SLOVENSKI GOVOR NA INTERNETU

Tomaž Šef, Aleš Dobnikar, Matjaž Gams, Marko Grobelnik
Odsek za inteligentne sisteme
Institut Jožef Stefan
Jamova 39, 1000 Ljubljana, Slovenija
Tel: +386 61 1773419, fax: +386 61 1251038
e-mail: tomaz.sef@ijs.si

POVZETEK

Predstavljamo sintetizator slovenskega govora, ki je sposoben samodejnega pretvarjanja poljubnih slovenskih besedil v govor. Sistem temelji na združevanju osnovnih govornih enot s pomočjo algoritma TD-PSOLA, ki smo ga dopolnili z linearno interpolacijo s spremenljivim številom interpoliranih period. Zasnovan je modularno, kar omogoča enostavno popravljanje in spreminjanje posameznih delov sistema.
Sintetizator smo uporabili v zaposlovalnem agentu EMA in sicer za govorno posredovanje obvestil o prostih delovnih mestih na internetu.

ABSTRACT

This paper presents a text-to-speech (TTS) system, capable of synthesising continuous Slovenian speech. The system is based on the concatenation of basic speech units, diphones, using TD-PSOLA technique improved with a variable length linear interpolation process. Input text is processed by a series of independent modules. That enables an easy improvement of separate parts of the system.
Our system is used in an employment agent EMA that provides employment information through the Internet. It is the most often visited and used intelligent system in Slovenia.

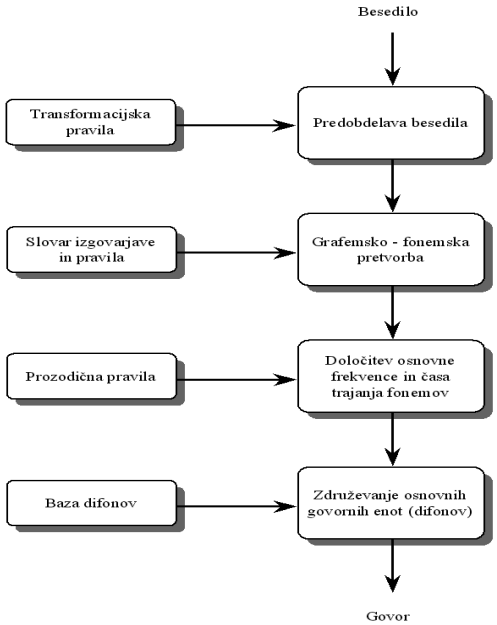
1 UVOD

Z razvojem in vse večjo uporabo računalniške tehnologije se je začela kazati čedalje večja potreba po sistemih, ki omogočajo komunikacijo med človekom in strojem v naravnem jeziku. Za to so potrebni tako sistemi za razpoznavanje in razumevanje govora, kakor tudi sistemi za sintezo govora. Raziskave na teh področjih potekajo v svetu že preko 30 let. Trenutni razvoj je prišel že tako daleč, da so takšni sistemi uporabni tudi v praksi.
V Sloveniji se trenutno ukvarjata s sintezo govora predvsem dve raziskovalni skupini: Odsek za inteligentne sisteme (vodja prof. dr. Ivan Bratko), ki deluje v okviru Institututa Jožef Stefan v Ljubljani in Laboratorij za umetno zaznavanje (vodja prof. dr. Nikola Pavešić) na Fakulteti za elektrotehniko v Ljubljani.

Na Institutu Jožef Stefan je 1993. leta S. Weilguny izdelal sistem za izgovorjavo izoliranih slovenskih besed [1]. A. Dobnikar je razvil modul za napovedovanje slovenske stavčne intonacije in trajanja premorov [2, 3]. Leta 1995 je bil zasnovan nov sintetizator slovenskega govora, sposoben samodejnega pretvarjanja poljubnih slovenskih besedil v govor [4, 5, 6, 7]. Razvoj podobnega sistema poteka vzporedno tudi na Fakulteti za elektrotehniko v Ljubljani [8, 9].

2 ZGRADBA SISTEMA

Sistem STTS – Slovenian Text-to-Speech System (http://zlatoust.ijs.si/stts/stts.htm) je zasnovan modularno [5, 6], kar omogoča enostavno popravljanje posameznih delov sistema, ki so povsem neodvisni drug od drugega. Vhod v sintetizator je poljubno slovensko besedilo predstavljeno v digitalni obliki. V hierarhični arhitekturi sistema [7] (slika 1) je na prvem mestu modul za predobdelavo besedila, sledita mu grafemsko - fonemski modul in modul za nastavljanje prozodičnih parametrov.



