Pridobivanje govorne zbirke za korpusni sintetizator govora Phonectic

Aleš Mihelič*, Jerneja Gros*, Nikola Pavešić†, Mario Žganec*

*Masterpoint razvoj in raziskave
Baznikova 40, Ljubljana, Slovenija, www.masterpoint.si
info@masterpoint.si

†Fakulteta za elektrotehniko, Univerza v Ljubljani
Tržaška c. 25, Ljubljana, Slovenija

Povzetek

V članku opisujemo snovanje in ustvarjanje govorne zbirke za korpusno sintezo slovenskega govora. Sprva smo izvedli frekvenčno analizo pogostosti pojavljanja glasovnih sklopov za slovenski jezik nad obsežnim vhodnim besedilom, ki smo ga pretvorili v fonetični prepis. Nadalje opisujemo postopek, ki iz množice slovenskih besedil v pisni obliki izbere kompaktno množico povedi, ki vsebujejo vsa želena zaporedja glasov. Z bistveno manjšo začetno zbirko besedil nam je uspelo ustvariti štirikrat krajše besedilo kot je to uspelo konkurenčnemu postopku. Prav tako vsebuje naša govorna zbirka več različnih glasovnih nizov. Sledi opis snemanja in označevanja govorne zbirke, kjer smo uporabili programsko orodje SIGMARK, ki omogoča segmentacijo in označevanje glasovnih segmentov ter samodejno označevanje osnovne periode.

1. Uvod

S hitrim razvojem računalniške tehnologije se je zmožnost razvoja sistemov, ki omogočajo komunikacijo med človekom in strojem v naravnem jeziku, močno povečala. V modernem času je razvoj govornih tehnologij, predvsem to velja za sintezo in prepoznavanje govora, doživel izjemen razcvet in v vsakdanjem življenju se že uporabljajo raznovrstne rešitve s tega področja – v samodejnih informacijskih centrih, v govornih portalih, za glasovno prebiranje elektronske pošte.

Vse več govornih zbirk in rezultatov jezikovnih študij je dostopnih tudi v našem prostoru. Jezikovne tehnologije nezadržno prodirajo v vsakdanje življenje. Na tržišču je kar nekaj solidnih sintetizatorjev govora, nekaj med njimi jih podpira tudi slovenski jezik. Razvoj in raziskave s področja sinteze slovenskega govora se odvijajo na Fakulteti za elektrotehniko Univerze v Ljubljani (Gros, 1997; Vesnicer et al., 2001), na Inštitutu Jožef Stefan (Šef, 2001), na Fakulteti za elektrotehniko, računalništvo in informatiko na Univerzi v Mariboru (Rojc et al., 1999; Rojc et al., 2000) ter v podjetju Masterpoint iz Ljubljane (Gros et al., 2001). Z izjemo Masterpointovega korpusnega sintetizatorja govora Phonectic so v okviru nam poznanih in dostopnih objavljenih informacij ostali sistemi za sintezo slovenskega govora zasnovani na difonski sintezi govora, z načrti za prehod na korpusno sintezo govora (Rojc et al., 2000; Vesnicer et al., 2001; Šef, 2001)..

V nadaljevanju opisujemo postopek zasnove in realizacije nove, kvalitetne zbirke osnovnih govornih enot, ki je potrebna za delovanje korpusnega sintetizatorja govora za slovenski jezik *Phonectic*. *Phonectic* (poznan tudi pod imenom Mp-Synth, verzija 3.0) razvija podjetje Masterpoint, v raziskavah slovenskega govorjenega jezika pa tesno sodeluje z Laboratorijem za umetno zaznavanje, sisteme in kibernetiko na Fakulteti za elektrotehniko Univerze v Ljubljani.

Za korpusno sintezo govora potrebujemo glasovno zbirko, ki vsebuje posnete osnovne enote govora (Conkie, 1999; Conkie et al., 2000). Kvaliteta na ta način sintetiziranega govora je v precejšnji meri odvisna od kvalitetno posnete govorne zbirke (Beutnagel et al, 1999).

