između njih, kako bi se umanjili čujni diskontinuiteti visine glasa. Dobijena f_0 kriva je naknadno iskošena ka kraju rečenice, a varijacije f_0 su pred kraj rečenice umanjene, kako bi se, makar grubo, uzela u obzir tendencija govornika da opušta vokalni aparat kako se približava kraju govorne celine [1, 3]. Ovakav metod za sada ne vodi računa o sintaksnoj strukturi rečenice i njenom uticaju na prozodiju, već se oslanja isključivo na znake interpunkcije.

Treba ipak napomenuti da opisano rešenje generatora prozodije ne otklanja sve probleme, jer ne obuhvata kompletnu sintaksnu analizu, niti razrešava slučajeve sintaksnih dvosmislenosti, gde se mora pribeći semantičkoj analizi, ma koliko primitivna ona bila.

4.3. Eksperimentalna provera uticaja pravilne intonacije na razumljivost govora

U okviru ovog rada izveden je eksperiment čiji je cilj bio da pokaže koliki je uticaj ispravne intonacije rečenice na razumljivost govora, odnosno, da pokaže da se slušalac u velikoj meri oslanja na intonaciju rečenice kada u otežanim uslovima slušanja treba da rekonstruiše sintaksno i semantički ispravnu rečenicu koju je čuo.

U ovu svrhu uz pomoć AlfaNum sintetizatora govora kreirano je 12 sintaksno i semantički ispravnih rečenica, čija dužina nije prelazila 8 reči. Ove rečenice bile su namenjene prepoznavanju od strane slušalaca. Međutim, intonacija ovih rečenica nije bila ista u svim slučajevima. Svaka od ovih rečenica bila je kreirana u tri varijante. Prva varijanta rečenice bila je okarakterisana potpuno ravnom intonacijom, svojstvenom sistemima za sintezu govora slabijeg kvaliteta. Druga varijanta rečenice je imala prozodiju generisanu na osnovu pogrešne akcentuacije rečenice, kreirane ručno. Treća varijanta imala je ispravnu intonaciju.

S obzirom da je razumljivost sve tri varijante rečenica u normalnim uslovima bila vrlo visoka, eksperiment je sproveden u otežanim uslovima. U snimke sintetizovanog govora dodati su Gausov i impulsni šum, a prepoznavanje je vršeno u uslovima jakog ambijentalnog šuma.

Eksperiment je obuhvatio 12 slušalaca, i svakom od njih su, različitim redosledom, emitovane različite kombinacije od po četiri rečenice sa ravnom intonacijom, četiri rečenice sa pogrešnom intonacijom i četiri rečenice sa ispravnom intonacijom. Od svakog slušaoca zahtevalo se da ponovi rečenicu koju je čuo, a u slučaju da je nije razumeo, emitovana je ponovo. Za svakog slušaoca i svaku rečenicu evidentirano je da li je prepoznata iz prvog pokušaja, posle ponovljenog emitovanja, ili uopšte nije prepoznata. Svaki od slušalaca bio je unapred upoznat sa činjenicom da su rečenice koje će čuti sintetizovane, a ne izgovorene od strane

čoveka, i svakom od slušalaca srpski je maternji jezik.

Rezultati eksperimenta dati su na slici 3, i iz njih se jasno može videti doprinos ispravne intonacije razumljivosti govora. Dok su rečenice sa neutralnom (ravnom) intonacijom iz prvog pokušaja prepoznavane u 52% slučajeva, rečenice sa ispravnom intonacijom su iz prvog pokušaja prepoznate u čak 83% slučajeva, a ostale su neprepoznate posle drugog emitovanja u svega 8% slučajeva.

