UTICAJ KARAKTERISTIKA GLASA NA KVALITET GOVORA SINTETIZOVANOG NA OSNOVU TEKSTA

Milan Sečujski, Darko Pekar, Vlado Delić
Fakultet tehničkih nauka, Novi Sad

Sadržaj – *U ovom radu je diskutovana problematika snimanja govorne baze za potrebe TTS sistema sa on-line selekcijom segmenata. Naročito se obratila pažnja na izbor govornika(-ce), a date su i uporedne karakteristike dve snimljene govornice.*

1. UVOD

Sinteza govora na osnovu teksta već vekovima predstavlja predmet pažnje velikog broja istraživača. Radi se o veoma složenom, multidisciplinarnom problemu, čije rešavanje posebno otežava činjenica da je govor u svim svojim aspektima još uvek nedovoljno istražen. Uz to, za rešavanje konkretnog problema sinteze govora na određenom jeziku neophodno je na odgovarajući način uključiti i specifičnosti tog jezika. Najkvalitetniji postojeći sistem za sintezu govora na srpskom jeziku realizovan je u okviru projekta AlfaNum na Fakultetu tehničkih nauka u Novom Sadu. U daljem tekstu biće opisano snimanje govornih baza kontinualnog govora namenjenim korišćenju u okviru AlfaNum TTS sistema. Reč je o bazama identičnog sadržaja, ali izgovorenog od strane dveju različitih govornica, u daljem tekstu označenih sa G1 i G2, tako da će u okviru ovog rada poseban osvrt biti dat na uticaj karakteristika glasa na kvalitet sintetizovanog govora.

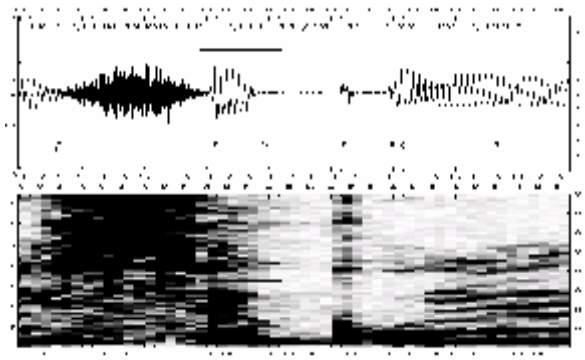
2. DIGITALNA OBRADA SIGNALA U SINTEZI GOVORA

Zahvaljujući razvoju računarske tehnologije, danas raspolažemo memorijama dovoljnog kapaciteta i procesorima dovoljnih snaga, da se mogu koristiti algoritmi za sintezu govora na osnovu teksta koji se oslanjaju na velike količine unapred snimljenog govornog materijala. Iz ovog materijala se, u trenutku sinteze, kada je poznato koju govornu celinu treba sintetizovati, biraju segmenti snimljenog govora nad kojima se vrši dodatna obrada, iz dva razloga. Prvi cilj ove obrade je da približi akustičke parametre svakog od odabranih segmenata iz baze odgovarajućim parametrima ciljnih segmenata koje treba sintetizovati, s obzirom da se u bazi retko kad nalazi tačno ono što je u datom trenutku potrebno. Drugi cilj obrade je da ublaži čujne prelaze između pojedinih segmenata koji će biti spajani, odnosno da ujednači vrednosti akustičkih parametara na krajevima pojedinih segmenata. Sam metod obrade zavisi od konkretnog implementiranog modela govornog signala. Segmenti se biraju iz govorne baze na osnovu kriterijuma minimizacije ukupnog zahvata digitalne obrade signala, pošto veća modifikacija akustičkih parametara narušava postojeću boju glasa i unosi veće degradacije u govorni signal. Ovo je pogotovo tačno u slučaju da govorni signal poseduje izvesna oštećenja koja u spontanom govoru mogu delovati prirodno u postojećem kontekstu (mljackanje, nepravilnosti u radu glasnih žica, nepravilnosti u izgovoru visoko tranzijentnih glasova...). Međutim, podvrgnute digitalnoj obradi signala, a pogotovo premeštene u neki drugi kontekst, ove manifestacije mogu prerasti u velika oštećenja sintetizovanog govornog signala. Neke od ovih manifestacija su karakteristika samog glasa i na njih govornik ne može da utiče, dok se druge mogu u manjoj ili većoj meri izbeći odgovarajućim načinom izgovora teksta. Faza realizacije TTS sistema u kojoj se može učiniti najviše na otklanjanju ovakvih problema je upravo snimanje govorne baze.