Najkvalitetnejši govor bi pridobili iz zbirke, kjer bi bile posnete vse možne besede in besedne zveze, ki se pojavljajo v slovenskem jeziku, v vseh možnih kontekstih in besednih zvezah. Za sintezo govornega signala bi v govorni zbirki samo poiskali ustrezno frazo in jo brez spreminjanja njenih prozodičnih lastnosti uporabili. Vendar bi bila takšna zbirka preobsežna, pa tudi posneti bi jo bilo zelo težko.

Drugo skrajnost predstavlja glasovna zbirka, v kateri so posneti samo vsi glasovi oziroma izbrani alofoni, ki nastopajo v slovenskem govoru. Takšne zbirke v praksi obstajajo in zadoščajo za zadovoljivo kvaliteto sinteze govora. Pri sintezi v taki zbirki poiščemo ustrezen alofon, mu spremenimo prozodične lastnosti in ga uporabimo. Pri tem na meji posameznih alofonov po lepljenju pogosto zasledimo popačenja, ki pa jih je možno deloma odpraviti z ustreznimi postopki (MBR-PSOLA, FD-PSOLA).

V splošnem velja, da je sintetizirani govor kvalitetnejši, če uporabljamo za sintezo daljše osnovne segmente s čim manj spremembami prozodičnih parametrov (Kopeček, 2001; Yi, 1998). Čim daljše osnovne segmente uporabljamo, tem večje je število različnih segmentov, ki jih moramo uvrstiti v zbirko. Število segmentov povečuje tudi želja po minimalnih spremembah prozodičnih parametrov, saj te vnašajo popačenja v sintetizirani govor. Pri tem imamo enako zaporedje fonemov posneto večkrat, z različnimi prozodičnimi lastnostmi, kar posledično zopet pomeni obširnejšo glasovno zbirko.

Zasnovo zbirke za korpusno sintezo govora lahko logično razstavimo v tri zaporedne korake, ki jih podrobneje predstavljamo v nadaljevanju članka:

- izbira besedila, potrebnega za snemanje govorne zbirke,
- snemanje govornega gradiva,
- segmentacija in označevanje govornega gradiva.

Pri načrtovanju vsebine glasovne zbirke za novi sintetizator slovenskega govora smo precejšnjo pozornost

posvetili prav izbiri optimalnega besedila. Vloženi trud se je poplačal, saj smo dobili razmeroma majhno in z raznolikimi glasovnimi nizi bogato vhodno besedilo. Daljše besedilo namreč pomeni zamudnejše snemanje, predvsem pa se poveča težavnost segmentacije in označevanja obsežnega posnetega govornega gradiva.

2. Izbira besedila

Besedilo za snemanje smo izbirali iz obsežne zbirke slovenskih besedil, ki so pokrivala različne zvrsti, kot so dnevni časopisi, revije, leposlovje. Izmed vseh povedi v zbirki smo najprej izločili vse, ki niso vsebovale vsaj petih besed. Dobili smo preko 200,000 različnih povedi (25 MB v navadnem ASCII zapisu), ki so ustrezale našim željam. Besedila teh povedi smo pretvorili v fonetični prepis. Fonetični prepis besedila smo izvedli z modulom za grafemsko fonemsko pretvorbo sintetizatorja Phonectic (Gros et al., 2001).

2.1 Analiza pogostosti pojavljanja glasovnih nizov

Večina dosedaj izvedenih analiz pogostosti zaporedij glasov v slovenskem jeziku je bila izvedena nad grafemskim zapisom ter se je ukvarjala z zaporedji grafemov. Nam pa je pretvorba besedila v fonetični prepis omogočila preučevanje zaporedij posameznih glasov oz. alofonov.

Nad fonetičnim prepisom besedila smo izvedli frekvenčno analizo pojavljanja glasovnih nizov - alofonov, difonov, trifonov in štirifonov. S tem smo dobili vpogled v pogostost pojavljanja posameznih sklopov alofonov v slovenskem govorjenem jeziku.

Kot zanimivost v sliki 1 predstavljamo rezultate frekvenčne analize slovenskih alofonov. Alofoni so predstavljeni v SAMPA notaciji. V analizi razlikujemo med naglašenimi in nenaglašenimi alofoni.