Slika 1: Sistem STTS za sintezo slovenskega govora.

Kot zadnji nastopa modul za združevanje osnovnih govornih enot, katerega rezultat je digitalni zapis umetnega govora, ki ga predvajamo preko zvočne kartice računalnika.

1.1 Predobdelava besedila

V fazi predobdelave vhodnega besedila najprej pretvorimo različne formate besedila v ASCII zapis. Nato odstranimo vse odvečne simbole, zapišemo okrajšave v njihovi polni obliki, cifre razvijemo v številke ter poiščemo pomen raznim ostalim zapisom in znakom (datumi, ure, ideogrami). Pri tem si pomagamo z najrazličnejšimi seznami pravil in slovarji, ki jih lahko po potrebi tudi dopolnjujemo. Opisani postopki se izvajajo poved za povedjo. Na koncu celotno besedilo razdelimo na posamezne besede in osnovna ločila.

Beseda	Izgovorjava
angleški	**angle:Ski**
cesta	**ce:sta**
čas	**Ca:s**
delavec	**de:lav@c**
delo	**de:lO**
delovnih	**de:lOwnih**
do	**dO:**
določen	**dOlo:C@n**
govorno	**gO:vOrnO**
izkušenj	**izku:S@n**
jezik	**jE:zik**
jeziki	**jEzi:ki**
let	**le:t**
mesecev	**me:sEc@w**
nedoločen	**nEdOlo:C@n**
ostali	**Osta:li**
pisno	**pi:snO**
pogoji	**pOgO:ji**
pomožni	**pOmo:Zni**
slovenski	**slOve:nski**
število	**StEvi:lO**
ulica	**u:lica**

Slika 2: Seznam nekaterih najpogostejših besed s področja zaposlovanja in njihova izgovarjava.

1.2 Grafemsko-fonemska pretvorba

Grafemsko-fonemsko pretvorbo lahko opišemo v treh korakih:

- pogledamo, ali se iskana beseda nahaja v slovarjih izgovarjav,
- če jo najdemo, prepišemo prevod besede v fonetični zapis, pri čemer dodatno upoštevamo koartikulacijo med sosednjimi besedami [9],
- besedam, ki jih ni v slovarjih izgovarjav, določimo mesto naglasa, temu sledi pretvorba v fonetični zapis.

Slovarje najbolj pogostih besed smo razdelili na več podslovarjev: števniki, lastna imena, akronimi, kolokacije, splošne besede (najobsežnejši del), besede iz danega področja uporabe (v našem primeru je to zaposlovanje oz. prosta delovna mesta; slika 2) [7].

Avtomatsko pretvorbo v fonetični zapis opravimo z večjim številom kontekstno odvisnih pravil [9, 10, 11]. Pri nastavljanju naglasnega mesta besed si pomagamo še s seznami naglašenih in nenaglašenih enot [7, 9]. Glede na to, da je za slovenski jezik značilno prosto mesto naglasa, ki se ga naučimo hkrati z učenjem jezika in besed, je to vse prej kot enostavno opravilo. Ta del zato trenutno predstavlja najšibkejšo točko našega sistema.

1.3 Določitev prozodičnih parametrov

Pravilna nastavitev prozodičnih parametrov, podanih v obliki osnovne frekvence in časa trajanja fonemov, močno vpliva tako na naravnost kot tudi razumljivost umetnega govora. Sintetizirani govor, pri katerem so upoštevane le srednje vrednosti osnovne frekvence in trajanja posameznih glasov, zveni močno nenaravno [9].

Postopek nastavljanja prozodičnih parametrov obsega tri korake [7]:

- nastavljanje trajanja,
- nastavljanje osnovne frekvence,
- določitev mesta in trajanja premorov.

Modeliranje trajanja

Govornim enotam v trajanju ene besede sprva priredimo inherentne dolžine, ki jih dobimo kot vsoto inherentnih dolžin glasov, vsebovanih v besedi. Dejavniki, ki jih obravnavamo so: ime in vrsta glasu, glasovni kontekst, položaj glasu znotraj besede, vrsta zloga, naglašenost zloga, položaj zloga v besedi. Ko se besede vključujejo v večje govorne enote, se skrajšujejo ali podaljšujejo v skladu z zahtevami višjenivojskih prozodičnih pojavov. Pri tem upoštevamo različne parametre, kot so položaj besede v frazi, izbrano hitrost govora in dolžino besede, izraženo s številom zlogov v besedi [7, 9].