3. FAKTORI SNIMANJA GOVORNE BAZE KOJI UTIČU NA KVALITET TTS-a

Faktori tehničke prirode

Za kvalitet govorne baze, a time i kvalitet sintetizovanog govora, veoma su bitni tehnički uslovi u kojima je snimanje izvedeno. Potrebno je da snimanje bude obavljeno u što kvalitetnijem studiju, sa po mogućstvu što slabije izraženim ehom. Eho je, u ovom slučaju, posebno nepoželjan, zato što dovodi do efekata sličnih koartikulaciji u govornom signalu. Drugim rečima, u talasnom obliku koji odgovara jednom glasu prisutni su i tragovi prethodnih glasova. Ovaj efekat ilustrovan je slikom 1, gde je posebno označen glas *e* u reči *sekunde*, u kom se oseća prisustvo prethodnog glasa *s*. Ovo se može registrovati slušanjem, kao i posmatranjem spektrograma u kom postoje značajne komponente na visokim učestanostima, netipične za vokal *e*. Ova pojava sigurno nije posledica koartikulacije, jer se po završetku izgovaranja glasa *s* gornja i donja vilica razilaze i frikcija u potpunosti prestaje.



Slika 1. *Uticaj eha na govorni signal i njegov spektrogram*

U kontinualnom govoru, u odgovarajućem kontekstu, ta pojava zvuči sasvim prirodno, međutim, u nekom drugom kontekstu deluje kao oštećenje govornog signala, jer nema razloga da se u okviru jednog glasa oseća prisustvo nekog glasa koji uopšte nije izgovoren.

Poželjno je da snimanje bude izvedeno što kvalitetnijim mikrofonom. Kao i kod svakog snimanja govora, treba voditi računa o konstantnom rastojanju mikrofona od govornika (da

ne bi dolazilo do prevelikih odstupanja u nivou snimljenog signala, kao i boji glasa, s obzirom da čovek nije tačkasti izvor zvuka) i pravilnom položaju mikrofona u odnosu na govornika. Poželjno je u tom cilju pre snimanja analizirati talasne oblike nekoliko test-snimaka. Ostala oprema, a pogotovo zvučna kartica računara, treba da bude dovoljno kvalitetna (naročito sa aspekta odnosa signal-šum), kako dobitak ostvaren zahvaljujći kvalitetu ostale studijske opreme ne bi bio anuliran. Što se tiče formata zapisa snimaka govornog signala na digitalni medijum, ranije je ovaj izbor u velikoj meri bio uslovljen pristupačnošću medijuma velikih kapaciteta, s obzirom da su govorne baze za tadašnje pojmove u nekomprimovanom obliku bile izuzetno velike. Izlaz je tražen u raznim tehnikama kompresije (jedna od prednosti uvođenja govornih modela bila je upravo ta da omogućuju značajnu kompresiju baze), kao i u smanjenju učestanosti odabiranja. Za primene u telefonskim aplikacijama dovoljnom je smatrana učestanost odabiranja od 8 kHz, a za ostale primene korišćena je učestanost odabiranja od 16 kHz. Danas ova ograničenja više ne postoje, pa je za potrebe AlfaNumTTS sistema usvojena učestanost odabiranja od 22 kHz. Snimci su u *mono* tehnici, i nad njima nije primenjena nikakva kompresija.

Način izgovora teksta u celini

Osim faktora tehničke prirode, na kvalitet sintetizovanog govora utiču i određene karakteristike glasa govornika, kao i način izgovaranja. Govornik može da utiče na način izgovaranja teksta samo u određenoj meri. Od njega se ne može tražiti da izbegava određena karakteristična oštećenja pojedinih glasova svojstvena za njega, pogotovo ne da to dosledno čini nekoliko sati, koliko sesija snimanja može da potraje, ali se kvalitet sintetizovanog govora ipak može podići tako što se od govornika traži da pri izgovaranju teksta poštuje pravila čiji je cilj da minimizuju potrebe za budućom obradom govornog signala u trenutku sinteze. Naime, od govornika se traži:

- Da izgovara tekst ujednačenom jačinom glasa,
- Da izgovara tekst ujednačenom brzinom,
- Da ne dopušta prevelike oscilacije visine glasa u toku čitanja, odnosno, da tekst izgovara prirodno, ali bez prevelikih emocija, onako kako se obično čitaju tekstovi informativnog sadržaja,
- Da izgovara tekst ujednačenom bojom glasa, do čije promene može doći ako se, primera radi, promeni nivo emocija koje govornik unosi, ili ako govornik „primeti“ da nešto nije dobro čitao ranije,
- Da u dovoljnoj meri artikuliše glasove.