Iz grafov na slikah 2 in 3 je razvidno, da se zelo pogosto pojavlja le nekaj trifonov, nekateri trifoni pa so razmeroma redki. Prav zaradi tega je v govorno zbirko nesmiselno uvrstiti vse trifone. Odločili smo se za prvih 500 trifonov, kar znaša le 1% vseh možnih trifonov, na katere smo naleteli pri analizi obsežne zbirke besedila. Vseh trifonov v zbirki povedi skupaj je 17784, vseh različnih pa je teoretično možnih 50653, vendar se jih kar precejšnje število v slovenskem jeziku nikoli ne pojavi. Na ta način smo pokrili skoraj polovico vseh pojavljanj trifonov v naši zbirki povedi. Podobno smo obravnavali tudi štirifone.

Odločili smo se, da govorno zbirko sestavimo iz vseh difonov ter najpogosteje uporabljanih trifonov ter štirifonov. Sklopi trifonov in štirifonov krepko pripomorejo h kakovostno sintetiziranemu besedilu, saj je potrebno manj lepljenj, kar v končni fazi privede do manj vnesenih popačenj. Sklenili smo, da v govorno zbirko uvrstimo prvih 500 v besedilih najpogosteje ponavljajočih se trifonov in 300 štirifonov.

2.2 Algoritem za optimalno izbiro povedi

Za cilj smo si zadali, da zasnujemo kompaktno govorno zbirko s čim manj besedila, ki bi ga morali posneti, hkrati pa je moralo besedilo vsebovati vse pogoste glasovne sklope (difoni, trifoni, štirifoni), ki smo jih želeli v njej imeti. Zato smo razvili poseben algoritem za optimalno izbiro povedi in s tem tudi minimizacijo obsega govorne zbirke (Mihelič, 2002).

Vsem povedim smo določili ceno, in sicer na podlagi bogatosti z glasovnimi skupinami, ki smo jih potrebovali. Štirifonom, ki jih v govorni zbirki še nismo imeli in jih je trenutno obravnavana poved vsebovala, smo predpisali najvišjo ceno, trifonom nižjo in difonom najnižjo. Posamezne vrednosti smo sešteli in dobili skupno ceno. Da ne bi vedno izbrali najdaljše povedi iz korpusa (taka ima namreč statistično gledano največjo možnost, da vsebuje največ ustreznih glasovnih skupin), smo cene normirali s številom alofonov v povedi.

Poved z najvišjo ceno smo iz korpusa premestili v seznam povedi za snemanje, prav tako pa smo iz seznama želenih glasovnih skupin izločili vse glasovne skupine, ki jih je ta poved vsebovala. Preostalim povedim v korpusu smo na novo izračunali ceno in zopet izbrali poved z najvišjo ceno. Celoten postopek smo ponavljali v zanki, dokler nismo z izbiro povedi pokrili prav vseh želenih glasovnih zvez.

Algoritem za optimalno izbiro povedi je izbral razmeroma maloštevilčno skupino povedi, bogatih z vsemi različnimi glasovnimi skupinami, ki smo jih zahtevali. Vendar smo poleg teh pridobili še vrsto drugih glasovnih skupin, ki niso bile navedene v zahtevah.

Najbolje bi bilo poiskati optimalno kombinacijo uteži, ki bi generirala najkrajšo izhodno zbirko povedi po kateri izmed uveljavljenih optimizacijskih metod (metoda simpleksov, optimizacija z nevronskimi mrežami, najmanjša napaka kvadratov, ipd.). Zaradi časovne zahtevnosti optimizacije nismo izvedli. Začetne nabore uteži smo določili iz predhodnih izkušenj ob razvijanju sintetizatorja slovenskega govora. Nabore uteži smo spreminjali in na koncu izbrali najustreznejšega.

Opisani način izbire povedi, primernih za uvrstitev v glasovno zbirko, se je izkazal za zelo učinkovitega in uporabnega. Iz zbirke preko 200,000 povedi, iz katere so bile že izločene take, ki so bile bodisi neprimerne ali prekratke, smo dobili vsega skupaj le 297 povedi. Te povedi smo uporabili kot besedilo pri snemanju govorne zbirke.