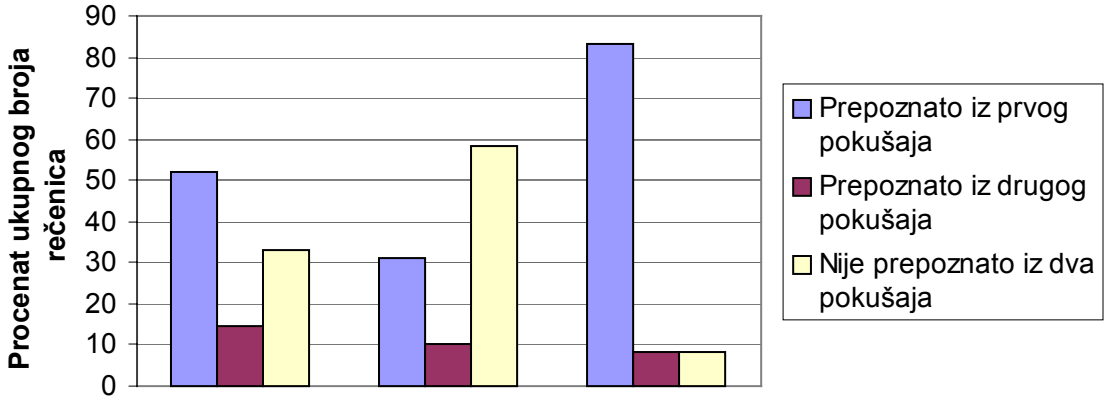
Još jedan zanimljiv rezultat eksperimenta jeste da su slušaoci znatno lošije prepoznavali rečenice sa pogrešnom intonacijom čak i od rečenica sa neutralnom (ravnom) intonacijom. Naime, u odnosu na 52% rečenica s neutralnom intonacijom prepoznatih iz prvog pokušaja, procenat rečenica s pogrešnom intonacijom prepoznatih iz prvog pokušaja je znatno niži, i iznosi svega 31%. Ovaj rezultat samo ukazuje na to koliko se slušalac oslanja na intonaciju rečenice. Naime, u slučaju da je intonacija rečenice neutralna, slušalac je svestan da ne može da se oslanja na intonaciju i na to što zna „kako koja reč zvuči“, već maksimalnu pažnju usredsređuje na prepoznavanje fonema i na njihovo kombinovanje u smislene reči koje čine smislenu rečenicu. U slučaju da je intonacija pogrešna, ali da ipak postoje osetne varijacije f_0 krive i trajanja pojedinih vokala, slušalac stiče pogrešan utisak da je rečenica pravilno intonirana, i pokušava da u njoj pronađe reči „po zvučanju“, ali ne uspeva u tome, i zbog toga postiže lošije rezultate nego da varijacija f_0 krive nije ni bilo.

Iz rezultata opisanog eksperimenta mogao bi se izvući još jedan zaključak. Naime, „pravilna intonacija“ svake od rečenica korišćenih u ovom eksperimentu nije svaki put preuzeta iz iste rečenice koje je izgovorio čovek, već je zapravo dobijena korišćenjem f_0 krivih iz kodne knjige, koje su bile unapred izdvojene iz pogodnih akcenatskih konteksta i naknadno ručno modifikovane. Samim tim, ovaj eksperiment ujedno predstavlja i potvrdu da su f_0 krive u kodnoj knjizi adekvatne.

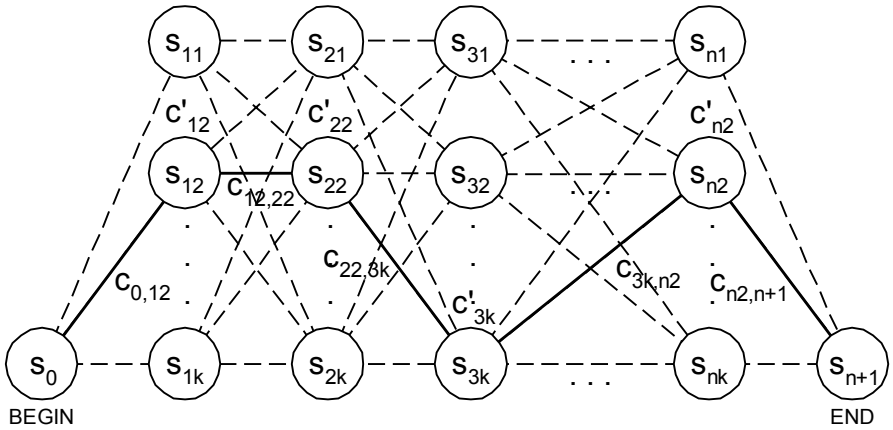
5. ON-LINE SELEKCIJA SEGMENATA

Polovine fonema smatraju se najkraćim jedinicama ispod kojih nije moguća dalja segmentacija. Međutim, poželjno je koristiti što duže govorne segmente, kako bi se očuvala razumljivost. U skladu sa razlikama između raspoloživih i zahtevanih vrednosti unapred utvrđenih parametara, svakom govornom segmentu koji bi se mogao izvući iz baze i upotrebiti za sintezu dodeljena je *cena korišćenja*, a u skladu s razlikama na granicama dva segmenta, svakom paru segmenata koji bi se mogao spajati dodeljena je *cena spoja* [2]. Cena korišćenja je mera odstupanja vrednosti određenih prozodijskih parametara, koji obuhvataju trajanje, f_0, energiju i odmerke obvojnice spektra, između ciljnog i raspoloživog segmenta. Cena spoja je mera odstupanja istih parametara na granicama između glasova. Različite grupe fonema tretirane su pri tome na različit način. Neke grupe, kao što su bezvučni plozivi, pogodnije su za segmentaciju od ostalih, pa imaju niže cene spoja.