Modeliranje osnovne frekvence

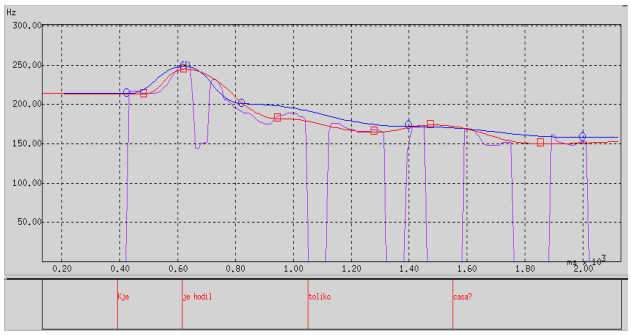
Pri nastavljanju osnovne frekvence smo uporabili tako imenovani "superpozicijski" model, ki definira potek osnovnega tona v intonacijskem segmentu kot vsoto [12]:

- globalne (nosilne) komponente in
- lokalnih komponent.

Globalna komponenta osnovnega tona je v intonacijskih segmentih povezana in relativno ravna ter odvisna od: vrste intonacijskega segmenta, položaja intonacijskega segmenta v sestavljenih oblikah povedi in dolžine intonacijskega segmenta [2, 3, 13].

Nad globalni potek osnovnega tona je dodana lokalna komponenta, ki je definirana za poudarjene besede oz. naglašene zloge v intonacijskem segmentu [13]. Pri tonskem naglaševanju je značilno naraščanje tona znotraj naglašenega zloga. Po dosegu tonskega vrha sledi

upadanje, ki je odvisno od vrste besede (oksiton ali bariton) ter od vrste naglasa (akut ali cirkumfleks) [8].



Slika 3: Rezultat modeliranja osnovne frekvence za stavek: "Kje je hodil toliko časa?". Kvadratki predstavljajo posnetek naravnega govora v sistemu INTSINT, krogci pa karakteristične točke umetno generiranega poteka osnovne frekvence.

Ker so prehodi med posameznimi fonemi skoraj vedno postopni in se govorni signal zvezno spreminja [2], je bilo potrebno za simuliranje naravnega poteka osnovnega tona poiskati zvezne funkcije brez nenadnih sprememb ali nezveznosti na stikih med posameznimi fonemi. Za globalno komponento smo izbrali eksponentno funkcijo, za lokalne komponente pa kosinusne funkcije (slika 3) [2, 13].

Določitev mesta in trajanja premorov

Glede na dolžino premorov in položaj v besedilu ločimo štiri skupine: premori pri naslovih (so najdaljši), premori na koncu povedi, premori v povedi na mestih ločil in premori v povedi na mestih ritmičnih delitev (običajno pred vezniki *in*, *pa*, *ter* in sicer v daljših stavkih) [2].

1.4 Združevanje osnovnih govornih enot

Baza difonov

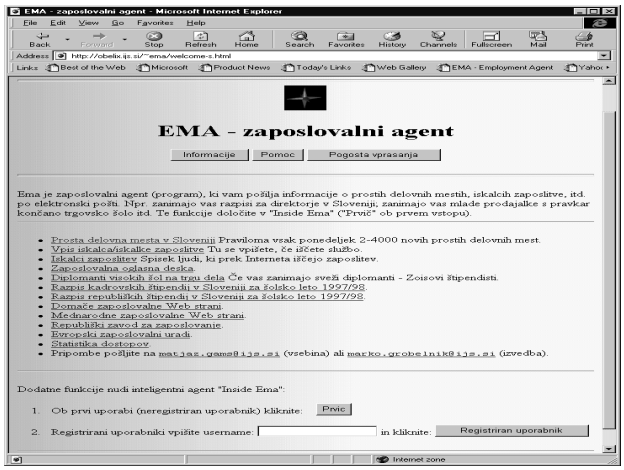
V posneti bazi osnovnih enot se nahaja 1155 difonov (ločimo med 33 različnimi glasovi) [5, 6]. Difoni so bili izrezani iz brezpomenskih besed, logatomov. Snemanje smo opravili v studijskih razmerah, pri čemer je govor prispeval profesionalni govorec s slovenskega nacionalnega radia. Razčlenjevanje in označevanje difonov je potekalo ročno, kar je precej dolgotrajen postopek [7]. V preteklosti je bilo narejenih že več poizkusov samodejnega pridobivanja zbirk difonov, vendar še nobeden do sedaj ni dal zadovoljivih rezultatov.