2.3 Primerjava postopkov za izbiro povedi

Podoben algoritem za izbiro povedi so razvili na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (Rojc et al., 2000). Za cilj so si postavili kar največjo razgibanost končnega besedila, pri čemer naj bi se posamezna glasovna zaporedja kar najmanjkrat ponavljala.

V nadaljevanju podajamo primerjavo obeh postopkov:

1. Postopek za izbiro povedi – Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru (Rojc et al., 2000):

- Velikost uporabljene zbirke besedil: 31 milijonov besed, ~2 milijona povedi.
- Cilj: Zbirka s kar največjim možnim naborom glasovnih nizov, ki naj se v njej čim manjkrat ponovijo. Zbirka naj bo po obsegu čim manjša. Vsebuje naj 1200 povedi.

- Postopek: Iz celotnega besedila izloči vse povedi, ki so krajše od 15 besed in daljše od 25 besed. Generiraj 4 manjše zbirke s po 5000 povedmi, ki ustrezajo postavljenim omejitvam. Posamezne zbirke se uporabijo za nadaljnjo obdelavo. Vsakemu glasovnemu nizu določi pogostost pojavljanja v vsaki od 4 majhnih zbirk. Izloči povedi, ki vsebujejo podvojene glasovne nize. Pri tem pazi, da število ponavljanj posameznega trifona ne pade pod vnaprej določen prag. Prag znižuj toliko časa, dokler ni v govorni zbirki samo 1200 povedi. Preglej strukturo vseh štirih zreduciranih zbirk in izberi najboljšo. Povedi, ki najmanj doprinesejo k raznolikosti zbirke, zamenjaj z najboljšimi povedmi iz preostalih treh zbirk.

- Rezultat: Končni rezultat predstavlja zbirka s 1200 povedmi, ki vsebuje vsa glasovna zaporedja iz večje zbirke. Povedi vsebujejo **9391** različnih trifonov.

2. Postopek za optimalno izbiro povedi – sintetizator govora Phonectic:

- Velikost uporabljene zbirke besedil: 3.2 milijona besed, ~200,000 povedi, v fonetičnem prepisu ~17.3 milijonov glasov.

- Cilj: Zbirka z vsemi izbranimi glasovnimi nizi, vsemi kombinacijami difonov in najpogostejšimi trifoni ter štirifoni. Zbirka naj ima kar najmanjši obseg.

- Postopek: Statistično obdelaj celotno vhodno besedilo (in ne po delih, kot je bilo to izvedeno pri predhodno opisanem postopku) in določi pogostost pojavljanja posameznih glasovnih nizov v besedilu.

Izberi število najpogostejših trifonov in štirifonov, ki jih želimo imeti v zbirki. V našem primeru smo se odločili za 500 najpogostejših trifonov in 300 štirifonov.

Oceni doprinos glasovnih nizov za vsako poved. Izberi tisto z najvišjim doprinosom. Iz spiska želenih glasovnih nizov odstrani vse, ki jih izbrana poved vsebuje.

Ponovno oceni vsako poved in izberi najboljšo. Postopek ponavljaj, dokler besedilo zbirke ne vsebuje vseh želenih glasovnih nizov.

- Rezultat: Končni rezultat predstavlja zreducirana zbirka povedi (izhodiščno število povedi je bilo **200,000**). V njej se nahaja najmanjše možno število povedi, ki še vsebujejo vse želene glasovne nize (vse difone ter najbolj pogoste trifone in štirifone). V našem primeru se je število povedi zreduciralo na **297**. Te povedi vsebujejo skupno 10.218 alofonov.

Povedi vsebujejo **1,132** različnih difonov, **17,784** različnih trifonov ter **120,425** različnih štirifonov. Povprečna dolžina povedi v zbirki rahlo presega šest besed oz. 34.4 alofonov.