Pri izdvajanju segmenata iz baze vodilo se računa i o stepenu oštećenja pojedinih glasova. Delimično i potpuno oštećeni glasovi, koji su prilikom labeliranja posebno označeni, ne mogu se koristiti za sintezu, osim u slučajevima kada je oštećenje relativno malo, a potrebno je sintetizovati glas (vokal ili konsonant) u izrazito nenaglašenom delu naglasne celine, tipično na njenom kraju. Naime, tu se, zbog lencije govornika, odnosno njegove težnje da opusti govorni trakt pre kraja govorne celine, pošto je preneo njen naglašen deo, artikulacija glasova u određenoj meri



Slika 3. Rezultati eksperimenta: (a) Rečenice sa neutralnom intonacijom, (b) Rečenice sa pogrešnom intonacijom, (c) Rečenice sa ispravnom intonacijom.



Slika 4. Traženje puta kroz trelis koji predstavlja rečenicu

menja. Konsonanti su slabije artikulisani, a vokalima se boja modifikuje u smislu da se i prvi i drugi formant u spektru pomeraju ka srednjim učestanostima.

Zadatak sintetizatora je da pronađe najbolji put kroz trelis koji predstavlja govornu celinu koju treba sintetizovati. To je onaj put kojim se akumuliše najmanja ukupna cena, koja se računa kao zbir cena korišćenja upotrebljenih segmenata i cena spojeva između njih. Odabirom puta, odabrani su i konkretni segmenti koji će biti upotrebljeni, kao što je prikazano na slici 4.

6. ZAKLJUČAK

U ovom radu detaljno je opisana koncepcija kvalitetnog TTS sistema za srpski jezik. Prikazani su principi generisanja prozodije, u svetlu određenih specifičnosti srpskog jezika. Objašnjen je način selekcije govornih segmenata iz govorne baze u realnom vremenu, kojim je postignuta relativno visoka razumljivost i prirodnost sintetizovanog govora. Nakon nekoliko pokušaja realizacije TTS-a na srpskom jeziku, koji su uglavnom bili orijentisani na spajanje difona i nisu uzimali prozodiju u obzir, ovo je prvi kompletan TTS na srpskom jeziku čiji je kvalitet dovoljno visok da može računati na širu primenu.

LITERATURA

[1] T. Dutoit: *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, Dordrecht/ Boston/London, 1997.

[2] M. Beutnagel, M. Mohri, M. Riley: Rapid Unit Selection from a Large Speech Corpus for Concatenative Speech Synthesis, Proceedings of *EUROSPEECH '99*, pp.607-610. Budapest, Hungary, 1999.

[3] I. Lehiste, P. Ivić: Word and Sentence Prosody in Serbocroatian, The Massachusetts Institute of Technology, 1986.

[4] V. Delić, S. Krčo, D. Glavatović: Basic Elements for ASR and TTS in Serbian Language, *DOGS*, pp. 32-37, Fruška Gora, 1998.

[5] M. Sečujski: Text-to-Speech with Respect to Serbian Language, graduation thesis, School of Engineering, Novi Sad, 1999.

[6] M. Sečujski: Akcenatski rečnik srpskog jezika namenjen sintezi govora na osnovu teksta, *DOGS*, Bečej, 2002.

[7] D. Pekar, R. Obradović, V. Delić: AlfaNum System for Continuous Speech Recognition, *TSD 2002*, Brno, Czech Republic, 2002.

[8] R. Obradović, D. Pekar: C++ Library for Signal Processing – SLIB, *DOGS,* Novi Sad, 2000.

ABSTRACT

This paper presents some basic criteria for conception of a concatenative TTS synthesizer in Serbian language, using variable length speech segments, and gives its detailed description. The paper describes the prosody generator which was used, and reflects upon several peculiarities of Serbian language which led to its adoption. The paper also describes the method of on-line selection of appropriate segments from a large speech corpus.