Ako govornik u dovoljnoj meri ispoštuje prva četiri uslova, smanjuje se potreba za kasnijom modifikacijom glasnosti, trajanja i osnovne učestanosti segmenata, čime se smanjuje i nivo degradacije koji bi na taj način bio unet u signal. Od govornika zavisi i u kojoj će meri uspeti da održi koncentraciju tokom snimanja, kako česte intervencije stručne osobe koja nadzire tok snimanja ne bi suviše usporile snimanje i zamorile govornika. Značaj poslednjeg uslova biće detaljnije objašnjen u nastavku teksta.

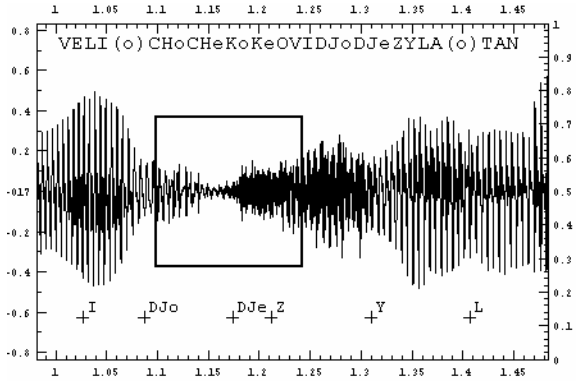
Karakteristike glasa i specifičnosti u izgovoru pojedinih glasova

Svetska iskustva u sintezi govora na osnovu teksta govore da je izbor glasa jedan od presudnih faktora sinteze govora koji zvuči prirodno. Nije dovoljno samo naći govornika bez uočljivih govornih mana, već je poželjno i da glas bude prijatan za slušanje. Drugim rečima, pri izboru glasa se treba rukovoditi sličnim kriterijumima kao i prilikom izbora glasa televizijskog ili radijskog spikera. Pored toga, izbor glasa može zavisiti i od primene TTS sistema.

Među važnijim karakteristikama glasa koje utiču na kvalitet sintetizovanog govora ističu se tipičan opseg kretanja osnovne učestanosti glasa (poželjno je da ne bude prevelik), kao i način izgovora pojedinih glasova. Kao što je navedeno, bitno je da izgovoren tekst bude artikulisan u dovoljnoj meri. Problem nedovoljno artikulisanog govora nije samo u tome što je značajan broj glasova oštećen, već i što je koartikulacija znatno izraženija, pa su pojedini glasovi u različitim kontekstima artikulisani na potpuno različit način. Ako usvojen sistem obeležavanja glasova u bazi (labeliranja) ne obuhvata mogućnost da se ta pojava na neki način evidentira, različite instance istog fonema biće obeležene na isti način, pa je moguće da će, u nekom trenutku, biti upotrebljene za spajanje, što će zvučati neprirodno, jer se, primera radi, dotičan glas artikuliše ili na jedan ili na drugi način, ali nikad na način koji je "između" ta dva. Jasna artikulacija izgovorenog teksta vodi ka doslednosti u izgovaranju pojedinih glasova, što obezbeđuje postojanje dovoljno velikog broja potencijalnih kvalitetnih spojeva u bazi, a samim tim i viši kvalitet sintetizovanog govora. Problem s oštećenjem pojedinih fonema nije samo u tome što se takvi fonemi ne mogu koristiti u sintezi. Veći problem je što često pojavljivanje oštećenih fonema smanjuje širinu konteksta koji se u celini može preuzeti iz baze.