Iz povedanega sledi, da nam je z manjšo začetno zbirko besedil uspelo ustvariti štirikrat krajše besedilo, iz katerega smo posneli govor, potreben za govorno zbirko. Prav tako je naša govorna zbirka bogatejša, saj sestoji iz večjega števila različnih glasovnih nizov, pri tem da vključuje vse najbolj pogoste. Objektivno primerjavo uspešnosti obeh postopkov bi lahko dobili le z obdelavo iste vhodne zbirke slovenskih povedi.

3. Snemanje in označevanje govorne zbirke

3.1 Snemanje govorne zbirke

Besedilo, ki vsebuje vsa želena zaporedja alofonov, je najpriporočljiveje prebrati naenkrat. Zelo pomembno je namreč, da govorec skozi vse besedilo govori na enak način, z enakim glasom, z enako intonacijo, skratka z enakimi parametri govora. Snemanje besedila po kosih v daljšem časovnem obdobju ni priporočljivo, saj se govorcu lahko glas zaradi različnih zunanjih (vreme, drugačne nastavitve pri snemanju, spremenjen spekter in intenziteta motenj iz okolice) ali notranjih (razpoloženje, bolezen) vzrokov spremeni, govorna zbirka pa ni v celoti posneta, kar oteži kvalitetno sintezo govora.

Snemanje govornega gradiva je potekalo ob prisotnosti izkušenega snemalnega operaterja z namenom, da bi preprečili neustrezne izgovorjave besed in napake pri snemanju govora. Govorca smo prosili, da povedi prebira razločno in razmeroma počasi.

3.2 Označevanje govorne zbirke

Govorni zbirki je bilo potrebno dodati še oznake o začetku ter koncu posameznih glasov oziroma fonemov ter oznake osnovne periode. Za fonetične prepise posnetega govornega gradiva smo uporabili 36 različnih alofonskih oznak.

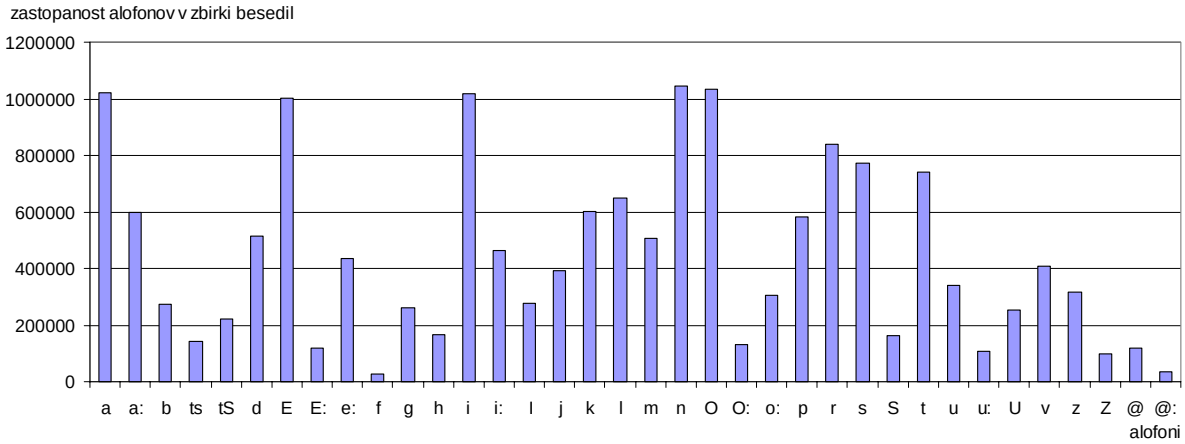
S postopkom vsiljenega prileganja posnetkov govora z grafi modelov glasov, ki so bili določeni iz fonetičnih prepisov izgovorjenih različic besed, smo si močno olajšali dolgotrajno in zamudno ročno označevanje glasov (Dobrišek, 2001).

Za ročno pregledovanje in označevanje govorne zbirke ter popravljanje oznak smo v Masterpointu razvili poseben računalniški uporabniški vmesnik - SIGMARK, ki omogoča prikaz in spreminjanje posnetih govornih signalov in izbranih akustičnih značilk ter poslušanje poljubnih odsekov signala (slika 4).