Sinteza umetnega govora

Za združevanje osnovnih govornih enot smo izbrali algoritem TD-PSOLA [14], ki smo ga dopolnili z linearno interpolacijo s spremenljivim številom interpoliranih period [16]. Algoritem TD-PSOLA je hiter in zagotavlja visoko kakovost sintetiziranega govora [15]. Omogoča sočasen nadzor tako nad osnovno frekvenco kot nad trajanjem sintetiziranega signala. Njegove pomanjkljivosti se kažejo v obliki spektralnih nezveznosti na mestih lepljenja. Te nezveznosti odpravimo s postopkom linearne interpolacije. Interpolacija poteka v časovnem prostoru, zato tudi ni računsko preveč zahtevna [17].

2 ZAPOSLOVALNI AGENT EMA

Opisani sistem za sintezo slovenskega govora smo uporabili v zaposlovalnem agentu EMA (http://www-ai.ijs.si/~ema/) [7, 18], ki posreduje informacije, vezane na zaposlovanje prek interneta (slika 4).
Sistem za sintezo slovenskega govora smo uporabili v najbolj obiskanem modulu, ki posreduje informacije o prostih delovnih mestih v Sloveniji. Besedilo, ki ga sintetiziramo, je deloma omejeno (npr. vrsta poklica), deloma neomejeno (npr. naslov). Ob izbiri govorne opcije sistem prek interneta posreduje uporabniku datoteko tipa WAV s prebranim obvestilom. Tipično obvestilo o prostem delovnem mestu vsebuje od 150 do 250 znakov, dolžina govorne datoteke pa znaša okoli 800 KB.



Slika 4: Zaposlovalni agent EMA.

3 PREIZKUS SINTETIZATORJA

Pri preizkusu kakovosti sintetiziranega govora je sodelovalo 11 poslušalcev v starostnem razponu med 22 in 53 leti. Od tega so bili trije poslušalci ženskega spola in osem moškega spola. Večina poslušalcev pred preizkusom še ni slišala sintetiziranega govora [7].
Pri prvem testu smo preverjali razumljivost posameznih besed. Poslušalcem smo predvajali 54 besed in jih prosili, da na pripravljene pole zapišejo vse, kar so slišali. Pri tem so zapisovali vsako besedo sproti. Iz napisanega besedila smo izračunali odstotek pravilno razpoznanih besed.
Pri drugem testu smo poslušalcem predvajali obvestila o prostih delovnih mestih. Zopet smo jih naprosili, da si

zapišejo vse, kar so slišali. Testiranci so imeli možnost sprotnega ustavljanja govora, da so lahko vpisali besedilo na papir. Nato smo preverili pravilnost posameznih delov obvestila.

Razumljivost izoliranih besed pri prvem testu je znašala 94.6 %, kar je primerljivo s podobnimi sintetizatorji v svetu [7]. Težave so bile predvsem pri krajših besedah, in še tam so poslušalci velikokrat zamešali le po en znak na besedo (npr. namesto stop so napisali stot); ponavadi sta bila to dva sorodna glasova.

Poslušalci tudi niso imeli večjih problemov pri razumevanju prebranih obvestil o prostih delovnih mestih. Bistvo sporočila so vselej pravilno razumeli. Nekaj več težav je bilo le pri razumevanju imena in naslova podjetja oziroma organizacije, ki nudi zaposlitev (slika 5) [7].

Deli besedila, na katere smo razdelili sporočila o prostih delovnih mestih	Odstotek pravilno razpoznanih delov besedila [%]
Poklic	100
Vrsta zaposlitve	100
Ostali pogoji	98
Datum, do katerega je potrebno poslati prošnjo	98
Ime podjetja oziroma organizacije, ki nudi zaposlitev	74
Naslov podjetja oziroma organizacije, ki nudi zaposlitev	87
Število prostih delovnih mest	100

Slika 5: Odstotki pravilno prepoznanih delov obvestil o prostih delovnih mestih.