Po iskustvima stručnih osoba koje su radile na korekciji automatski labeliranog snimljenog materijala, pojedini fonemi su osetljiviji na način izgovora od drugih. Tu su, pre svega, laterali i poluvokali, kao što su *l*, *v* i *j*, i vibrant *r*. Konsonant *j* je posebno zanimljiv, s obzirom da se, kad se nađe između vokala, u nekim situacijama potpuno gubi. Deo ovih problema može se rešiti softverski, tako da se, ukoliko očekujemo da se glas izgubi, on automatski izbacuje prilikom fonetske transkripcije teksta koji treba prevesti u govor. Primera radi, ukoliko znamo da će se *j* uvek izgubiti između *i* i *a*, ne treba ga obeležavati u bazi, a kad je potrebno sintetizovati govornu celinu koja ga sadrži u tom kontekstu (primera radi, reč *majica*), u bazi treba tražiti segmente čijim će se spajanjem generisati govorna celina bez tog glasa (dakle, [maica]). Fonetska transkripcija teksta može da reši i ozbiljnije probleme, kao što je, primera radi, jednačenje po zvučnosti. Ako treba sintetizovati govornu celinu *pet_banana*, u bazi umesto spoja *t_b* tada se traži spoj [db], čak i ako je u okviru iste reči, npr. *su**db**ina*. To je moguće zahvaljujući tome što se reči u okviru veće govorne celine tipično izgovaraju bez ikakve pauze. Na slici 2 prikazana je pojava jednačenja po zvučnosti na granici između reči *Veličković_**Z**oran*. Da bi se mogla definisati što efikasnija pravila fonetske transkripcije i uvećao broj kvalitetnih spojeva u bazi, poželjno je da govornik bude što dosledniji u tome da li će vršiti jednačenje po zvučnosti na granicama između reči ili ne. Slično se odnosi i na druge glasovne

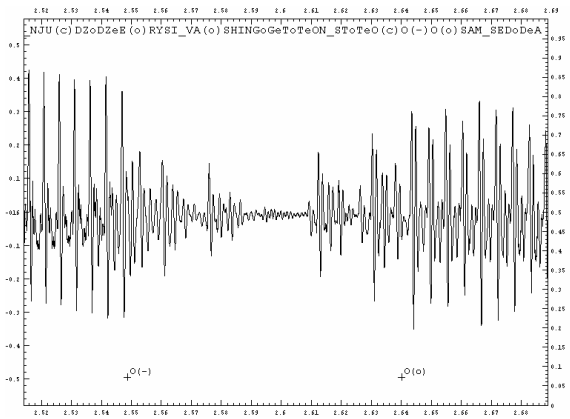
promene, kao što je potpuno gubljenje suglasnika u određenom kontekstu.



Slika 2. *Jednačenje po zvučnosti na granici između reči (G1)*

Jedan od problema s kojim su se autori ovog rada susretali prilikom kompletne realizacije ovih govornih baza je što ne mali broj govornika ima tendenciju ka ubacivanju kratkih pauza između pojedinih grupa reči u rečenici (obično između sintagmi, kao i gde god im se čini da bi to doprinelo razumljivosti govora). U tome po pravilu nisu dosledni, tako da opredeljivanje za jednoznačnu fonetsku transkripciju teksta dovodi do toga da je broj potencijalnih spojeva u bazi značajno smanjen. Primera radi, kada treba sintetizovati sintagmu ***sto_osam***, nije jasno da li u bazi treba tražiti niz glasova [sto#osam] ili [stoosam] (gde # označava tišinu). Štaviše, u bazi su tada vrlo česte pojave kao što je ova prikazana na slici 3. Pri izgovoru sintagme ***sto_osam***, govornica G2 namerno je ubacila pauzu, u nameri da poveća razumljivost govora (u čemu je, s aspekta slušaoca, i uspela), ali je time oštetila i završetak prvog i početak drugog glasa *o*. Time je onemogućila korišćenje dotičnog spoja u sintezi govora spajanjem segmenata, osim u identičnom kontekstu i bez ikakve dodatne obrade, a i to bi zahtevalo korišćenje složenijih konvencija labeliranja. To bi, s druge strane, dodatno usložnilo pretragu govorne baze i usporilo čitav postupak.

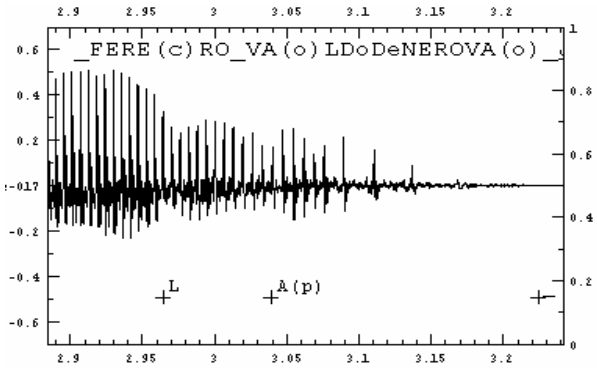
Zanimljivo je da su ovakve pojave znatno češće kod govornika sa iskustvom televizijskog ili radijskog spikera, odnosno, govornika sa boljom dikcijom.