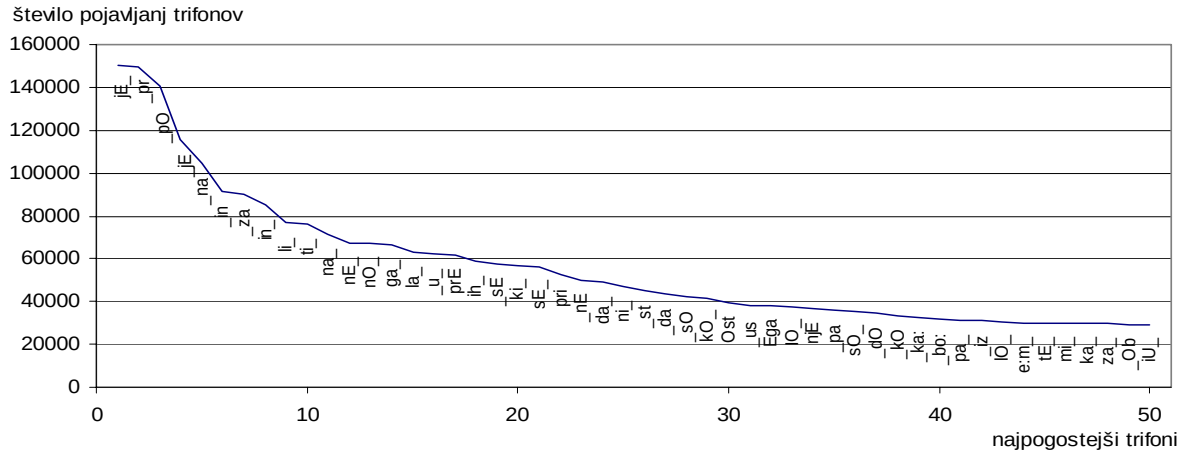
Orodje SIGMARK omogoča sočasni prikaz časovne in kratkočasovne frekvenčne karakteristike signala, kar močno olajša preverjanje in popravljanje mej med glasovi in oznak za glasove. Druga prednost orodja SIGMARK je samodejno in konsistentno postavljanje značk osnovne periode.

3.3 Obseg govorne zbirke

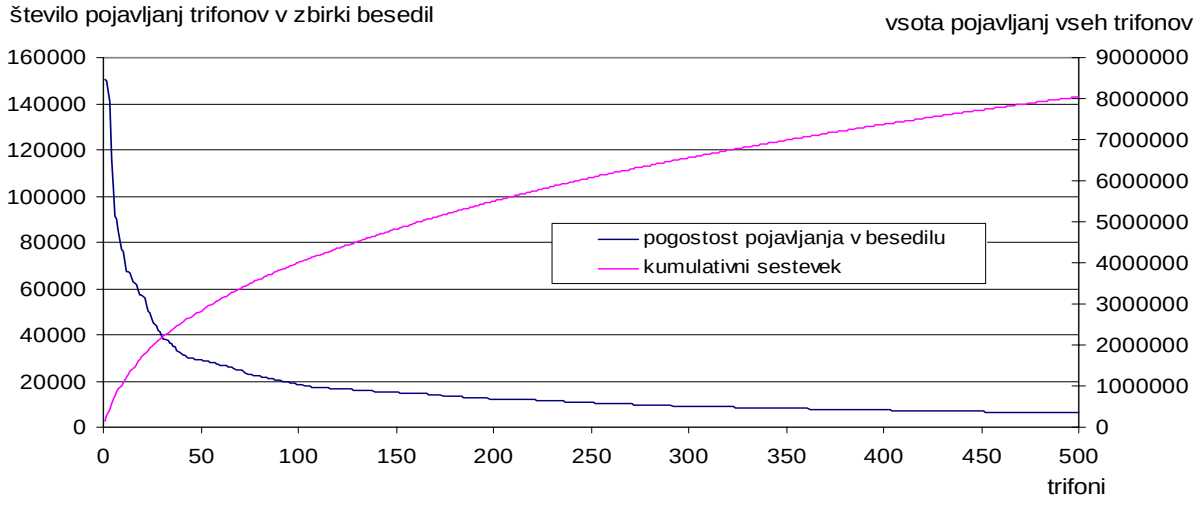
Končna govorna zbirka za korpusno sintezo govora vsebuje 297 različnih povedi s 1814 besedami ter vsemi potrebnimi difoni oziroma trifoni, ki smo jih izolirali (pridobili) iz logatomov. Vse povedi v korpusu smo opremili s kanoničnimi fonetičnimi prepisi. Obseg in razdelitev zbirke sta razvidna iz tabele 1. V tabeli so za posamezne dele zbirke podani časi trajanja posnetkov ter število vsebovanih besed in glasov. Po delih zbirke je podano še število vseh besed, število različnih besed in število glasov.