4 ZAKLJUČEK

Razvili smo sistem STTS, ki je sposoben samodejnega pretvarjanja poljubnih slovenskih besedil v govor. Modularna zgradba sistema omogoča njegovo enostavno popravljanje in dograjevanje. Preizkus sistema je pokazal, da je tako dobljeni umetni govor dovolj razumljiv in povsem primeren za uporabo.

Sistem smo uporabili v zaposlovalnem agentu EMA in sicer v njegovem najbolj obiskanem modulu, ki posreduje informacije o prostih delovnih mestih v Sloveniji. Sistem je na razpolago tudi v obliki programskih knjižnic, kar omogoča njegovo vključitev v poljubno aplikacijo. Ena od takih aplikacij, ki jih na Odseku za inteligentne sisteme trenutno razvijamo, je sistem, ki bo slepim in slabovidnim omogočil delo v okolju Windows.

5 LITERATURA

[1] S. Weilguny, *Grafemsko fonemski modul za sintezo slovenskega jezika*, magistrsko delo, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani, 1993.

[2] A. Dobnikar, *Določevanje stavčne intonacije pri sintezi slovenskega govora*, doktorska disertacija, Fakulteta za elektrotehniko, Univerza v Ljubljani, 1997.

[3] A. Dobnikar, *Modeling Segment Intonation for Slovene TTS System*, Proc. ICSLP'96, str. 1864-1867, Philadelphia, 1996.

[4] Dobnikar, J. Bakran, *A New Approach for Slovene Text-to-Speech Synthesis*, Proceedings of Microcomputers in Intelligent Information Systems MIS, MIPRO'95, Opatija, str. 312-318, 1995.

[5] T. Šef, A. Dobnikar, M. Gams, *Text-to-Speech Synthesis in Slovenian Language*, Proceedings of the EUSIPCO'98, Greece, 1998, sprejeto v objavo.

[6] T.Šef, A. Dobnikar, Recent Improvements in Slovene Text-to-Speech Synthesis, Proceedings of the ICSLP'98, Sydney, Australia, 1998, sprejeto v objavo.

[7] T. Šef, *Sistem za govorno posredovanje obvestil o prostih delovnih mestih*, magistrska naloga, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, 1998.

[8] J. Gros, N. Pavešić, S. Dobrišek, M. Erpič, B.Grenc, A. Rakar, T. Šef, V. Vračar, F. Mihelič, *Sistem za sintezo slovenskega govora*, Zbornik četrte Elektrotehniške in računalniške konference ERK'95, str. 265-268, Portorož 1995.

[9] J. Gros, *Samodejno tvorjenje govora iz besedil*, doktorska disertacija, Fakulteta za elektrotehniko, Univerza v Ljubljani, 1997.

[10] J. Hribar, *Sinteza umetnega govora iz teksta*, magistrsko delo, Fakulteta za elektrotehniko in računalništvo, 1984.

[11] J.Toporišič, *Slovenska slovnica*, Založba Obzorja, Maribor, 1984.

[12] H. Fujisaki, S. Ohno, *Analysis and Modeling of fundamental Frequency Contour of English Utterances*, Proceedings of the EUROSPEECH'95, str. 985-988, Madrid,1995.

[13] A. Dobnikar, T. Šef, *Modeliranje intonacijskih krivulj slovenskega branega govora*, Zbornik šeste Elektrotehniške in računalniške konference ERK'97, str. 217-220, Portorož 1997.

[14] T. Šef, *Spreminjanje prozodičnih lastnosti osnovnih enot sintetiziranega govora*, diplomska naloga, Fakulteta za elektrotehniko in računalništvo, Univerza v Ljubljani, 1995.

[15] E. Moulines, F. Charpentier, *Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones*, Speech Communications (9), str. 453-467, 1990.

[16] T. Šef, A. Dobnikar, *Izpopolnjeni algoritem za spreminjanje osnovnega tona in časa trajanja difonov pri sintezi govora*, Zbornik šeste Elektrotehniške in računalniške konference ERK'97, str. 221-224, Portorož, 1997.

[17] T. Dutoit, H. Leich, *MBR-PSOLA: Text-To-Speech synthesis based on an MBE re-synthesis of the segments database*, Speech Communication (13), str. 435-440, 1993.

[18] M. Gams, V. Križman, T. Šef, *An Employment Agent with a NL Interface*, International Conference on Systems, Signals, Control, Computers (SSCC'98), Južna Afrika, 1998, sprejeto v objavo.