Slika 3. *Oštećenje vokala pre i posle namerno ubačene pauze (G2)*

Od interesa je i da su glasne žice govornika dovoljno pokretljive, te da nepravilnosti u njihovom radu, među kojima je najizrazitiji primer tzv. *vocal fry* efekat, nastaju što ređe. Ove nepravilnosti javljaju se kada je dotok vazduha iz pluća slabiji, što znači, tipično na kraju govornih celina. *Vocal fry* efekat se često manifestuje tako što glasne žice ne propuštaju vazduh u dovoljnoj meri svaki put kada pritisak u ždrelu dostigne potrebnu vrednost, već svaki drugi ili treći put, kao što je prikazano na slici 4, što se registruje kao nagli pad osnovne učestanosti, a subjektivno kao karakteristična promuklost. Ovakav govorni signal nije preporučljiv za obradu u cilju modifikacije prozodijskih obeležja.

Pored ovog, česta su i oštećenja vokala do kojih dolazi pri bržem, nedovoljno artikulisanom govoru. Tada vokali imaju tendenciju da zvuče kao poluglas [∂]. Ovakvi vokali u originalnom kontekstu najčešće dobro zvuče i teško ih je uočiti, ali ako se nađu u drugom kontekstu (naročito u okviru naglašenog sloga) mogu značajno da umanje kvalitet sintetizovanog govora, pa i da ga učine nerazumljivim.



Slika 4. *Vocal fry ekekat (G2)*

Još jedan veoma bitan faktor pri izboru govornika je i talasni oblik govornog signala (engl. *waveform*). Naime, poznato je da sa veštačkom promenom osnovne učestanosti govornog signala dolazi i do promene boje glasa govornika. Izraženost te pojave zavisi od primenjene metode sinteze, tako da je kod vremenske metode ona velika, dok neke složenije, hibridne metode donekle potiskuju ovu pojavu. U svakom slučaju, poželjno je imati što jednostavniji talasni oblik koji će pretrpeti minimalne izmene nakon promene osnovne učestanosti. U tom smislu bi pikovi u zvučnim delovima govora (naročito vokalima) morali biti jasno izraženi, a što veća snaga signala koncentrisana oko njih. Nepovoljno je kad postoje značajne promene tokom čitave periode. Ovo će biti ilustrovano kod poređenja karakteristika glasa dveju govornica.

4. POREĐENJE KARAKTERISTIKA GLASA GOVORNICA G1 I G2

Kao što je već rečeno, za potrebe AlfaNumTTS sistema do sada su snimljeni glasovi dveju govornica (G1 i G2). Govornica G1 odabrana je u ranijoj fazi projekta, kada svi opisani problemi još nisu bili poznati. Govornica G2 odabrana je između pet profesionalnih spikerki, čiji su snimci prethodno bili pažljivo proučeni. Otud i razlike koje će biti navedene u daljem tekstu. Poređenje će biti obavljeno po nekoliko već navedenih kriterijuma od značaja.

Studijski uslovi. U oba slučaja snimanje je obavljeno u studijskim uslovima (mada ne u istom studiju), a zvučni

fajlovi snimljeni u formatu 22kHz, 16 bita po odmerku, te s tog aspekta nema razlika.

Tekst koji su čitale govornice je takođe isti u oba slučaja.

Ravnomernost brzine čitanja. Obe govornice su čitale tekst relativno ujednačenom brzinom. Ponekad je, naravno, bila potrebna intervencija prisutnog tehničara, ali su oscilacije ostale u prihvatljivim okvirima. G1 je govorila nešto sporije, pošto pri sporijem izgovoru ređe dolazi do oštećivanja vokala, a i pri obradi signala u fazi sinteze glasove je lakše skratiti nego produžiti (naročito vokale). G2 nije oštećivala vokale ni pri bržem izgovoru, tako da je mogla da čita skoro prirodnom brzinom.

Ravnomernost jačine glasa i osnovne učestanosti. Obe govornice su čitale tekst relativno ujednačenom jačinom i visinom glasa.