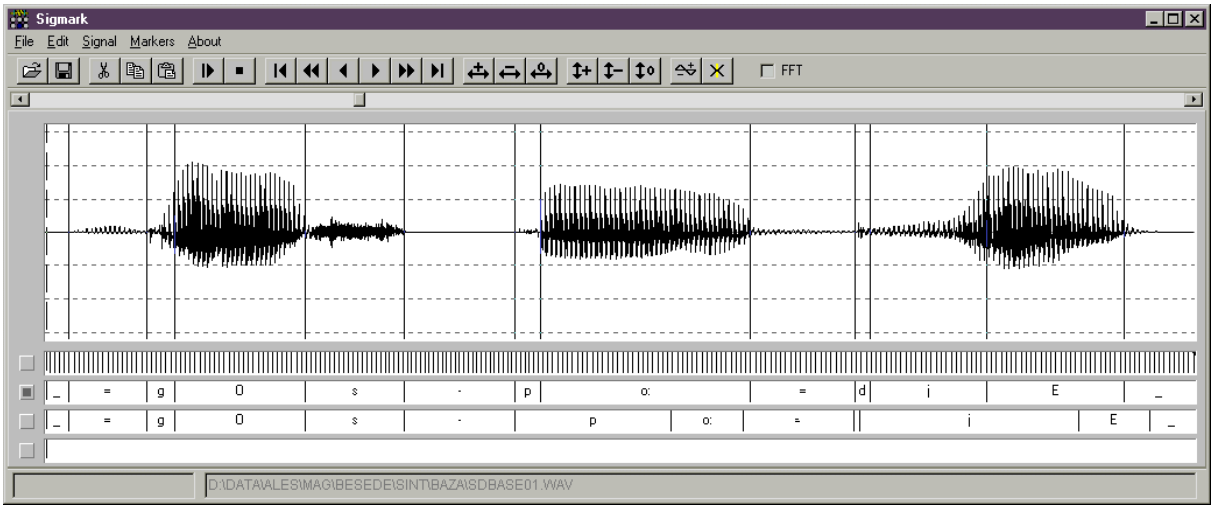
Slika 1: Prikaz frekvence zastopanosti alofonov v zbirki povedi, alofoni si sledijo v abecednem vrstnem redu.



Slika 2: Prikaz frekvence pojavljanja 50 najbolj pogostih trifonov v zbirki povedi.



Slika 3: Prikaz frekvence zastopanosti trifonov v zbirki povedi.



Slika 4: Orodje SIGMARK omogoča opremljanje signala z oznakami o mejah med posameznimi glasovi. Prva skupina oznak predstavlja potek osnovne frekvence signala, druga skupina oznak so ročno popravljene oznake mej med glasovi, tretja skupina pa prikazuje samodejno določene oznake mej med glasovi.

	trajanje	število besed		število glasov
		vseh	različnih	
	naraven govor			
zbirka naravnega govora	3622 s	1814	1354	10218
	logatomi			
zbirka logatomov	508 s	1169	1169	2338
zbirka logatomov (dvoglasniki)	1088 s	1668	1668	5004
celotna zbirka logatomov	1596 s	2837	2837	7342
celotna govorna zbirka za korpusno sintezo govora	5218 s (1h 27 min)	4651	4191	17560

Tabela 1: Obseg govorne zbirke za korpusno sintezo slovenskega govora.

4. Sklep

Pri snovanju govorne zbirke za korpusno sintezo govora smo izvedli frekvenčno analizo pogostosti pojavljanja glasovnih sklopov za slovenski jezik nad obsežnim vhodnim besedilom. Nadalje smo razvili postopek, ki iz množice slovenskih besedil v pisni obliki izbere kompaktno množico povedi, ki vsebuje vsa želena zaporedja glasov. Z bistveno manjšo začetno zbirko besedil nam je uspelo ustvariti štirikrat krajše besedilo, kot je to uspelo konkurenčnemu postopku. Prav tako vsebuje naša govorna zbirka več različnih glasovnih nizov. Iz tega lahko povlečemo zaključek, da smo postopek optimizacije besedila za govorno zbirko opravili uspešno. Končni rezultat, govorna zbirka za korpusno sintezo govora z vsemi potrebnimi oznakami, dobro služi svojemu namenu. Sintetizator Phonectic se že nekaj časa uporablja tudi v komercialne namene.

V razvoju je že naslednja verzija sintetizatorja slovenskega govora, ki bo vključevala algoritem za odpravo spektralnih nezveznosti na mestih lepljenja glasovnih nizov. Dodali bomo tudi izpopolnjeni algoritem za izbiro najprimernejšega govornega segmenta iz glasovne zbirke ter algoritem za iskanje optimalnega mesta lepljenja znotraj glasu.

5. Literatura

Beutnagel M., Mohri M., Riley M., 1999. *Rapid unit selection from a large speech corpus for concatenative speech synthesis*. In Proc. Eurospeech '99, Budapest.

Conkie A., 1999. *Robust unit selection system for speech synthesis.* Proc. Eurospeech '99, Budapest.

Conkie A., Beutnagel M., Syrdal A., Brown P., 2000. *Preselection of candidate units in a unit selection-based Text-to-Speech synthesis system*. In Proc. ICSLP '00, Peking.

Dobrišek S., 2001. *Analiza in razpoznavanje glasov v govornem signalu.* Doktorska disertacija, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Gibbon D., Moore R. in Winski R. (uredniki), 1997., Handbook of Standards and Resources for Spoken Language Systems, Mouton de Gruyter, Walter de Gruyter Publishers. Berlin.

Gros J., Mihelič F., Pavešić N., Žganec M., Mihelič A., Knez M., Merčun A., Škerl D., 2001. *The Phonectic SMS reader*. Proceedings. TSD 2001, Železna Ruda, Czech Republic.

Gros J., 1997. *Samodejno tvorjenje govora iz besedil*. Doktorska disertacija, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Hamon C., Moulines E., Charpentier C., 1998. *A diphone synthesis system based on time-domain prosodic modifications of speech.* Proceedings of the ICASSP'89, 238-241.

Kopeček I., 2001. *Algebraic Models of Speech Segment Databases.* Proceedings TSD 2001, Železna Ruda, Czech Republic.

Mihelič A., 2002. *Zbirka govornih signalov za sintezo slovenskega govora*. Magistrsko delo, Fakulteta za elektrotehniko, Univerza v Ljubljani.

Rojc M., Stergar J., Wilhelm R., Hain Horst- U., Holzapfel M., Horvat B., 1999. *A multilingual text processing engine for the Papageno text-to-speech synthesis system*. Proceedings Eurospeech '99, Budapest.

Rojc M., Kačič Z., 2000. *Design of optimal Slovenian speech corpus for use in the concatenative speech synthesis system.* Proceedings of the Second international conference on language resources and evaluation. Athens, Greece, 321-325.

Šef T., 2001. *Analiza besedila v postopku sinteze slovenskega govora*. Doktorska disertacija, Univerza v Ljubljani.

Toporišič J., 1991. *Slovenska slovnica*. Založba Obzorja, Maribor.

Vesnicer B., Pavešič N., Mihelič F., 2001. *Korpusna sinteza govora.* Zbornik ERK'01, B: 253-255, Portorož.

Yi J., 1998. *Natural-sounding speech synthesis using variable-length units*. MEE Thesis, Massachussetts Institute of Tehnology.