Ravnomernost boje glasa. Ovde su se javljala odstupanja, s tim da su ona kod G1 bila ređa i ne toliko izražena, dok se kod G2 mogla primetiti značajna promena u nivou emocija nastala posle nekoliko desetina rečenica i jedne intervencije tehničara. Nakon toga, G2 je održavala prilično ujednačenu boju. Ovaj prvi deo je izostavljen iz baze. Preporučljivo je da se pre svake nove sesije snimanja (nakon pauze) govorniku puštaju raniji snimci, kako bi što tačnije imitirao raniju boju, brzinu i visinu glasa.

Artikulacija glasova. G2 znatno bolje artikuliše glasove od G1 i ne oštećuje vokale čak ni pri bržem izgovoru. G1 ima neznatnu govornu manu, pri izgovoru kombinacije glasova [šć], dok kod G2 nisu primećeni nikakvi problemi pri izgovoru.

Greške. Greške pri čitanju nisu velik problem za samu sintezu, pošto govornica u tim slučajevima ili ponovo pročita rečenicu, ili se za sintezu koristi ono što je pročitala umesto onog što je planirano. Međutim, greške mogu značajno da produže (i poskupe) snimanje i obradu baze, jer se studio duže koristi i zahteva se više intervencija od ljudi koji naknadno obrađuju bazu. G1 je pravila umerenu količinu grešaka, dok ih G2 skoro uopšte nije pravila.

Boja glasa i talasni oblik. G1 ima dosta promenljivu boju glasa, i ima tendenciju da je menja čak i u okviru istog vokala. Ni za G2 se ne može reći da ima klasičnu spikersku boju glasa, ali je svakako mnogo bliža tome. Što se tiče talasnog oblika, on je kod G2 daleko jednostavniji i utisak je da je to možda i glavni razlog višeg kvaliteta sinteze sa G2.

Vocal fry. Još jedno od preimućstava G2. Od pet pregledanih spikerki, ona je bila jedina kod koje se ovo oštećenje gotovo uopšte nije javljalo. G1 ga je relativno često imala, i to ne samo na kraju rečenice, već često i u okviru reči.

Jednačenje po zvučnosti. G1 je skoro uvek jednačila po zvučnosti glasove na granicama između reči, dok G2 skoro nikad to nije radila. I jedno i drugo je dobro (dogod je konzistentno), samo ga treba uzeti u obzir pri fonetskoj transkripciji.

Pauze između reči. Obe govornice su imale tendenciju da ponekad naprave pauzu između reči u rečenici, s tim što se to kod G2 dešavalo znatno češće. Osim toga, često je dolazilo do koartikulacije dva vokala i pored pauze koja se nalazila između njih. Ovo je u osnovi loše, ali se može delimično prevazići pažljivim labeliranjem baze.

Na kraju treba napomenuti da, kao posledica svih ovih razlika, sinteza govora na osnovu baze snimljene sa G2 zvuči znatno prirodnije od one na osnovu baze snimljene sa G1, barem što se tiče segmentnog nivoa.

5. ZAKLJUČAK

U radu je obrađena problematika snimanja kvalitetne baze za potrebe TTS sinteze povezivanjem segmenata odabranih u realnom vremenu, sa osvrtom na probleme na koje je AlfaNum tim nailazio tokom snimanja ovih baza. Dato je mnoštvo zapažanja i ukazano je na razne probleme koji su se javljali. Date su i uporedne karakteristike dveju govornica sa kojima je AlfaNum tim u proteklom periodu obavio snimanje govorne baze.

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] M. Sečujski, *Sinteza govora na osnovu teksta sa osvrtom na srpski jezik*, Diplomski rad, Fakultet tehničkih nauka, Novi Sad, 1999.

[4] M. Sečujski, *Prozodijski elementi u sintezi govora na srpskom jeziku*, Magistarski rad, Fakultet tehničkih nauka, Novi Sad, 2002.

[5] S. Jovičić, *Govorna komunikacija fiziologija, psihoakustika i precepcija*, Nauka, Beograd 1999.

Abstract – This paper discusses problems of speech database recording for the purposes of TTS system with on-line segments selection. Special attention is given to speaker selection, and some comparative characteristics of two recorded female speakers were given.

VOICE CHARACTERISTICS INFLUENCE ON SYNTHESIZED SPEECH QUALITY

Milan Sečujski, Darko Pekar, Vlado Delić